Simplificação e análise de redes com dados multivariados

Markus Diego Sampaio da Silva Dias

Tese de Doutorado do Programa de Pós-Graduação em Ciências de
Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Markus Diego Sampaio da Silva Dias

Simplificação e análise de redes com dados multivariados

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Luis Gustavo Nonato

USP – São Carlos
Dezembro de 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

D541s Dias, Markus Diego Sampaio da Silva
Simplificação e análise de redes com dados
multivariados / Markus Diego Sampaio da Silva Dias;
orientador Luis Gustavo Nonato. -- São Carlos, 2018.
81 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2018.

1. Redes. 2. Agrupamento. 3. Fatoração de
matrizes não negativas. 4. Visualização. 5.
Emparelhamento em um grafo. I. Nonato, Luis
Gustavo, orient. II. Título.

Markus Diego Sampaio da Silva Dias

Simplification and analysis of network with multivariate data

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Luis Gustavo Nonato

USP – São Carlos
December 2018

Este trabalho é dedicado à minha família e minha namorada.

AGRADECIMENTOS

Meus primeiros e principais agradecimentos são para meus pais e meu irmão, por terem me apoiado, sempre acreditarem em mim e por terem sido peça fundamental para que eu conseguisse terminar esta tese. Diante de todos os problemas de saúde, psicológicos e toda jornada nos últimos anos, não seria possível concluir esse trabalho sem o apoio emocional, financeiro e sem eles estarem ao meu lado em todos os momentos. Agradeço meus pais por toda a educação e pela criação durante toda a minha vida que fizeram me tornar o homem que sou.

Agradeço a minha namorada, Karina, que nesse quase um ano juntos vem me dando apoio incondicional para que eu possa alcançar meus objetivos e possamos caminhar juntos. Agradeço por todos os momentos vividos, pelos dias alegres e por ter me mostrado um mundo diferente, mais colorido e mais bondoso ao que eu me acostumei a viver. Nos encontramos em uma época difícil para mim e foi uma pessoa fundamental em minha melhora, uma verdadeira companheira.

Agradeço a todos meus amigos que de alguma forma, mesmo que pequena, me ajudaram. Aos amigos antigos de USP nos mais de dez anos pela amizade, pelas conversas, pelos estudos e pela vida que dividimos durante estes anos em São Carlos. Agradeço a todos aos que conheci durante esses anos, aos que ainda vejo e aos que hoje eu não tenho mais contato.

Agradeço a todos amigos, alunos e professores do grupo de pesquisa pelos anos de trabalho. Agradeço ao Fabiano, Moussa, Luzia e Fábio que colaboraram de forma mais próxima nos trabalhos durante o doutorado.

Agradeço ao meu orientador Luis Gustavo Nonato pelos anos de orientação, por todos os ensinamentos, toda a ajuda e pela paciência com todos os acontecimentos.

Agradeço a todos os professores da USP que durante mais de uma década contribuíram para minha formação. Principalmente aos meus orientadores anteriores Marcelo Saia e Maria Aparecida Ruas.

Agradeço a todos os funcionários do ICMC e da USP São Carlos que nos proporcionam ótimas condições de trabalho e estudo.

Agradeço a CAPES pelo auxílio financeiro durante o doutorado.

*“Que os vossos esforços desafiem as impossibilidades,
lembrai-vos de que as grandes conquistas do homem
foram conquistadas do que parecia impossível.”*
(Charles Chaplin)

RESUMO

DIAS, M. D. S. S. **Simplificação e análise de redes com dados multivariados**. 2018. 81 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

As técnicas de visualização desempenham um papel importante na assistência e compreensão de redes e seus elementos. No entanto, quando enfrentamos redes massivas, a análise tende a ser prejudicada pela confusão visual. Esquemas de simplificação e agrupamento têm sido algumas das principais alternativas neste contexto. No entanto, a maioria das técnicas de simplificação consideram apenas informações extraídas da topologia da rede, desconsiderando conteúdo adicional definido nos nós ou arestas da rede.

Neste trabalho, propomos dois estudos. Primeiro uma nova metodologia para simplificação de redes que utiliza tanto a topologia quanto o conteúdo associado aos elementos de rede. A metodologia proposta baseia-se na fatoração de matriz não negativa (NMF) e emparelhamento para realizar a simplificação, combinadas para gerar uma representação hierárquica da rede, agrupando elementos semelhantes em cada nível da hierarquia. Propomos também um estudo da utilização da teoria de processamento de sinal em grafos para filtrar os dados associados aos elementos da rede e o seu efeito no processo de simplificação.

Palavras-chave: Redes, Agrupamento, Fatoração de matrizes não negativas, Visualização, Emparelhamento de grafos.

ABSTRACT

DIAS, M. D. S. S. **Simplification and analysis of network with multivariate data**. 2018. 81 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Visualization tools play an important role in assisting and understanding networks and their elements. However, when faced with larger networks, analytical tasks can be hindered by visual clutter. Schemes of simplification and clustering have been a main alternative in this context. Nevertheless, most simplification techniques consider only information extracted from the network topology, disregarding additional content defined in nodes or edges.

In this paper, we propose two studies. First, a new methodology for network simplification that uses both topology and content associated with network elements. The proposed methodology is based on non-negative matrix factorization (NMF) and graph matching to perform the simplification, combined to generate a hierarchical representation of the network, grouping the most similar elements at each level of a hierarchy. We also provide a study of the use of the graph signal processing theory to filter data associated to the elements of a network and its effect in the process of simplification.

Keywords: Networks, Clustering, Non-negative matrix factorization, Visualization, Graph matching.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de grupos gerados por remoção de arestas. Imagem retirada de (NEWMAN; GIRVAN, 2004)	28
Figura 2 – Comparação entre as simplificações derivadas de três diferentes medidas de centralidades. Imagem retirada de (CORREA <i>et al.</i> , 2009)	29
Figura 3 – Exemplo de simplificação do tipo Matching Neighbors. Imagem retirada de (DWYER <i>et al.</i> , 2013)	29
Figura 4 – Grafo original / layout do grafo híbrido / layout do grafo híbrido distorcido. Imagem retirada de (GANSNER; KOREN; NORTH, 2005)	31
Figura 5 – Colapso de arestas.	35
Figura 6 – Visão geral da nossa metodologia.	36
Figura 7 – SMM x MWM. O custo do emparelhamento de custo máximo é 15, mas a aresta entre os nós adjacentes mais similares não é selecionada. O emparelhamento maximal ordenado sempre escolhe as arestas com maior peso a cada etapa, representando os nós mais similares, embora resultando em um custo total reduzido e possivelmente um número menor de arestas colapsadas. . .	41
Figura 8 – Uma visualização de nó-aresta da hierarquia da rede artificial. O tamanho do metanó representa o número de nós do nível zero que foram agrupados nele. As arestas do emparelhamento colapsadas em cada etapa foram coloridas de vermelho. Os nós foram coloridos a partir do nível mais baixo para os níveis acima. Isto é, os nós que foram agrupados em algum nível da hierarquia são nós que são relativamente similares no nível zero da hierarquia.	43
Figura 9 – A rede College Football e os cinco níveis da representação hierárquica da nossa abordagem (no sentido horário a partir do canto superior esquerdo). A figura inferior esquerda compara os metanós no nível último nível (grandes nós) com aqueles no nível zero / original (nós pequenos). Os nós são coloridos a partir do nível mais baixo para os níveis acima.	43
Figura 10 – De cima para baixo, medições de qualidade (mediana / melhores valores) utilizando a função de modularidade, a Δ -measure modificada e a k-way Ratio Cut Cost aplicadas nas redes Artificial, College Football e VIS. . . .	47

Figura 11 – Comparação com outras métricas de similaridade. O SMM representa o emparelhamento maximal ordenado e o MWM é o emparelhamento de custo máximo. SMM_f e MWM_f representam os testes em que o passo da filtragem de arestas foi realizado. As linhas horizontais são as medianas (linha azul) e o melhor resultado (linha vermelha) do nosso método indicados na figura 10. Para o dataset artificial a mediana e o melhor resultados são os mesmos em cada métrica de validação.	49
Figura 12 – Comparação com outras técnicas de simplificação. Cada coluna (dataset) é dividida em quatro faixas de cor. As três primeiros são do nosso método, um para cada método de decomposição. Cada faixa de cores contém dois box plots, que correspondem ao esquema de emparelhamento SMM proposto sem (à esquerda em cada faixa de cor) e com (à direita em cada faixa de cor) o mecanismo de filtragem de arestas. Faixas acastanhadas (faixa de cor da direita em cada coluna) correspondem aos resultados do agrupamento tradicional da NMF. Os métodos de Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994) são determinísticos e seus resultados são mostrados nas linhas horizontais.	50
Figura 13 – Visualização de três etapas do dataset Open Flight ao ser simplificado utilizando nosso método.	52
Figura 14 – Na figura central temos um grafo com sinal definido por uma função degrau (o sinal nos nós é 0 ou 1) com alguns outliers (sinal 0,5). Nas figuras laterais, temos o efeito do filtro passa-baixa (esquerda) e do filtro de realce (direita).	56
Figura 15 – Etapas da simplificação da rede do dataset ecoli usando os dados brutos (parte superior) e os dados filtrados com o filtro passa-baixa de Fourier (parte inferior).	61
Figura 16 – Etapas da simplificação da rede do dataset Wine usando os dados brutos (parte superior) e os dados filtrados com o filtro passa-baixa de Fourier (parte inferior).	61
Figura 17 – Comparação da simplificação usando os dados brutos e os dados filtrados utilizando a métrica de modularidade (topo), Δ -Measure (centro) e a métrica k -Way Ratio Cut Cost (abaixo) para cada rede. A barra azul representa os resultados dos dados brutos. A barra verde representa os resultados dos dados filtrados com o filtro passa-baixa. A barra amarela representa os resultados do filtro de realce. O melhor resultado na métrica de modularidade é o maior número, enquanto nas métricas Δ -Measure e k -Way Ratio Cut Cost, os melhores resultados são os menores.	63
Figura 18 – Comparação da metodologia do Filtro Espectral do Grafo aliada a HNMF com a HNMF sem o filtro utilizando a rede VIS Conference.	64
Figura 19 – <i>Pipeline do Vizlattes.</i>	66
Figura 20 – <i>Representação dos caminhos relevantes através da matriz de termo-frequência.</i>	67

Figura 21 – <i>Função de transferência de peso da frequência das palavras.</i>	68
Figura 22 – Elementos que serão somados para calcular a relevância do tópico l em uma comunidade c	69
Figura 23 – Elementos que serão somados para calcular a relevância do um autor i em uma comunidade c relativa a um tópico l	70
Figura 24 – <i>Construção do método de visualização. Cada barra representa uma comunidade. As cores representam diferentes tópicos.</i>	71
Figura 25 – <i>Método de visualização do Vizlattes.</i>	72

LISTA DE TABELAS

Tabela 1 – Datasets	44
Tabela 2 – Datasets	60
Tabela 3 – Esta tabela resume os tópicos encontrados pela decomposição e respectivos autores proeminentes extraídos com o método proposto da rede de colabora- ção do corpo docente do ICMC em 2015.	73

SUMÁRIO

1	INTRODUÇÃO	23
2	REVISÃO BIBLIOGRÁFICA	27
3	HIERARCHICAL NON-NEGATIVE MATRIX FACTORIZATION . .	35
3.1	Decomposição de matrizes	36
3.1.1	<i>Fatoração de matrizes não-negativas</i>	36
3.1.2	<i>Fatoração de matrizes não-negativas com grafo regularizador</i>	37
3.1.3	<i>Fatoração de matrizes não-negativas simétricas</i>	38
3.2	Similaridade dos nós	38
3.3	Emparelhamento e agrupamento	39
3.4	Hierarquia	41
3.5	Datasets e métricas	42
3.5.1	<i>Datasets</i>	42
3.5.2	<i>Métricas de validação</i>	44
3.6	Resultados e comparações	46
3.6.1	<i>Emparelhamento de grafos e filtragem de arestas</i>	46
3.6.2	<i>Similaridade dos nós e pesos das arestas</i>	47
3.6.3	<i>Comparações com outras técnicas de simplificação</i>	48
3.7	Discussões e Limitações	51
4	FILTRAGEM ESPECTRAL EM GRAFOS APLICADA À SIMPLIFI- CAÇÃO DE REDES	53
4.1	Processamento de sinais em grafos	53
4.1.1	<i>Transformada de Fourier em grafos</i>	53
4.1.2	<i>Filtro espectral em grafo</i>	55
4.2	Filtro espectral em dados multidimensionais e simplificação da rede	55
4.2.1	<i>Filtro Spectral em dados multidimensionais</i>	56
4.2.2	<i>Similaridade dos nós</i>	57
4.2.3	<i>Simplificação hierárquica da rede</i>	58
4.3	Resultados e comparações	59
4.3.1	<i>Datasets</i>	59
4.3.2	<i>Visualização nó-aresta e comportamento do agrupamento</i>	60
4.3.3	<i>Simplificação da rede</i>	62

4.4	Discussões e limitações	62
5	VIZLATTES	65
5.1	Metodologia	66
5.1.1	<i>Redes de colaboração científica</i>	66
5.1.2	<i>Caminhos relevantes</i>	66
5.1.3	<i>NMF e métricas</i>	68
5.1.4	<i>VizLattes</i>	70
5.2	Experimentos	72
5.3	Discussões e limitações	73
6	CONCLUSÃO	75
	REFERÊNCIAS	77

INTRODUÇÃO

As redes são estruturas importantes para modelar conjuntos cujos elementos possuem uma relação par a par. Exemplos típicos são redes sociais, onde as pessoas interagem de acordo com sua relação de amizade; redes científicas, onde os pesquisadores estão conectados de acordo com seu nível de colaboração; redes de aeroportos que estão conectados se possuem uma rota entre dois aeroportos.

Ferramentas de visualização são cruciais para a análise de redes, revelando informações como formação de grupos e padrões nas relações entre os elementos. No entanto, a manipulação e visualização de redes massivas trazem vários tipos de desafios. Por exemplo, quando redes são visualizadas como diagramas nó-aresta, elas apresentam uma visualização confusa com vários nós e arestas se cruzando.

Desta forma, simplificar de forma hierárquica a rede tem sido uma das principais alternativas para que redes massivas possam ser visualizadas e manipuladas. Só para dar uma ideia da importância de técnicas de simplificação e hierarquização, dos 81 métodos de visualização de redes analisados por Vehlow et al. ([VEHLOW; BECK; WEISKOPF, 2015](#)), 43 utilizam uma estrutura hierárquica para auxiliar nas tarefas de visualização.

Muito já foi feito na tentativa de simplificar redes, tanto na área de teoria de grafos, quanto nas áreas de reconhecimento de padrões e de visualização de redes. Porém a grande maioria das técnicas considera apenas a topologia, desconsiderando possíveis informações associadas aos elementos da rede. Negligenciar essas informações durante a simplificação pode gerar resultados e estruturas cujo conteúdo mal reflete os originais. Além disso, gerar uma simplificação hierárquica utilizando dados associados aos elementos da rede traz novos desafios como, por exemplo, definir os dados referentes aos elementos gerados em cada nível da hierarquia de forma a preservar propriedades relevantes.

No capítulo 3 propomos uma nova metodologia para simplificação de redes chamada Hierarchical Non-Negative Matrix Factorization ([DIAS *et al.*, 2017](#)) baseada na informação

topológica e também no conteúdo associado aos elementos de rede, utilizando fatoração de matrizes não-negativas (NMF) (LEE; SEUNG, 1999; CAI *et al.*, 2011; KUANG; PARK; DING, 2012) e emparelhamento de grafos (BONDY; MURTY, 2007). O método pode ser aplicado repetidamente para gerar uma representação hierárquica da rede, agrupando elementos semelhantes em cada nível da hierarquia. Primeiro utilizamos a NMF para decompor os dados e estabelecer uma métrica de similaridade entre os nós. Depois a técnica de emparelhamento do grafo resulta em um conjunto de arestas ligando nós semelhantes que ao serem colapsadas dão origem à rede simplificada. A decomposição da matriz é realizada apenas no início do processo, reduzindo o custo computacional sem comprometer a qualidade do resultado. Os atributos da nova rede são definidos através do resultado da NMF nos nós agrupados. Repetidamente aplicamos a metodologia acima para gerar uma hierarquia.

No capítulo 4 propomos uma metodologia para melhorar a simplificação da rede usando Graph Signal Processing (GSP) (SHUMAN *et al.*, 2013). A metodologia proposta baseia-se em mecanismos de filtragem espectral para evitar possíveis problemas na simplificação causados pelos dados associados aos elementos da rede. A etapa de filtragem da nossa metodologia torna a distribuição de atributos mais homogênea ao longo da rede, de modo que os nós terão atributos mais semelhantes aos seus vizinhos. Além disso, adaptamos o filtro espectral a dados multidimensionais. Assim, podemos filtrar a matriz de atributos da rede. Nossa metodologia também melhora a visualização dos atributos reduzindo o ruído.

Em resumo, as principais contribuições desta tese são:

- Um método de simplificação de redes com dados multivariados utilizando fatoração de matrizes não-negativas e emparelhamento em grafo.
- Uma representação hierárquica da rede classificando os nós em todos os níveis da hierarquia gerada.
- Um conjunto de testes e comparações garantindo a eficácia da metodologia proposta.
- Uma adaptação da teoria de filtro espectral de grafo para redes com dados multidimensionais.
- Um estudo da utilização da filtragem do sinal como forma de fazer um pré-processamento dos dados para simplificar uma rede comparando a topologia da rede e a qualidade da simplificação final.

A tese está estruturada da seguinte forma:

- No capítulo 2 fazemos uma revisão dos trabalhos relacionados a simplificação de redes.
- No capítulo 3 descrevemos a técnica HNMF (DIAS *et al.*, 2017) de simplificação de redes com dados multivariados.

- No capítulo 4 mostramos um estudo sobre a utilização da teoria de processamento de sinais em grafos como ferramenta de pré-processamento dos dados voltada para a simplificação de uma rede (DIAS *et al.*, 2018).
- No capítulo 5 apresentamos o Vizlattes (DIAS *et al.*, 2015), uma aplicação da NMF para visualização de redes de colaboração científica.
- No capítulo 6 fazemos as conclusões sobre o trabalho.

Durante o doutorado foram publicados os trabalhos “VizLattes: a Tool for Relevance Analysis from Scientific Co-Authorship Networks” (DIAS *et al.*, 2015), “A Hierarchical Network Simplification via Non-Negative Matrix Factorization” (DIAS *et al.*, 2017) e “Graph Spectral Filtering for Network Simplification” (DIAS *et al.*, 2018). O segundo recebeu o prêmio de menção honrosa na premiação de melhor trabalho de computação gráfica e visualização na track principal do SIBGRAPI 2017.

REVISÃO BIBLIOGRÁFICA

Apresentaremos nesse capítulo trabalhos relevantes sobre simplificação de redes e trabalhos sobre visualização de redes que fazem uso de simplificação na rede no processo de visualização. Os trabalhos e resultados discutidos aqui foram escolhidos por apresentarem uma sólida formulação matemática ou computacional. Para efeito de notação, redes e grafos serão sinônimos nessa tese. Utilizaremos rede sempre que possível, utilizando a palavra grafo em técnicas e teorias com grafo em seu nome.

Alguns métodos de simplificação se baseiam em remover elementos da rede enquanto preservam informações importantes e de interesse do usuário. Por exemplo, a técnica Pathfinder (DEARHOLT; SCHVANEVELDT, 1990) e suas variantes (QUIRIN *et al.*, 2008; HAUGUEL; ZHAI; HAN, 2009; BOTE *et al.*, 2006) tem como foco preservar caminhos na rede enquanto removem as arestas não importantes. Todos os caminhos mínimos entre dois nós são calculados. As arestas que não pertencem a nenhum desses caminhos são removidas, mantendo então apenas os caminhos de menor distância entre dois nós. Em (NOCAJ; ORTMANN; BRANDES, 2014) e (NOCAJ; ORTMANN; BRANDES, 2016) é utilizada uma métrica nas arestas para decidir quais serão excluídas. A métrica depende apenas da topologia da rede e tentam preservar as comunidades da rede. Uma etapa de filtragem remove todas as arestas abaixo de um determinado valor resultante da métrica.

Ainda neste contexto temos os trabalhos de Girvan e Newman (NEWMAN; GIRVAN, 2004; GIRVAN; NEWMAN, 2002), baseados em remoção de arestas utilizando a medida de centralidade da intermediação. Essa medida pode ser calculada para cada nó ou aresta. A centralidade da intermediação no nó é definida como a proporção entre o número de caminhos mínimos que o nó participa pelo número total de menores caminhos da rede. É uma medida importante, pois nós com intermediação alta são os nós que ligam regiões distantes da rede, ou seja, grupos diferentes. A definição de centralidade da intermediação para arestas é análoga. Em (NEWMAN; GIRVAN, 2004), a centralidade da intermediação é calculada para todas

as arestas e a aresta com o maior valor é removida. Depois as centralidades de todas as arestas são calculadas novamente e o processo se repete. O agrupamento é definido pelos grupos que restam após a remoção de um determinado número de arestas (figura 1). O número ideal de grupos em que deve ser dividida a rede é encontrado utilizando a métrica de modularidade (definida com detalhes em 3.5.2) que nos fornece a qualidade de um agrupamento. Ela é calculada nos diferentes agrupamentos encontrados removendo as arestas e o número ideal de grupos é o que resultar o agrupamento com a maior qualidade segundo essa métrica. Em outro trabalho (GIRVAN; NEWMAN, 2002), as arestas são removidas com o intuito de gerar uma árvore mínima. Seguindo um caminho diferente dos trabalhos citados, Jia et al. (JIA *et al.*, 2008) remove as arestas com a menor centralidade. O objetivo é gerar uma rede com um número menor de arestas que preserve os caminhos entre grupos diferentes.

Figura 1 – Exemplo de grupos gerados por remoção de arestas. Imagem retirada de (NEWMAN; GIRVAN, 2004)

Diferentes medidas de centralidade são utilizadas para medir a importância dos nós. A partir da centralidade, a sensibilidade pode ser definida como a derivada da centralidade e é geralmente utilizada para indicar a taxa de variação ao adicionar ou remover algum elemento da rede (CORREA; CRNOVRSANIN; MA, 2012). Com isso podemos avaliar a simplificação de uma rede medindo a preservação da centralidade ao adicionar/remover elementos da rede. Seguindo essa linha, a sensibilidade é utilizada para remover arestas até criar uma árvore geradora mínima da rede (CORREA *et al.*, 2009). Na figura 2 é exibido alguns resultados em algum passo da simplificação para diferentes medidas de centralidade.

Métodos de remoção de arestas/nós não são apropriados para certas aplicações, principalmente quando a conectividade do nó é relevante para a análise. Porém, muitas dessas técnicas consideram apenas a topologia, desconsiderando conteúdo adicional associado às arestas e nós de rede, restringindo sua utilidade.

Outra forma de simplificar uma rede é através do colapso de nós/arestas. Estes métodos geralmente buscam agrupar os nós baseados em alguma medida de similaridade entre eles. Técnicas como Matching Neighbors e Power Graph Analysis tem como objetivo visualizar redes substituindo vários nós por um metanó, baseando-se na vizinhança dos nós, de modo a facilitar a compreensão da rede como um todo. As técnicas do tipo Matching Neighbors (figura 3) são as

Figura 2 – Comparação entre as simplificações derivadas de três diferentes medidas de centralidades. Imagem retirada de (CORREA *et al.*, 2009)

mais simples, pois agrupam os nós que possuem a mesma vizinhança, diferindo apenas por eles próprios.

Figura 3 – Exemplo de simplificação do tipo Matching Neighbors. Imagem retirada de (DWYER *et al.*, 2013)

Dois trabalhos baseados em Matching Neighbors são os de Ham (HAM; WATTENBERG; VIEGAS, 2009) e o de Dinkla (DINKLA; WESTENBERG; WIJK, 2012). O trabalho de Ham é o sistema de visualização de textos Phrase Nets, onde nós com vizinhanças iguais são agrupados e dois grupos são ligados por uma aresta se existe uma aresta ligando dois nós de cada um dos grupos na rede original. Dinkla propôs uma técnica similar para agrupar nós com conjuntos de vizinhos idênticos, no contexto de análise de redes genéticas, para produzir representações utilizando matrizes compactas. A sua grande desvantagem é que entender a representação por matrizes não é muito intuitivo.

A grande vantagem das técnicas do tipo Power Graph Analysis (YOGHOORDJIAN *et al.*, 2016; DWYER *et al.*, 2013; DWYER *et al.*, 2014) é que elas agrupam os nós baseados nas suas vizinhanças, mas possuem algum relaxamento na sua formulação. Não há mais a obrigatoriedade de que os nós tenham vizinhanças idênticas. A desvantagem dessas técnicas com relação ao que buscamos no nosso trabalho é que elas se baseiam na vizinhança do nó, ou seja, utilizam apenas a topologia da rede, não levando em consideração possíveis atributos.

Entretanto, há outros métodos que colapsam elementos baseados em critérios de otimização ou no conteúdo associado aos elementos da rede. O método proposto por Newman ([NEWMAN, 2003](#)) é um exemplo típico. É uma técnica simples, mas importante por utilizar a função de modularidade, já citada acima, que consegue calcular a qualidade de um agrupamento da rede. A ideia consiste em colapsar uma aresta de forma a achar o agrupamento de maior modularidade. Calculam o valor da função de modularidade da rede caso uma aresta seja colapsada. Fazem isso para todas arestas e colapsam a que produziu o maior valor. Repetem o processo novamente calculando a modularidade para todas arestas e colapsando uma delas. Assim geram uma estrutura hierárquica do grafo, porém agrupam apenas dois nós a cada passo.

Teoria espectral é muito utilizada para realizar agrupamentos de redes. Em seu trabalho, Chan et al. ([CHAN; SCHLAG; ZIEN, 1994](#)) elaboraram uma técnica baseada em minimizar a métrica k -Way Ratio Cut Cost, que calcula o custo de um agrupamento de uma rede em k grupos. Na técnica é demonstrado que os k menores autovetores/autovalores minimizam essa métrica. Com esses k autovetores é realizado um mergulho dos dados em um espaço k -dimensional dado pelo espectro do grafo. Os dados são agrupados de acordo com esse mergulho no espaço espectral. Que seria a melhor partição dos nós.

A técnica K-SNAP ([TIAN; HANKINS; PATEL, 2008](#)) utiliza uma função de custo entre grupos baseada nos atributos dos elementos da rede e também em suas relações. Além disso, atributos utilizados pra criar a partição são definidos por escolha do usuário. Porém os atributos são categóricos, e não numéricos. Logo não há uma relação de distância. Os agrupamentos são feitos com nós de mesmos atributos e as relações baseadas nas ligações dos nós. Baseando-se na K-SNAP, Zhang et al. ([ZHANG; TIAN; PATEL, 2010](#)) propôs uma técnica que classifica os grupos utilizando uma ordenação que categoriza dados numéricos. Primeiro, nós com mesmos atributos numéricos são agrupados. Em uma segunda fase, nós são agrupados de acordo a semelhança gerada pela ordenação criada (os que estão seguidos na ordenação). Em uma última fase, os nós são agrupados de acordo com a qualidade do agrupamento de forma que minimize a perda da qualidade. A métrica de qualidade Δ -Measure (que será utilizado nos nossos resultados) é utilizada para esse cálculo.

Descrição de comprimento mínimo também é utilizada por Navlkha et al. ([NAVLAKHA; RASTOGI; SHRIVASTAVA, 2008](#)) para gerar a simplificação de uma rede. A ideia é gerar uma representação simplificada da rede original de forma que ela possa ser recuperada com custo mínimo. Isso é feito de duas formas. Uma em que o melhor par de nós é escolhido para se agrupar e outra em que dado um nó aleatório, é escolhido o melhor nó adjacente para agrupar com ele.

Como nesta tese, outros trabalhos também utilizam emparelhamento em grafo. Karpis e Kumar ([KARPIS; KUMAR, 1998](#)) exploram o fato do emparelhamento conseguir fornecer uma hierarquia natural ao ser utilizado de forma repetida. Os autores propõem outra forma de calcular um agrupamento maximal de forma eficiente. O emparelhamento aleatório proposto

é feito escolhendo um nó aleatório u e depois outro nó v , adjacente a u , é escolhido de forma também aleatória. Se ambos não estiverem no emparelhamento, a aresta (u, v) é adicionada ao emparelhamento. Porém, não existem atributos nos nós, então essas escolhas podem ser realizadas de forma aleatória. O emparelhamento de grafos tem sido um mecanismo importante para criar representações hierárquicas em outras áreas, incluindo desenho de grafos (WALSHAW, 2001; HADANY; HAREL, 1999). No entanto, a maioria desses métodos baseados em emparelhamento usam apenas na topologia da rede.

A técnica Topological Fisheye (GANSNER; KOREN; NORTH, 2005) também utiliza o emparelhamento para construir uma representação hierárquica de um grafo. Para escolher as arestas do emparelhamento que serão colapsadas, são utilizadas métricas de semelhança entre nós que combinam a informação topológica e geométrica do grafo. Descreveremos melhor essas métricas em 3.6.2, as quais iremos comparar com nossa métrica de similaridade entre nós. Com a hierarquia gerada, os autores definem um grafo híbrido da seguinte forma. Um nó é escolhido como foco. Conforme você se afasta do foco, os nós são substituídos por grupos da representação hierárquica. A visualização desse grafo será tipo olho de peixe, ou seja, a região em volta desse nó é vista mais detalhadamente. Para isso é realizada uma distorção para melhorar a visualização em volta da região em volta do foco. (veja figura 4).

Figura 4 – Grafo original / layout do grafo híbrido / layout do grafo híbrido distorcido. Imagem retirada de (GANSNER; KOREN; NORTH, 2005)

Decomposições de matrizes também são técnicas muito utilizadas para realizar a simplificação de uma rede. Wang et al (WANG *et al.*, 2008) utiliza a fatoração de matriz não-negativa (NMF (LEE; SEUNG, 1999)) para decompor a matriz de similaridade de um grafo, A , em duas matrizes W e H . As colunas de H representam as colunas da matriz de similaridade. A matriz de similaridade do grafo é construída a partir de medidas de similaridade baseadas na topologia do grafo. Cada coluna h_u é referente a um nó u . Os nós são agrupados de acordo com o maior valor encontrado na sua coluna h_u . Se o maior valor for h_{iu} , o nó u pertencerá ao grupo i . Falaremos com mais detalhes sobre classificação de uma rede baseada na NMF na seção 3.1.

A técnica Vegas (SHI *et al.*, 2015) utiliza a NMF simétrica (KUANG; PARK; DING, 2012) para resolver o problema de sumarização de grafos de influência e simplificar redes de citação. A GNMF (CAI *et al.*, 2011) é uma variação da NMF que utiliza um grafo regularizador para guiar a decomposição das matrizes. Após ter as matrizes W e H , é feito o agrupamento

do conjunto de dados (mais detalhes em 3.1.2). A técnica Multivis (SUN *et al.*, 2009) utiliza informação de e-mails como atributos da rede. Nessa técnica, ao invés de decompor uma matriz, essa informação dos e-mails é utilizada para gerar um tensor. Esse tensor é decomposto para então simplificarem a rede. Embora a NMF seja eficaz para definir grupos de elementos similares, ela não foi devidamente explorada no contexto de gerar representações hierárquicas onde a similaridade entre os elementos é aplicada em cada nível de uma hierarquia, como pretendemos fazer.

Embora os recursos visuais sejam alternativas, como edge bundling (ZHOU *et al.*, 2013), remoção ou colapso de nós e arestas ainda permanecem como abordagens essenciais no contexto de visualização de redes. Auber *et al.* (AUBER *et al.*, 2003) propõem uma técnica de visualização de redes que consiste em eliminar arestas “fracas”. Isso é feito calculando uma métrica nas arestas. Métrica que é baseada somente na topologia da rede. Arestas que possuem um valor abaixo de um limiar dado são excluídas, e as componentes conexas restantes são vistas como grupos. Semelhante ao que vimos em algumas das primeiras técnicas citadas nesse capítulo. Variando esse limiar, os grupos variam. Com um limiar maior, mais arestas são eliminadas e mais grupos são formados, gerando assim uma representação hierárquica da rede. A técnica Ask-Graph View (ABELLO; HAM; KRISHNAN, 2006) é uma técnica de visualização de redes que também utiliza apenas a informação topológica da rede. Ela se utiliza de técnicas de agrupamento para construir uma hierarquia. Bastian *et al.* (RIECK *et al.*, 2017) realiza a análise e visualização de uma rede utilizando homologia persistente para simplificá-la. Com a homologia se detecta as comunidades da rede formadas por cliques do grafo.

Outros mecanismos para auxiliar a visualização de uma rede foram propostos, como o OnionGraph (LIAO *et al.*, 2014) e Pivot Graph (WATTENBERG, 2006), que consideram a topologia e os atributos para definir alguma similaridade entre os nós, permitindo um agrupamento. A Pivot Graph, por exemplo, é uma técnica interativa que une nós que possuem informações de atributos iguais ou que usa a interatividade para criar grupos com as informações desejadas. No entanto, nenhuma dessas técnicas que consideram atributos dos dados associados aos elementos das rede utilizam métodos baseados na NMF para guiar a construção da hierarquia. Além disso, a maioria das técnicas de visualização com redes agrupadas hierarquicamente já assume uma hierarquia como dado de entrada, ou utilizam técnicas prontas para gerar a mesma, realizando apenas a visualização (VEHLOW; BECK; WEISKOPF, 2016; BLANCH; DAUTRICHE; BISSON, 2015). Além disso, as técnicas estudadas que utilizam emparelhamento se baseiam apenas na topologia para realizar a simplificação.

Nossa abordagem nesse trabalho é um passo para preencher essa lacuna entre topologia e conteúdo associado aos elementos da rede, combinando fatoração de matrizes não-negativas, processamento de sinais e emparelhamento de grafos para produzir representações hierárquicas de redes com dados multivariados de forma que as similaridades e padrões sejam preservados. O que difere nessa proposta das técnicas existentes, é que construímos a hierarquia a partir

da combinação de decomposição de matrizes e emparelhamento. Essa combinação utiliza um mecanismo sólido que utiliza as informações dos atributos para definir semelhanças entre nós em cada nível da hierarquia a partir da NMF, com colapsos/agrupamento guiados pelo emparelhamento, permitindo uma representação para fins de visualização. Além disso, utilizamos a teoria de filtragem espectral como ferramenta de pré-processamento da simplificação melhorando essa simplificação e sua visualização.

HIERARCHICAL NON-NEGATIVE MATRIX FACTORIZATION

Seja $\mathcal{G} = (V, E, X)$ uma rede em que V é o conjunto de nós, E é o conjunto de arestas e X é uma matriz em que cada coluna corresponde ao conteúdo (atributos) associado aos nós. A metodologia proposta para simplificar e posteriormente analisar a rede se baseia em construir uma representação hierárquica de $\mathcal{G}$. A hierarquia é construída por meio de colapso de arestas (veja figura 5) agrupando (meta)nós semelhantes e adjacentes gerando metanós que representam pares de (meta)nós colapsados.

Figura 5 – Colapso de arestas.

Algumas questões importantes que devem ser consideradas no processo de simplificação via colapso de arestas são:

1. Como medir a similaridade entre nós adjacentes a fim de decidir quais arestas devem ser colapsadas?
A proposta utiliza a NMF como base para definir a similaridade entre os nós, considerando a informação dada pela matriz X .
2. Como decidir a ordem com que as arestas que serão colapsadas?
Utilizaremos o emparelhamento em grafos para encontrar o conjunto de arestas que devem ser colapsadas, permitindo múltiplos colapsos em cada nível da hierarquia.
3. Como combinar o conteúdo dos (meta)nós colapsados de modo a gerar o conteúdo associado aos metanós resultantes?

O conteúdo dos metanós gerados pelo colapso de arestas é definido pela NMF no início do processo de simplificação e sem a necessidade de calcular a NMF em cada nível da hierarquia.

Como visto na figura 6, os três principais passos de nossa metodologia são utilizados recorrentemente para definir uma representação hierárquica da rede.

Figura 6 – Visão geral da nossa metodologia.

A nossa escolha de agrupar apenas nós adjacentes a cada passo é para não agrupar nós que estejam distantes no sentido da topologia da rede, já que isto poderia acontecer na formulação original da NMF se fosse utilizada apenas a matriz de atributos para realizar a decomposição e agrupar a rede.

Diferentemente de técnicas como single linkage que agrupa um único par de nós a cada passo do processo de simplificação, nossa metodologia agrupa vários pares de nós a cada passo. Por isso foi feita a escolha do emparelhamento para decidir quais arestas serão colapsadas. Além disso, o fato do emparelhamento nos fornecer pares de vértices disjuntos garantimos que cada metanó representa apenas 1 ou 2 (meta)nós do nível abaixo na hierarquia, não mais que isso.

A seguir descreveremos a teoria necessária de NMF e emparelhamento de grafos, as quais fundamentam a metodologia da nossa simplificação proposta.

3.1 Decomposição de matrizes

Fatoração de matrizes é uma ferramenta muito utilizada em análise de dados, com aplicações em modelagem de tópicos (CHOO *et al.*, 2013) e extração de características (FENG *et al.*, 2002). Como nos baseamos na NMF (LEE; SEUNG, 1999) para criar uma representação hierárquica do grafo, vamos revisar rapidamente os conceitos básicos necessários da teoria.

3.1.1 Fatoração de matrizes não-negativas

Dada uma matriz $X = [x_1, \dots, x_n] \in \mathbb{R}^{m \times n}$, onde cada x_i é um vetor coluna, e um inteiro $k \leq \min(m, n)$ (em geral $k \ll \min(m, n)$), a fatoração da matriz não-negativa, NMF (LEE; SEUNG, 1999), busca encontrar duas matrizes $W \in \mathbb{R}^{m \times k}$ e $H \in \mathbb{R}^{k \times n}$ tal que $X \approx W \cdot H$, em que k é um parâmetro definido pelo usuário. Essa decomposição é obtida a partir de matrizes iniciais W_0

e H_0 , aplicando técnicas de otimização que buscam minimizar uma função objetivo sujeita a restrição de termos não negativos. Comumente se utiliza a norma de Frobenius a seguir como função objetivo:

$$\min_{W, H} \|X - WH\|_F^2 \quad (3.1)$$

A principal vantagem de usar a NMF é que somente operações aditivas são feitas no produto $W \cdot H$. Essa restrição de não negatividade é compatível com a noção intuitiva de combinar partes para formar um todo, que é como a NMF aprende uma representação baseada em partes. As colunas de W tendem a serem formadas por pedaços de informação que, quando combinados com H , aproximam X . Desta forma, W leva a uma interpretação intuitiva de X (LEE; SEUNG, 1999).

Um exemplo típico dessa interpretação é a extração de tópicos a partir de documentos textuais. Nessa teoria cada matriz tem seu significado. Neste contexto, as colunas de X são vetores onde cada entrada x_{ij} corresponde ao número de vezes que uma palavra i aparece no documento j . Na decomposição, as colunas de X são expressas em termos da base definida pelas colunas de W , mostrando a relação entre cada palavra e os tópicos. Quanto maior o valor de w_{is} , mais importante a palavra i é para o tópico s .

A j -ésima coluna de H corresponde à j -ésima coluna de X e a s -ésima entrada h_{sj} multiplica a s -ésima coluna de W ; quanto maior for o valor de h_{sj} , mais relevante é o tópico w_s para o documento j . Tradicionalmente o agrupamento de documentos via NMF é realizado utilizando o maior valor de cada coluna de H , agrupando documentos relacionados ao mesmo tópico.

Pode-se realizar uma normalização de tal forma que a soma dos elementos de cada coluna de H seja igual a 1. Dividir cada coluna de H pela soma de suas entradas torna possível interpretar h_{sj} como a probabilidade do documento j ser representado pela coluna w_s , ou seja, a probabilidade de j pertencer ao tópico w_s . Chamaremos h_j de *vetor de probabilidade*, recorrendo a essa noção de probabilidades para definir a semelhança entre dois nós no contexto de redes.

3.1.2 Fatoração de matrizes não-negativas com grafo regularizador

Embora efetiva em vários contextos, a formulação clássica da NMF como descrita na equação 3.1 possui limitações. Quando calculada diretamente na matriz de atributos, ela não leva em consideração relações de vizinhança da rede. A GNMF (Graph regularized Non-negative Matrix Factorization) (CAI *et al.*, 2011) é uma variação da NMF que foi criada para capturar a topologia intrínseca dos dados cuja a versão clássica da NMF não é capaz de representar. A GNMF incorpora na decomposição um regularizador baseado na topologia dos dados.

Para realizar a GNMF em agrupamento de dados é construído uma rede onde cada nó corresponde a uma coluna de X e as arestas ligam os nós mais próximos, ou seja, as colunas de X são interpretadas como pontos em um espaço multidimensional. Essa rede corresponde

ao grafo de vizinhos mais próximos nesse espaço. No nosso caso onde já temos uma rede, não iremos utilizar o grafo de vizinhos mais próximos, mas sim a nossa rede.

A Graph Regularized Non-negative Matrix Factorization (CAI *et al.*, 2011) propõe a seguinte função de custo:

$$\min_{W,H} (\|X - WH\|^2 - \lambda \text{Tr}(WLW^t)) \quad (3.2)$$

sujeita a não negatividade de W e H . No termo de regularização, Tr é o traço de uma matriz, e L é a matriz laplaciana de $\mathcal{G}$. A matriz laplaciana é dada por $L = D - A$, onde A é matriz de adjacência ponderada e D é a matriz diagonal cujas entradas D_{ii} são a soma dos elementos da i -ésima linha de A . O parâmetro $\lambda \geq 0$ (parâmetro suavizador) é o que define a suavidade da nova representação, ou seja, a importância da regularização no processo.

3.1.3 Fatoração de matrizes não-negativas simétricas

Assim como feito na GNMF, a SymNMF (Symmetric Nonnegative Matrix Factorization) (KUANG; PARK; DING, 2012) considera a topologia dos dados. Pode ser mais conveniente utilizar a matriz de adjacência ponderada A de tamanho $n \times n$ ao invés de utilizar a matriz de dados X . Cada entrada na matriz A corresponde à similaridade entre dois nós; toda entrada a_{ij} , diferente de zero, corresponde à similaridade entre os nós i e j de $\mathcal{G}$. A matriz A pode ser definida por alguma medida de similaridade. Uma forma popular é utilizar $A = X^T \cdot X$ que corresponde à medida baseada no produto escalar entre os vetores de atributos de cada nó.

A Symmetric Non-negative Matrix Factorization (KUANG; PARK; DING, 2012) tem como função de custo a seguinte expressão:

$$\min_H \|A - H^t H\|_F^2 \quad (3.3)$$

A SymNMF pode ser facilmente interpretada no contexto de agrupamento. Assim como anteriormente, podemos definir que se h_{ij} é a maior entrada da coluna h_j , o nó j pertence ao grupo i . Uma vantagem da SymNMF é que ela não precisa dos dados multivariados nos nós, precisaria apenas da similaridade entre eles.

3.2 Similaridade dos nós

Considerando a decomposição $X \approx W \cdot H$, cada coluna h_j da matriz H é um vetor com k coordenadas, pois H é uma matriz de tamanho $k \times n$, com entradas h_{ij} correspondendo a probabilidade do vetor de atributos associado ao nó j ser bem representado pelo vetor base w_i .

Dados dois nós j e s , e seu correspondente vetor h_j e h_s em H , dizemos que dois nós são similares se a probabilidade conjunta

$$p(h_j, h_s) = \sum_{i=1}^k h_{ij} h_{is} \quad (3.4)$$

é próxima de 1. Além disso, a probabilidade conjunta pode ser escrita da seguinte forma:

$$\sum_{i=1}^k h_{ij} h_{is} = \langle h_j, h_s \rangle = \|h_j\| \|h_s\| \cos(h_j, h_s), \quad (3.5)$$

onde $\cos(h_j, h_s)$ é o valor do cosseno do ângulo formado pelos vetores h_j e h_s .

Dessa forma a similaridade entre dois nós é dada pelo cosseno do ângulo entre os vetores colunas de H correspondentes, multiplicado pelas normas de tais vetores. Assim, definimos o cosseno entre os vetores de probabilidade h_j e h_s como o peso da aresta ligando os nós j e s . Com o peso definido, colapsar arestas com pesos maiores corresponde a unir os nós mais similares.

As arestas poderiam ser colapsadas uma de cada vez, simplificando a rede sem alterar muito a topologia do grafo a cada passo. Porém, esse procedimento não corresponde com o que se espera de um mecanismo hierárquico, que deve simplificar a rede significativamente em cada nível da hierarquia. Para definirmos o nosso mecanismo hierárquico, iremos colapsar as arestas baseando-se em um emparelhamento em grafo.

3.3 Emparelhamento e agrupamento

Existem várias vantagens de realizar o colapso de arestas via emparelhamento em grafos. Por exemplo, o emparelhamento geralmente encontra um grande número de arestas a serem colapsadas em cada etapa da simplificação. Além disso, o colapso de uma aresta não entra em conflito com os de outras, fazendo com que o procedimento seja computacionalmente mais simples.

Definição 1. Seja $\mathcal{G} = (V, E)$ um grafo. Dizemos que um subconjunto de arestas M de E é um emparelhamento (matching) no grafo $\mathcal{G}$ se não existem duas arestas em M com nós em comum. Um nó de $\mathcal{G}$ sobre o qual uma das arestas de M incide é dito ocupado com respeito a M . Contrariamente, um nó que não incide em nenhuma das arestas de M é dito livre com respeito a M . De forma similar, uma aresta de $\mathcal{G}$ é dita ocupada com respeito a M se ela pertence a M , ou livre com respeito a M , caso contrário. Para cada aresta em M , dizemos que seus dois nós extremos estão emparelhados com respeito a M .

Como descrito acima, o emparelhamento nos fornece um conjunto de arestas dentro da rede que não incidem no mesmo vértice. Desta forma, pode-se colapsar várias arestas emparelhadas de uma vez sem criar cortes na rede ou gerar conflitos.

Um grafo é dito ponderado se possui pesos $w(e)$ em suas arestas. Definimos como custo total do conjunto $X \subseteq E$ a função:

$$c(X) = \sum_{e \in X} w(e) \quad (3.6)$$

É de nosso interesse um emparelhamento que maximize este custo, já que estamos procurando uma simplificação que leve em consideração o peso de cada aresta e colapse primeiro as arestas de maior custo ligando os nós mais similares. Tais emparelhamentos são definidos da seguinte maneira:

Definição 2. Seja $\mathcal{G} = (V, E)$ um grafo ponderado. Seja C a coleção de todos os emparelhamentos em $\mathcal{G}$. Então, um emparelhamento $M \in C$ é de custo máximo (Maximum Weighted Matching - MWM), se, para todo emparelhamento $M' \in C$, temos que $c(M) \geq c(M')$.

Além do emparelhamento de custo máximo, existem outros tipos importantes de emparelhamento:

Definição 3. Seja $\mathcal{G} = (V, E)$ um grafo ponderado. Seja C a coleção de todos os emparelhamentos em $\mathcal{G}$. Então, um emparelhamento $M \in C$ é maximal, se, para todo emparelhamento $M' \in C$, temos que $M' \subseteq M$.

Definição 4. Seja $\mathcal{G} = (V, E)$ um grafo e M um emparelhamento em G . Então, o emparelhamento M é perfeito se todos os vértices de G estão ocupados com respeito a M .

No melhor caso, onde temos um emparelhamento perfeito, conseguiríamos simplificar praticamente metade dos nós da rede em apenas uma etapa de simplificação. Porém, realizar um emparelhamento buscando apenas que ele seja um emparelhamento perfeito seria muito restritivo, pois nem todos os grafos permitem um emparelhamento perfeito.

Desta forma, optamos pelo uso de dois tipos de emparelhamento. Idealmente deveríamos realizar o colapso baseado no emparelhamento de custo máximo (MWM). Entretanto, o MWM tem um custo computacional maior e não é recomendado para redes de larga escala. Para permitir a simplificação de redes maiores, elaboramos uma aproximação gulosa do MWM, o *emparelhamento maximal ordenado* (sorted maximal matching - SMM). O emparelhamento maximal ordenado é computado ordenando as arestas de E em ordem decrescente de peso. O emparelhamento M é construído adicionando arestas a M seguindo a ordenação anterior. Se uma aresta a ser adicionada é adjacente a uma aresta que já está em M , então ela é descartada e a próxima aresta na lista ordenada é avaliada. O processo segue até que todas arestas tenham sido avaliadas.

Não é garantido que o SMM seja um emparelhamento de custo máximo. Porém, ele sempre inclui a aresta de maior peso no emparelhamento, garantindo que os dois nós mais

similares irão ser sempre colapsados em cada passo da construção da hierarquia conforme a figura 7.

Figura 7 – SMM x MWM. O custo do emparelhamento de custo máximo é 15, mas a aresta entre os nós adjacentes mais similares não é seleccionada. O emparelhamento maximal ordenado sempre escolhe as arestas com maior peso a cada etapa, representando os nós mais similares, embora resultando em um custo total reduzido e possivelmente um número menor de arestas colapsadas.

Filtrando arestas de peso baixo. Como mencionado previamente, o ideal seria construir um emparelhamento M com o maior número de arestas possível. Entretanto, existem casos em que arestas correspondendo a nós pouco similares são adicionadas a M . Para evitar esse problema, filtramos arestas de acordo com um limiar δ antes de criar o emparelhamento, isto é, arestas cujo peso é menor que δ não são consideradas no emparelhamento M .

3.4 Hierarquia

Usamos índices sobrescritos para representar os níveis da hierarquia. Por exemplo, j^t corresponde a um (meta)nó no t -ésimo nível da hierarquia, onde $t = 0$ é o nível mais fino, a rede original. Nós denotaremos por $|j^t|$ o número de nós (da rede original) agrupados em j^t durante a construção da hierarquia.

Como a coluna h_j de H corresponde a coluna x_j na matriz de atributos X (veja equação 3.1), definimos a matriz H^t com colunas definidas como:

$$h_j^t = \frac{1}{|j^t|} \sum_{s \in j^t} h_s. \quad (3.7)$$

Em outras palavras, h_j^t é a média das colunas em H correspondendo aos nós pertencentes a j^t . Note que as entradas em h_j^t também podem ser interpretadas como vetores de probabilidade definidos anteriormente.

Dessa forma, todos os metanós j^t de um nível t qualquer da hierarquia, possuem um vetor de probabilidade h_j^t . Cada vetor h_j^t nos fornece a relação entre os vetores base (tópicos) da matriz W e o metanó j^t . Dessa forma, além da representação hierárquica e sua simplificação gerada,

também poderíamos classificar as redes de todos os níveis da hierarquia usando os vetores h_j^t conforme o agrupamento tradicional da NMF explicado na seção 3.1.1.

A metodologia de agrupamento descrita acima não tem a necessidade de recalculando a NMF em cada etapa da simplificação (a cada nível da hierarquia). Portanto, a metodologia proposta torna a construção hierárquica computacionalmente viável e matematicamente sólida.

3.5 Datasets e métricas

A eficácia da nossa metodologia será avaliada usando seis redes com dados multivariados associados aos nós. Esses conjuntos de dados, em conjunto com quatro métricas de qualidade, são usadas para avaliar e comparar nossa abordagem contra métodos da literatura.

3.5.1 Datasets

Artificial. Esta rede tem 30 nós, uniformemente distribuídos em 6 classes; os nós são classificados de acordo com suas classes. Esta rede foi projetada para ter muitas arestas entre nós da mesma classe e poucas arestas entre nós de classes diferentes. Cada nó corresponde a um documento de texto representado por um vetor de frequência de palavras. Nós da mesma classe compartilham as mesmas coordenadas diferente de zero no vetor de frequência de palavras, com valores diferentes. Além disso, vetores de frequência de palavras de diferentes classes são ortogonais, isto é, nós de diferentes classes não possuem similaridade.

A figura 8 descreve uma visualização nó-aresta da rede artificial e três níveis da hierarquia resultante da simplificação feita pela nossa metodologia. As arestas em vermelho são as arestas escolhidas pelo emparelhamento para serem colapsadas em cada etapa. As cores de cada nó correspondem às classes. A metodologia proposta preservou perfeitamente as classes. Esta rede fornece apenas um cenário perfeito para agrupamento baseado em similaridade, servindo como um requisito mínimo para todos os métodos de simplificação, isto é, qualquer método de simplificação deveria ser totalmente preciso neste conjunto de dados.

College Football. Este conjunto de dados, proposto por Girvan e Newman (GIRVAN; NEWMAN, 2002), fornece informações sobre a tabela de jogos da College Football Division, a liga universitária de futebol americano, na temporada do ano 2000. A rede tem 115 nós, cada um representando um time. Os times são divididos em 12 conferências e equipes da mesma conferência tendem a se enfrentar com mais frequência que as equipes de conferências diferentes. Com exceção a uma das conferências, que é composta por equipes independentes que jogam mais frequentemente contra times de outras conferências. A rede é construída criando arestas entre nós (times) que se enfrentaram durante a temporada.

Associamos dados multivariados a cada nó i criando um vetor de característica x_i com dimensão 115. Cada entrada x_{ij} armazena o número de vezes o time i jogou contra o time j .

Figura 8 – Uma visualização de nó-aresta da hierarquia da rede artificial. O tamanho do metanó representa o número de nós do nível zero que foram agrupados nele. As arestas do emparelhamento colapsadas em cada etapa foram coloridas de vermelho. Os nós foram coloridos a partir do nível mais baixo para os níveis acima. Isto é, os nós que foram agrupados em algum nível da hierarquia são nós que são relativamente similares no nível zero da hierarquia.

Figura 9 – A rede College Football e os cinco níveis da representação hierárquica da nossa abordagem (no sentido horário a partir do canto superior esquerdo). A figura inferior esquerda compara os metanós no nível último nível (grandes nós) com aqueles no nível zero / original (nós pequenos). Os nós são coloridos a partir do nível mais baixo para os níveis acima.

Como os times são divididos em 12 conferências, construímos uma estrutura hierárquica de redes onde o último nível da hierarquia possui exatamente 12 nós. A figura 9 mostra a rede original e cinco níveis da hierarquia resultante da simplificação da rede. As cores são definidas com base nos metanós do último nível da hierarquia e os nós de cada nível são coloridos a partir do nível mais baixo para os níveis acima. Nós que foram agrupados em algum nível são coloridos com a mesma cor nos níveis anteriores.

VIS Conference. O dataset da Conferência VIS (ISENBERG *et al.*, 2015) contém informações relacionados a artigos científicos publicados nas conferências IEEE VAST, InfoVis e SciVis,

incluindo o ano de publicação, a conferência, o título, autores, resumo, DOI, dentre outras informações.

Cada autor é representado por um nó e os nós são conectados se o os autores colaboraram um com o outro em pelo menos um artigo. O conteúdo adicional em cada nó é feito a partir dos títulos de todos os artigos de autoria do respectivo autor, sob a forma de uma matriz termo-frequência. Para o nosso estudo consideramos apenas a maior componente conexa dessa rede, que possuía 966 nós e 4.323 arestas.

Movies. O conjunto de dados "The Movies Dataset"(TMDb) é um popular conjunto de dados sobre filmes e programas de TV que pode ser editado pelo usuário. O conjunto de dados contém informações sobre filmes incluindo título, gênero, análises, atores, diretores, equipe e palavras-chave. Cada filme é representado por um nó e dois nós são conectado se os filmes tiverem pelo menos um ator em comum. Consideramos dois conjuntos diferentes de atributos, um em que a matriz X é uma matriz gênero $\times$ filme e outro onde X é dada como palavras-chave $\times$ filme. A rede possui 498 nós e 8076 arestas.

MovieLens. MovieLens é um site de pesquisa da Universidade de Minnesota. No site os usuários podem encontrar filmes que gostam, classificar os filmes e criar perfis. Cada usuário é representado por um nó e os usuários possuem uma aresta entre eles se ambos tiverem classificado o mesmo filme com pontuação máxima (5), resultando em uma rede com 478 nós e 39.991 arestas. O conteúdo adicional associado aos nós é gerado a partir das classificações. Especificamente, X é uma matriz filme $\times$ usuários, em que cada coluna corresponde à nota que um usuário em particular avaliou o filme.

Um resumo dos datasets com o número de nós e arestas pode ser visto na tabela 1.

Tabela 1 – Datasets

Dataset	Nós	Arestas
Artificial Dataset	30	61
College Footbal	115	613
VIS Conference	966	4323
Movies	498	8076
Movielens	478	39991

3.5.2 Métricas de validação

A eficácia da nossa metodologia é avaliada por três diferentes métricas vindas de outros trabalhos. Modularidade, Δ -measure e K-Way Ratio Cut Cost. Essas três métricas quantificam a qualidade do agrupamento de uma rede. Em nossos testes, cada metanó no último nível da hierarquia é considerado como um grupo constituído de nós da rede original, permitindo o uso dessas métricas.

Modularidade. Foi usada por Newman (NEWMAN, 2003) e Wang (WANG *et al.*, 2008) para validar seus métodos de simplificação. Redes com alta modularidade tem muitas conexões dentro dos grupos e poucas conexões entre grupos diferentes. Assumindo que os nós são classificados de acordo com seu grupo, tome e_{ij} como a somatória dos pesos das arestas conectando nós do grupo i com o grupo j e $a_i = \sum_j e_{ij}$. A modularidade Q é definida como:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (3.8)$$

em que $Q = 0$ seria a qualidade de agrupamentos aleatórios e $Q = 1$ é o valor máximo que a modularidade pode ter, quando a rede foi dividida em grupos bem estruturados.

Δ -Measure. A Δ -measure (TIAN; HANKINS; PATEL, 2008; ZHANG; TIAN; PATEL, 2010) é uma métrica que avalia a qualidade do agrupamento medindo as relações entre pares de grupos. Seja $\Phi = \{G_1, G_2, \dots, G_k\}$ uma partição de nós de $\mathcal{G}$ tal que $G_i \cap G_j = \emptyset$, para todo $i \neq j$, e

$$P_{G_j}(G_i) = \{u \mid u \in G_i \text{ e } \exists v \in G_j \text{ tal que } (u, v) \in E\}. \quad (3.9)$$

Tomando $p_{i,j} = (|P_{G_j}(G_i)| + |P_{G_i}(G_j)|) / (|G_i| + |G_j|)$, definimos a Δ -measure como:

$$\Delta(\Phi) = \sum_{G_i, G_j \in \Phi} = (\delta_{G_j}(G_i) + \delta_{G_i}(G_j)) \quad (3.10)$$

em que,

$$\delta_{G_j}(G_i) = \begin{cases} |P_{G_j}(G_i)| & \text{se } p_{i,j} \leq 0.5 \\ |G_i| - |P_{G_j}(G_i)| & \text{caso contrário} \end{cases} \quad (3.11)$$

Utilizando $\Delta = \Delta/k$ temos a média de contribuição dos grupos. Quanto menor o valor da Δ -measure, melhor é o agrupamento.

Essa definição não leva em consideração o peso das arestas, então nós aplicamos a equação (3.10) somente a arestas com peso maior que zero. Pois existem casos que os atributos são esparsos (documentos textuais) e alguns nós adjacentes podem possuir similaridade zero. Chamamos essa versão de Δ -measure modificada. As duas versões serão utilizadas nos resultados.

Métrica K-Way Ratio Cut Cost. A métrica K-Way Ratio Cut Cost (CHAN; SCHLAG; ZIEN, 1994) mede o custo do corte de uma rede gerado por uma agrupamento $\Phi = \{G_1, G_2, \dots, G_k\}$ com k grupos. Seja E_h a soma dos pesos de todas as arestas que tem exatamente um de seus nós no grupo G_h . O custo do corte pode ser definido como:

$$custo(\Phi) = \sum_{h=1}^k \frac{E_h}{|G_h|} \quad (3.12)$$

Assim como na Δ -measure, quanto menor o valor da K-Way Ratio Cut Cost, melhor é o agrupamento.

3.6 Resultados e comparações

Nesta seção, apresentaremos um conjunto abrangente de experimentos e comparações para validar e mostrar a eficácia da metodologia proposta. Dividimos os experimentos em três partes:

1. Validação do método proposto para encontrar o emparelhamento,
2. Validação do mecanismo utilizado para medir a similaridade entre meta(nós), o qual dá origem aos pesos das arestas,
3. Comparação com outras técnicas de simplificação existentes.

Os métodos de simplificação denotados com H maiúsculo (HNMF, HGNMF, HSNMF) na seção de resultados indicam os resultados quando utilizamos a nossa metodologia para simplificar a rede e também o método de decomposição de matriz utilizado durante a simplificação. Métodos de decomposição sem o H correspondem aos métodos tradicionais de agrupamento (NMF, GNMF, SNMF) feitos como a seguir; os nós são agrupados de acordo com as colunas de H definidos inicialmente. Mais especificamente, um nó j irá pertencer ao grupo i se h_{ij} for o maior valor de h_j , como é usual em agrupamentos de dados baseados na NMF (KUANG; PARK; DING, 2012; WANG *et al.*, 2008). Dessa forma, nos métodos tradicionais (NMF, GNMF, SNMF), nenhuma estrutura hierárquica é obtida, e a simplificação é realizada em apenas um passo.

Como dito previamente, os datasets Artificial e o College Football foram simplificados até obtermos 6 e 12 metanós no último nível. Os outros datasets foram simplificados até restar 10 metanóss no último nível.

3.6.1 Emparelhamento de grafos e filtragem de arestas

Para avaliar a eficácia do emparelhamento maximal ordenado (SMM), fizemos uma série de comparações com o emparelhamento de custo máximo (MWM). Mais ainda, avaliamos também o impacto de filtrar e descartar arestas com pesos pequenos antes do processo de calcular o emparelhamento (indicaremos esses casos como SMM_f e MWM_f). Nos casos em que o mecanismo de filtragem foi utilizado, foi empregado um limiar de $\delta = 0,3$; um valor que apresentou bons resultados de uma forma geral nos testes feitos com os datasets citados acima.

As tabelas representadas na figura 10 mostram a qualidade das redes simplificadas para os datasets Artificial (AD), College Football (CF) e VIS (VIS CCN), com e sem filtragem de arestas. As tabelas na figura 10 também mostram a qualidade da simplificação, usando três diferentes decomposições, HNMF, HGNMF e HSNMF. Métodos diferentes de decomposição geram diferentes matrizes H , afetando assim os pesos das arestas e o emparelhamento.

De acordo com a métrica modularidade, o mecanismo SMM proposto resultou em simplificações de maior qualidade para os datasets AC e CF, independentemente do método

		modularity				modified Δ -measure				k-way ratio cut cost			
SMM	VIS	CF	AD	$\delta=0.0$	$\delta=0.3$	CF	AD	$\delta=0.0$	$\delta=0.3$	CF	AD	$\delta=0.0$	$\delta=0.3$
				0.828 / 0.828	0.828 / 0.828			0.000 / 0.000	0.000 / 0.000			0.000 / 0.000	0.000 / 0.000
				0.828 / 0.828	0.828 / 0.828			0.000 / 0.000	0.000 / 0.000			0.000 / 0.000	0.000 / 0.000
				0.828 / 0.828	0.828 / 0.828			0.000 / 0.000	0.000 / 0.000			0.000 / 0.000	0.000 / 0.000
	VIS	CF	AD	0.761 / 0.823	0.700 / 0.784			12.08 / 10.42	11.42 / 9.33			8.298 / 4.844	12.42 / 7.134
				0.827 / 0.848	0.700 / 0.768			11.17 / 10.58	11.58 / 9.67			4.666 / 3.173	11.86 / 7.465
				0.251 / 0.542	0.630 / 0.719			19.35 / 5.00	33.40 / 6.30			7.273 / 4.963	2.678 / 1.763
				0.117 / 0.493	0.575 / 0.722			5.95 / 4.60	28.10 / 5.50			7.022 / 4.292	2.588 / 1.249
MWM	VIS	CF	AD	0.610 / 0.760	0.828 / 0.828			1.750 / 0.667	0.000 / 0.000			2.371 / 0.866	0.000 / 0.000
				0.828 / 0.828	0.828 / 0.828			0.000 / 0.000	0.000 / 0.000			0.000 / 0.000	0.000 / 0.000
				0.632 / 0.689	0.575 / 0.613			12.21 / 11.25	11.17 / 10.08			13.71 / 10.62	17.31 / 15.08
				0.748 / 0.818	0.571 / 0.613			10.50 / 9.33	11.13 / 10.08			10.57 / 8.86	17.12 / 14.98
	VIS	CF	AD	0.584 / 0.662	0.729 / 0.750			76.50 / 31.90	87.75 / 81.40			7.750 / 4.966	8.179 / 6.348
				0.574 / 0.674	0.731 / 0.764			69.85 / 49.40	89.60 / 68.50			7.426 / 4.741	8.485 / 5.579
				0.743 / 0.743	0.743 / 0.743			64.40 / 64.40	64.40 / 64.40			5.188 / 5.188	5.188 / 5.188
				0.743 / 0.743	0.743 / 0.743			64.40 / 64.40	64.40 / 64.40			5.188 / 5.188	5.188 / 5.188

Figura 10 – De cima para baixo, medições de qualidade (mediana / melhores valores) utilizando a função de modularidade, a Δ -measure modificada e a k-way Ratio Cut Cost aplicadas nas redes Artificial, College Football e VIS.

de decomposição. No conjunto de dados VIS, o MWM foi superior, embora o SMM com a SMNF obteve resultados semelhantes. Com relação à Δ -measure e à K-way Ratio Cut, onde os valores menores são melhores, o SMM proposto criou simplificações de qualidade melhor na maioria dos casos, independentemente do método de decomposição. Observe que o mecanismo de filtragem tende a melhorar ainda mais a simplificação, com algumas exceções. Embora não sejam mostrados, resultados semelhantes foram encontrados nos outros conjuntos de dados.

3.6.2 Similaridade dos nós e pesos das arestas

Definir corretamente a similaridade entre os nós é essencial para uma simplificação com alta qualidade. Nossa abordagem baseia-se na NMF e no cosseno do ângulo entre os vetores de probabilidade para realizar tal tarefa. No entanto, há uma infinidade de outras medidas que podem ser usadas para avaliar a similaridade entre os nós, algumas até com menor custo computacional. Para mostrar que a combinação proposta de NMF e cosseno é uma alternativa viável, comparamos contra outros quatro mecanismos comumente empregados para medir a similaridade entre nós. Em todos os casos, a similaridade entre os nós é usada apenas para atribuir pesos as arestas, executando o processo de simplificação de acordo com o pipeline proposto (Figura 6).

Similaridade da vizinhança. Seja N_i o conjunto de nós na vizinhança de um nó i . A similaridade da vizinhança entre dois nós i e j é dada por:

$$\mathcal{N}(i, j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \quad (3.13)$$

Como é uma métrica baseada na topologia (GANSNER; KOREN; NORTH, 2005), ela ignora os atributos nos nós.

Similaridade do tamanho dos grupos. Agrupamentos tipicamente tendem a ter grupos com um número uniforme de elementos. A similaridade do tamanho dos grupos entre dois (meta)nós j_1^i e j_2^i mede o quão uniforme estes (meta)nós são em termos do número de nós que eles contêm

e é dada por:

$$\mathcal{C}(j_1^t, j_2^t) = 1 - \frac{|j_1^t| + |j_2^t|}{\max_{j_i^t, j_k^t \in V^t} \{|j_i^t| + |j_k^t|\}} \quad (3.14)$$

em que $|j^t|$ é o número de nós no (meta)nó j^t e V^t é o conjunto de (meta)nós no t -ésimo nível da hierarquia. Ela se baseia apenas no tamanhos dos grupos gerados até então (GANSNER; KOREN; NORTH, 2005) e não leva em consideração os atributos dos nós.

Similaridade do grau. Seja d_i o grau do nó i , i.e, o número de arestas incidentes ao nó i . A similaridade do grau é dada por

$$\mathcal{D}(i, j) = 1 - \frac{d_i \cdot d_j}{\max_{r, s \in V} \{d_r \cdot d_s\}} \quad (3.15)$$

Ela também é uma métrica baseada em topologia.

Cosseno sem a NMF. Para investigar se o vetor de probabilidade gerado pela NMF realmente traz benefícios no cálculo da similaridade entre os nós, calculamos a similaridade do cosseno diretamente nos vetores de atributos dos nós:

$$\cos(i, j) = \cos(x_i, x_j) \quad (3.16)$$

em que x_i e x_j são os vetores de atributos dos nós i e j respectivamente.

A figura 11 mostra o resultado da utilização dessas métricas para comparar nós e atribuir pesos a arestas. As barras horizontais correspondem aos melhores resultados (vermelho) e melhor mediana (azul) da abordagem proposta (fatoração da matriz + cosseno dos vetores de probabilidade) utilizando o emparelhamento SMM. Também avaliamos todas as métricas com MWM para mostrar que nosso método com o SMM também funciona melhor que outras métricas com o MWM. Nossa métrica superou ou foi semelhante aos melhores resultados apresentados por outras métricas, em todos os casos e conjuntos de dados, incluindo os que não são mostrados na Figura 11.

3.6.3 Comparações com outras técnicas de simplificação

Concluimos os testes com um conjunto de comparações com métodos de simplificação existentes. Mais especificamente, comparamos nossa abordagem com as técnicas de simplificação propostas por Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994), além dos métodos tradicionais da NMF.

As comparações utilizam as métricas de qualidade definidas na seção 3.5, que também foram utilizadas por Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994). Note que os métodos do Newman (NEWMAN, 2003) e do Chan et al. (CHAN; SCHLAG; ZIEN, 1994) são otimizados para suas respectivas métricas. Estamos usando essas métricas de qualidade para avaliar todas as técnicas, fazendo as comparações as mais justas.

Figura 11 – Comparação com outras métricas de similaridade. O SMM representa o emparelhamento maximal ordenado e o MWM é o emparelhamento de custo máximo. SMM_f e MWM_f representam os testes em que o passo da filtragem de arestas foi realizado. As linhas horizontais são as medianas (linha azul) e o melhor resultado (linha vermelha) do nosso método indicados na figura 10. Para o dataset artificial a mediana e o melhor resultados são os mesmos em cada métrica de validação.

Os resultados obtidos nas comparações estão resumidos na figura 12, onde cada linha corresponde a uma métrica de qualidade e cada coluna corresponde a um dataset. Como a nossa abordagem utiliza a NMF com uma condição inicial aleatória, nós rodamos cada configuração 20 vezes, reunindo os resultados das métricas de qualidade nos box plots retratados na figura. Cada coluna na figura 12 é dividida em quatro tiras de cor, uma para cada método de decomposição. Cada barra de cores contém dois box plots, que corresponde ao esquema SMM proposto sem (à esquerda em cada faixa de cor) e com (à direita em cada faixa de cor) o mecanismo de filtragem da aresta baseado no peso. Os boxplots na faixa com cores acastanhadas (faixa de cor da direita em cada coluna) correspondem aos resultados do agrupamento tradicional da NMF; grupos de nós são calculados diretamente das matrizes decompostas em uma única etapa. Como os métodos de Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994) são determinísticos, suas medidas de qualidade são mostradas como linhas horizontais na figura 12.

Para o dataset artificial, composto por nós distribuídos em seis classes, a simplificação foi realizada até seis metanós no último nível. Da mesma forma, os outros métodos foram

Figura 12 – Comparação com outras técnicas de simplificação. Cada coluna (dataset) é dividida em quatro faixas de cor. As três primeiras são do nosso método, um para cada método de decomposição. Cada faixa de cores contém dois box plots, que correspondem ao esquema de emparelhamento SMM proposto sem (à esquerda em cada faixa de cor) e com (à direita em cada faixa de cor) o mecanismo de filtragem de arestas. Faixas acastanhadas (faixa de cor da direita em cada coluna) correspondem aos resultados do agrupamento tradicional da NMF. Os métodos de Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994) são determinísticos e seus resultados são mostrados nas linhas horizontais.

configurados para gerar seis grupos na representação final. Nossa abordagem foi superior ao método de Newman em todas as métricas, com os melhores resultados alcançados pelo HGNMF. O método de Chan apresentou uma tendência para obter melhores resultados neste conjunto de dados, embora, considerando o comportamento médio da nossa abordagem, ambos tiveram

performance bem semelhante.

Em relação às outras comparações, variando as métricas e datasets, a versão da HGNMF da nossa abordagem superou o método de Newman em 19 de 20 comparações. Considerando o desempenho de nossa abordagem independentemente do mecanismo de fatoração e filtragem, obtivemos melhores resultados em todos os datasets quando comparado ao método de Newman. A variante HGMNF da nossa abordagem superou o método de Chan em 11 de 20 comparações. Considerando todas as decomposições possíveis, nossa abordagem foi superior em 16 das 20 comparações.

De um modo geral, a variação HSNMF tem melhor desempenho quando a métrica de modularidade é usada, enquanto o HGNMF tende a ser superior com relação às outras métricas. Comparando a HNMF, HGNMF e HSNMF contra o seu homólogo tradicional, ou seja, NMF, GNMf e SNMF, mostramos claramente que as variantes hierárquicas apresentaram melhores resultados em todos os casos, com exceção do dataset College Football, em que a fatoração não negativa tradicional foi ligeiramente melhor.

3.7 Discussões e Limitações

A metodologia propostas para simplificação de redes apresentou bons resultados em boa parte dos experimentos e validações que executamos, mostrando que nossas escolhas sobre o procedimento de decomposição, definição dos pesos das arestas, emparelhamento, e redefinição dos atributos nos diferentes níveis da hierarquia funcionaram muito bem conjuntamente.

Outro fato interessante é que nossa metodologia superou os métodos de Newman (NEWMAN, 2003) e Chan et al. (CHAN; SCHLAG; ZIEN, 1994) na maioria dos casos, mesmo com o fato de esses dois métodos terem sido concebidos para otimizar algumas das métricas que usamos nas comparações.

A figura 13 ilustra o nosso método para o dataset Open Flight (<<http://openflights.org/data.html>>), que contém informações sobre aeroportos e voos de 3,253 aeroportos e 18,997 rotas. Apesar do dataset ter informações limitadas, considerando somente o número de voos e a distância geográfica, obtivemos alguns resultados interessantes preservando alguns dos principais aeroportos do mundo.

Uma limitação do nosso método é quando a rede possui a topologia de um grafo estrela. Um grafo estrela é uma árvore com um nó interno e todos os nós são folhas; todas as arestas compartilham o nó interno. Com essa configuração, nossa abordagem só pode colapsar uma aresta de cada vez, dificultando uma verdadeira representação hierárquica.

Figura 13 – Visualização de três etapas do dataset Open Flight ao ser simplificado utilizando nosso método.

FILTRAGEM ESPECTRAL EM GRAFOS APLICADA À SIMPLIFICAÇÃO DE REDES

Para dar continuidade aos nossos estudos sobre ferramentas matemáticas que possam melhorar a simplificação de uma rede contendo dados multivariados associados aos seus nós, iremos, nesse capítulo, apresentar a utilização de processamento de sinais em grafos como ferramenta de pré-processamento dos dados e, por conseguinte, seus efeitos na simplificação de redes.

4.1 Processamento de sinais em grafos

O processamento de sinal em grafos (Graph Signal Processing - GSP) (SHUMAN *et al.*, 2013) tem como objetivo desenvolver ferramentas para processamento de sinais definidos em grafos. Esta teoria já vem sendo aplicada na área de visualização de informação (COL *et al.*, 2017; VALDIVIA *et al.*, 2015). Primeiro apresentaremos a teoria matemática da transformada de Fourier em grafos (Graph Fourier Transform - GFT) para então descrevermos o processo de filtragem espectral, que é a base da metodologia proposta.

4.1.1 Transformada de Fourier em grafos

Seja $\mathcal{G} = (V, E)$ um grafo, em que $V = \{v_1, v_2, \dots, v_n\}$ é o conjunto de nós e $E = \{(v_i, v_j), i \neq j\}$ é o conjunto de arestas. A matriz de adjacência ponderada $A = (a_{ij})$ é uma matriz em que a entrada a_{ij} representa o peso da aresta (v_i, v_j) de E ; caso não exista aresta ligando os nós v_i e v_j , então $a_{ij} = 0$.

Um sinal em um grafo é uma função $f : V \rightarrow \mathbb{R}$ definida nos nós de $\mathcal{G}$ que associa um escalar $f(v_i)$ para cada nó $v_i \in V$. O sinal pode ser representado como um vetor em $\mathbb{R}^n$, em que a i -ésima componente do vetor representa o valor do sinal no nó v_i .

A matriz laplaciana não-normalizada do grafo é definida como $L = D - A$, onde D é a matriz diagonal com entradas d_{ii} iguais à soma dos elementos da i -ésima linha de A . A matriz laplaciana é uma matriz real, simétrica e positiva semi-definida. Então a matriz laplaciana L possui um conjunto completo de autovetores ortonormais $\{u_l\}_{l=1,2,\dots,n}$ com autovalores reais não negativos correspondentes $\{\lambda_l\}_{l=1,2,\dots,n}$. O número zero aparece como um dos autovalores e sua multiplicidade é igual ao número de componentes conexas do grafo. Considerando um grafo conexo (somente uma componente conexa), os autovalores podem ser ordenados como $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots \leq \lambda_n$ e zero é o autovalor cujo autovetor correspondente é um vetor constante. O conjunto de autovalores $\sigma = \{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n\}$ é chamado de domínio espectral (SHUMAN *et al.*, 2013).

A transformada de Fourier em um grafo (GFT), $\hat{f} : \sigma \rightarrow \mathbb{R}$, de um sinal f , é uma função definida no domínio espectral, dada pela expansão de f em termos dos autovetores da matriz laplaciana do grafo:

$$\hat{f}(\lambda_l) = \langle u_l, f \rangle = \sum_{j=1}^n u_l(v_j) f(v_j). \quad (4.1)$$

Dada a GFT $\hat{f}$, o sinal f pode ser recuperado através da transformada inversa de Fourier em grafo (inverse graph Fourier transform - iGFT), que é definida como:

$$f(v_i) = \sum_{l=1}^n \hat{f}(\lambda_l) u_l(v_i). \quad (4.2)$$

Se chamarmos de U a matriz (ortogonal) com as colunas dadas pelos autovetores u_l , a GFT e a iGFT podem ser calculadas apenas por multiplicação de matrizes da seguinte forma:

$$\hat{f} = U^T f \quad f = U \hat{f} \quad (4.3)$$

Os autovalores e autovetores da matriz laplaciana do grafo funcionam de forma semelhante às frequências e as funções de base na teoria de Fourier clássica. Como dito, zero é um autovalor cujo autovetor correspondente é um vetor constante. Os autovetores associados aos autovalores λ_l pequenos (mais próximos do autovalor zero, baixa frequência) variam lentamente ao longo do grafo; ou seja, se dois nós são conectados por uma aresta com um peso grande, é provável que os valores do autovetor em tais nós sejam semelhantes. Os autovetores associados a autovalores maiores (alta frequência) oscilam de forma mais abrupta e são mais propensos a terem valores dissimilares em nós adjacentes.

4.1.2 Filtro espectral em grafo

Um filtro espectral em um grafo é uma função $\hat{h} : \sigma \rightarrow \mathbb{R}$ que associa um número escalar $\hat{h}(\lambda_l)$ a cada autovalor $\lambda_l \in \sigma$. Definimos uma filtragem espectral em grafo, de um sinal f como:

$$\tilde{f}(\lambda_l) = \hat{f}(\lambda_l) \hat{h}(\lambda_l), \quad (4.4)$$

onde $\hat{f}$ é a GFT de f e $\hat{h}$ é um filtro espectral em grafo. Usando a iGFT podemos obter a versão filtrada $\tilde{f}$ do sinal f no domínio do grafo:

$$\tilde{f}(v_i) = \sum_{l=1}^n \hat{f}(\lambda_l) \hat{h}(\lambda_l) u_l(v_i). \quad (4.5)$$

Utilizando notação matricial, $\tilde{f}$ pode ser obtida como segue:

$$\tilde{f} = U H U^T f \quad (4.6)$$

onde H é uma matriz diagonal com entradas $\hat{h}(\lambda_1), \dots, \hat{h}(\lambda_n)$.

Iremos estudar o efeito de dois filtros diferentes, o filtro passa-baixa e o filtro de realce. Chamamos de filtro passa-baixa um filtro que não afeta significativamente o conteúdo de baixa frequência e atenua, isto é, reduz a magnitude das altas frequências do sinal. Analogamente, filtros passa-alta atenuam as baixas frequências. O filtro de realce enfatiza as frequências baixas e as frequências altas reduzindo a magnitude das frequências médias.

Seja λ_{cut} um número real tal que $0 < \lambda_{cut} < \lambda_n$. Um filtro passa-baixa $\hat{h}_l$ pode ser definido como:

$$\hat{h}_l(\lambda_l) = \begin{cases} 1, & \text{se } \lambda_l \leq \lambda_{cut}, \\ 0, & \text{se } \lambda_l > \lambda_{cut}. \end{cases} \quad (4.7)$$

Seja $0 < \lambda_{cut_1} < \lambda_{cut_2} < \lambda_n$. Um filtro de realce $\hat{h}_e$ pode ser definido como:

$$\hat{h}_e(\lambda_l) = \begin{cases} 1, & \text{se } \lambda_l \leq \lambda_{cut_1} \text{ ou } \lambda_l \geq \lambda_{cut_2} \\ 0, & \text{caso contrário.} \end{cases} \quad (4.8)$$

O efeito dos dois filtros podem ser vistos na figura 14.

4.2 Filtro espectral em dados multidimensionais e simplificação da rede

Dada uma rede $\mathcal{G}$, buscamos construir uma representação hierárquica de $\mathcal{G}$, onde os metanós de cada nível representam pares de (meta)nós do nível anterior. Assim como foi feito no capítulo anterior, a hierarquia é construída por colapso de arestas que conectam (meta)nós similares. Cada colapso leva a um novo metanó que representa um par de (meta)nós colapsados.

Figura 14 – Na figura central temos um grafo com sinal definido por uma função degrau (o sinal nos nós é 0 ou 1) com alguns outliers (sinal 0,5). Nas figuras laterais, temos o efeito do filtro passa-baixa (esquerda) e do filtro de realce (direita).

Toda a teoria da transformada de Fourier em grafo é definida com base em sinais, ou seja, o grafo possui apenas um único valor escalar para cada nó. Porém, no nosso caso, temos redes com atributos multidimensionais definidos nos nós. Em 4.2.1 iremos definir os conceitos para utilizar a filtragem espectral em dados multidimensionais e depois em 4.2.3 mostraremos como será feita a simplificação da rede utilizando o dado filtrado.

4.2.1 Filtro Spectral em dados multidimensionais

Precisaremos agora adaptar a teoria da transformada de Fourier do grafo para dados multidimensionais.

Seja $\mathcal{G} = (V, E, X)$ uma rede, em que V é o conjunto de nós com $|V| = n$ e E é o conjunto de arestas como de costume. Como no capítulo anterior, X será a matriz de atributos de tamanho $n \times m$, onde cada linha corresponde aos atributos numéricos associados a um nó. Cada entrada x_{ij} em X corresponde ao valor do atributo j no nó v_i . Cada coluna diferente de X representa um dos m atributos, portanto, temos um vetor n -dimensional contendo números reais para cada atributo. Denotaremos por f^j este vetor n -dimensional contendo o j -ésimo atributo (coordenada) de todos os n nós. Lembrando que um sinal em um grafo pode ser representado por um vetor em $\mathbb{R}^n$, $f^j : V \rightarrow \mathbb{R}$ será um sinal definido nos nós de $\mathcal{G}$. Este sinal será representado pelo vetor f^j e o valor do sinal no nó v_i é definido por $f^j(v_i) = x_{ij}$.

Agora podemos usar toda a teoria de processamento de sinais em grafo definida em 4.1 para manipular o sinal f^j definido pelo j -ésimo atributo da rede. Substituindo o sinal f por f^j nas definições, primeiro calculamos a GFT de f^j :

$$\hat{f}^j(\lambda_l) = \langle u_l, f^j \rangle = \sum_{k=1}^n u_l(v_k) f^j(v_k). \quad (4.9)$$

Representaremos por $\hat{X}$ a matriz de tamanho $n \times m$ em que cada coluna j representa a GFT calculada no j -ésimo atributo, $\hat{f}^j$. O valor da GFT, $\hat{f}^j$, para cada autovalor λ_l será

$\hat{X}_{lj} = \hat{f}^j(\lambda_l)$. Dessa forma $\hat{X}$ pode ser calculada através de uma multiplicação de matrizes como segue:

$$\hat{X} = U^T X \quad (4.10)$$

Dado um filtro espectral $\hat{h}$, podemos calcular a versão filtrada, $\tilde{f}^j$, de $\hat{f}^j$. Usando a transformada de Fourier inversa obtemos o sinal filtrado, $\tilde{f}^j$. Fazendo a filtragem dessa forma para todos sinais (atributos), filtraremos toda matriz X e teremos uma nova matriz filtrada $\tilde{X}$ de atributos. Para este propósito, usaremos o mesmo filtro $\hat{h}$ para filtrar cada sinal f^j . Desta forma, podemos calcular todos os sinais filtrados $\tilde{f}^j$ apenas calculando multiplicações entre matrizes. A matriz $\tilde{X}$ contendo todos os sinais filtrados $\tilde{f}^j$ pode ser obtida da seguinte forma:

$$\tilde{X} = U H U^T X \quad (4.11)$$

Com todas essas adaptações e definições, toda a teoria de transformada de Fourier em grafos definida na seção 4.1 foi generalizada para redes com dados multidimensionais definidos nos nós sem restrições quanto ao número de dimensões. Além disso, os sinais filtrados podem ser calculados somente realizando cálculos matriciais. Portanto a teoria não causa grande impacto no custo computacional.

4.2.2 Similaridade dos nós

Após o processamento do sinal da rede, temos que encontrar os nós similares para construir a representação hierárquica da rede colapsando arestas, assim como foi feito em 3.2. Ao invés de usar os dados brutos de X , utilizaremos os novos sinais filtrados de $\tilde{X}$.

Cada coluna $\tilde{f}^j$ da matriz $\tilde{X}$ é um vetor n -dimensional, uma vez que X é de tamanho n por m , com entradas correspondentes ao valor filtrado do sinal f^j associado a cada nó v_i . Do mesmo modo, a linha $\tilde{x}_i$ será um vetor m -dimensional que nos fornece o valor de todos os sinais filtrados nó v_i .

Dados dois nós k e s , e seus respectivos vetores (linha) $\tilde{x}_k$ e $\tilde{x}_s$ em $\tilde{X}$, podemos calcular a similaridade entre os nós usando uma medida de similaridade para calcular a relação entre os dois vetores. A medida de similaridade geralmente é escolhida dependendo de algumas propriedades dos dados. Na HNMF (ver seção 3), a semelhança de cosseno foi escolhida, uma vez que a NMF decompõe os dados em bases e pode ter um resultado muito esparso, principalmente quando aplicado em matrizes de adjacência. Mas como queremos encontrar diferenças (semelhanças) em cada sinal diferente, nos limitarmos a semelhança do cosseno não será a melhor opção. Para outros casos, podemos usar medidas de similaridade baseadas em distâncias (por exemplo, distância euclidiana ou Mahalanobis) ou a norma de Frobenius.

Dada uma aresta (v_i, v_j) , o peso dessa aresta será definido como o valor da similaridade entre os nós v_i e v_j . Portanto, colapsar arestas com pesos grandes corresponde a agrupar os nós semelhantes após a filtragem dos sinais.

Como filtramos os dados, a similaridade de dois nós pode ter um resultado muito diferente em comparação com o resultado que teria se tivéssemos usado os dados brutos. Por exemplo, o filtro passa-baixa considera apenas as baixas frequências. Quando reconstruirmos o sinal pelo iGFT depois de usar o filtro passa-baixa, os nós adjacentes terão atributos que não possuem uma grande variação. Eles serão semelhantes após a filtragem, mesmo se fossem diferentes nos dados brutos. Esse tipo de comportamento é bom para diminuir os efeitos de outliers e ruído (nós com atributos muito diferentes dos seus vizinhos) dentro das redes. Sem a filtragem, existia a chance de um nó com atributos diferentes dos vizinhos (outlier) possivelmente não se colapsar a nenhum outro nó na simplificação, ou a poucos nós criando assim grupos minúsculos. Com a filtragem, os outliers filtrados com o filtro passa-baixa serão mais semelhantes a seus vizinhos e não criarão grupos com apenas um nó no fim da simplificação.

4.2.3 Simplificação hierárquica da rede

Para analisarmos o efeito que a filtragem espectral causa na simplificação de uma rede, utilizaremos a mesma simplificação que fizemos na seção 3 com algumas modificações pertinentes.

Não utilizaremos a NMF para decompor a matriz de atributos nem antes e nem depois da filtragem. A NMF tem a restrição de que os dados não podem ser negativos e queremos testar a filtragem em dados mais gerais. Por isso, faremos a filtragem espectral dos dados logo no início de todo o processo. Depois calcularemos a similaridade como descrevemos em 4.2.2.

Após calcularmos a similaridade entre todos nós adjacentes, precisaremos definir os nós que serão colapsados. Não utilizaremos o emparelhamento de custo máximo nos testes agora, apenas o emparelhamento maximal ordenado que propusemos em 3.3. Ordenaremos as arestas em ordem decrescente de peso e depois checaremos se elas podem fazer parte do emparelhamento de acordo com essa ordem.

Hierarquia. Para adequar-se aos novos conceitos e definir-se as redes em todos os passos da hierarquia precisaremos fazer algumas mudanças que estejam de acordo com a nova notação.

Para representar os níveis da hierarquia, adotaremos índices subscritos. O meta(nó) $(v_j)_t$ corresponde a um (meta)nó no t -ésimo nível da hierarquia. Denotaremos por $|(v_j)_t|$ o número de nós da rede no nível zero que foram agrupados em $(v_j)_t$. Como a linha $\tilde{x}_j$ de $\tilde{X}$ corresponde a linha x_j da matriz de atributos X , definimos a nova matriz filtrada $(\tilde{X})_t$ calculando suas linhas da seguinte forma:

$$(\tilde{x}_j)_t = \frac{1}{|(v_j)_t|} \sum_{s \in (v_j)_t} \tilde{x}_s. \quad (4.12)$$

Em outras palavras, $(\tilde{x}_j)_t$ é a media das linhas de $\tilde{X}$ correspondente aos nós do metanó $(v_j)_t$. As entradas $(\tilde{x}_j)_t$ podem ser interpretadas como o sinal calculado no metanó $(v_j)_t$. Assim

como na HNMF, aqui também não precisaremos fazer a filtragem novamente a cada passo, o que é importante para o custo computacional.

4.3 Resultados e comparações

Para mostrar o efeito da metodologia proposta utilizando dados filtrados, apresentaremos um conjunto de experimentos comparando à teoria de simplificação utilizando os dados brutos. Nosso objetivo é responder às seguintes questões que surgem quando usamos o processamento de sinal da rede para filtrar os dados antes da simplificação:

1. Como a topologia da rede se comportará em cada nível de hierarquia e como isso afetará a simplificação?
2. Como o filtro afetará a simplificação da rede com relação a qualidade do agrupamento?

O efeito da nossa metodologia é avaliado utilizando diferentes redes com dados multivariados associados aos seus elementos, juntamente com as três métricas de qualidade descritas na seção 3.5.

4.3.1 Datasets

Utilizaremos os datasets College Football e o VIS Conference já citados na seção 3.5.1. O College Football será simplificado até encontrarmos 12 metanós. E o VIS Conference será simplificado até chegarmos a 10 metanós. Além destes, utilizaremos os seguintes datasets:

Ecoli. Criado por Kenta Nakai, este dataset contém localizações de proteínas (DHEERU; TANISKIDOU, 2017). Tem 336 instâncias e 7 atributos numéricos. A rede com os dados foi construída usando o grafo com os k vizinhos mais próximos (KNN) derivado do conjunto de dados usando 18 vizinhos. Nós executamos a simplificação até que o último nível tivesse exatamente 12 metanós.

Iris Data Set. Criado por Fisher, o dataset Iris é um banco de dados bem conhecido na literatura de reconhecimento de padrões (DHEERU; TANISKIDOU, 2017). O dataset contém 150 instâncias e 4 atributos numéricos. Cada instância é uma planta íris diferente. Os atributos representam comprimento/largura da sépala/pétala de uma planta íris. A rede com os dados foi construída novamente com os k vizinhos mais próximos utilizando 12 vizinhos. Nós executamos a simplificação até que o último nível tivesse exatamente 4 metanós.

Wine. Estes dados foram criados utilizando os resultados de uma análise química de vinhos cultivados em uma mesma região da Itália (DHEERU; TANISKIDOU, 2017). A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. Os dados possuem 178 instâncias e 13 atributos numéricos, um para cada constituinte. A rede foi

construída com o grafo KNN usando 13 vizinhos. Nós executamos a simplificação até chegarmos a 5 metanós.

Wine Quality. O dataset é composto por variantes do vinho branco português "Vinho Verde" (CORTEZ *et al.*, 2009). Os atributos numéricos representam os resultados dos testes físico-químicos de cada vinho. Por exemplo, acidez, ph e resíduos de açúcar dos vinhos. Este dataset tem 4898 instâncias e 11 atributos numéricos. A rede foi construída com o KNN usando 69 vizinhos. Realizamos a simplificação até termos 10 metanós.

A tabela 2 resume todos datasets acima com o número de nós, atributos e grupos.

Tabela 2 – Datasets

Dataset	Nós	Atributos	Grupos
College Football	115	115	12
Ecoli	336	7	8
Iris	150	4	4
Wine	178	13	5
Wine Quality	4898	11	10
VIS Conference	966	458	10

Nosso objetivo nessas comparações é mostrar o efeito da filtragem dos dados. Como dito anteriormente, todos os conjuntos de dados foram simplificados até que um pequeno número de metanós fossem obtidos no nível final.

4.3.2 Visualização nó-aresta e comportamento do agrupamento

Nesta seção, apresentaremos as diferenças de uma visualização do tipo nó-aresta da simplificação de uma rede com dados multivariados quando usamos os dados brutos em comparação com quando usamos os dados filtrados. Todos os nós são coloridos de acordo com o metanó do último nível da hierarquia, o nível em que interrompemos o método. Se dois nós pertencerem ao mesmo metanó no último nível, eles serão coloridos com a mesma cor em todos os níveis. O tamanho de cada metanó corresponde ao número de nós agrupados para formá-lo.

A figura 15 representa a simplificação da rede do dataset ecoli com os dados brutos (acima) e com os dados filtrados com um filtro passa-baixa (abaixo), respectivamente. A primeira diferença que podemos notar é o tamanho dos metanós no último nível. Como dissemos anteriormente, o filtro passa-baixa transforma os dados e deixa-os com uma distribuição mais homogênea ao longo do grafo. Observamos que quando a rede com os dados brutos foi simplificada, quase todos os nós no topo da visualização acabaram no mesmo grupo, com alguns nós próximos a eles formando pequenos grupos.

Esse tipo de comportamento não aconteceu na simplificação com o filtro passa-baixa, onde o grande grupo foi dividido em dois grupos com quase o mesmo tamanho. Desta ma-

Figura 15 – Etapas da simplificação da rede do dataset ecoli usando os dados brutos (parte superior) e os dados filtrados com o filtro passa-baixa de Fourier (parte inferior).

Figura 16 – Etapas da simplificação da rede do dataset Wine usando os dados brutos (parte superior) e os dados filtrados com o filtro passa-baixa de Fourier (parte inferior).

neira podemos observar a rede de uma perspectiva diferente, com grupos de tamanhos mais homogêneos.

O mesmo comportamento pode ser observado na visualização da simplificação da rede do dataset Wine (figura 16). Quando os dados brutos foram utilizados, um grupo muito pequeno aparece no último nível da hierarquia. Utilizando os dados filtrados com o filtro passa-baixa, a simplificação final apresentou metanós com tamanhos mais homogêneos.

Esse tipo de comportamento é interessante se levarmos em consideração o número de grupos desejado. No caso da simplificação do dataset ecoli com os dados brutos, é como se os dados tivessem sido divididos em seis grupos, sendo um dos seis muito pequeno, e não oito como desejado.

4.3.3 Simplificação da rede

A partir do momento que os atributos foram transformados pela filtragem espectral tornando nós que originalmente não eram similares mais parecidos uns com os outros, é esperado que a simplificação tenha um resultado pior quando fosse avaliada por métricas de qualidade. Mas não foi o que aconteceu. A simplificação com o dado filtrado apresentou resultados melhores ou iguais a simplificação com os dados brutos.

Na figura 17 (parte superior) temos os resultados das simplificações avaliadas pela métrica de modularidade. Lembrando que na modularidade, quanto maior o resultado, melhor é a simplificação. Comparando, vimos que usando os dados filtrados obtivemos melhores resultados nos datasets *ecoli*, *íris* e *wine*. No dataset *íris*, os valores são praticamente os mesmos. Para os datasets *wine quality* e *college football*, os resultados apresentaram uma diferença muito pequena. O dataset *college football* tem atributos esparsos, por isso era esperado que os resultados não fossem muito bons.

Usando a Δ -measure (centro da Figura 17) os dados filtrados apresentaram os melhores resultados nos dataset *ecoli*, *íris* e *wine* novamente. Nesta métrica, a diferença é mais perceptível. Tanto para o filtro passa-baixa quanto para o filtro de realce. A diferença neste caso é que o filtro passa-baixa apresentou um resultado melhor no dataset *wine quality* e também no dataset *college football*, mesmo sendo uma diferença muito pequena.

Novamente, os resultados com os dados filtrados foram melhores nos datasets *ecoli*, *íris* e *wine* se avaliados com a métrica *k*-Way Ratio Cut Cost (parte inferior da figura 17). Agora os dois filtros apresentaram melhores resultados no dataset *wine quality*. Porém, eles praticamente empataram no dataset *college football*.

4.4 Discussões e limitações

A metodologia proposta obteve bons resultados na maioria dos experimentos e avaliações consideradas, mostrando que nossa escolha de como filtrar os dados funcionou bem com a metodologia de simplificação de redes. Isso não quer dizer que é sempre melhor usar o filtro. Uma vez que ao transformar os dados sempre se perde algo. Além do usuário poder querer visualizar os dados brutos. Mas os dados filtrados nos fornecem outra maneira de analisar e visualizar a rede e o conteúdo adicional. Os dois dados, o filtrado e o bruto, poderiam ser utilizados juntos, por exemplo. Mas como visto, a filtragem espectral tem a sua importância e em determinados contextos ela melhora a simplificação da rede.

Assim como citado na seção 3.7, mesmo com a filtragem ainda temos problemas com grafos do tipo estrela, pois a filtragem é feita nos atributos e não altera a topologia do grafo. Para redes que possuem um grande grupo central e pequenos grupos isolados como a rede VIS Conference (figura 18), a metodologia proposta para simplificação pode não funcionar

Figura 17 – Comparação da simplificação usando os dados brutos e os dados filtrados utilizando a métrica de modularidade (topo), Δ -Measure (centro) e a métrica k -Way Ratio Cut Cost (abaixo) para cada rede. A barra azul representa os resultados dos dados brutos. A barra verde representa os resultados dos dados filtrados com o filtro passa-baixa. A barra amarela representa os resultados do filtro de realce. O melhor resultado na métrica de modularidade é o maior número, enquanto nas métricas Δ -Measure e k -Way Ratio Cut Cost, os melhores resultados são os menores.

perfeitamente, já que dois pequenos grupos ainda sobraram no último nível quando simplificamos a rede até que tenhamos dez metanós. Mas comparando nosso resultado utilizando a metodologia da HNMF, a HNMF sem o filtro espectral apresentou um número maior de pequenos metanós isolados. Basicamente, dividindo o dataset em apenas quatro grupos. Usando nossa metodologia para filtrar os atributos e depois utilizar a HNMF, podemos identificar oito grupos bem definidos.

Figura 18 – Comparação da metodologia do Filtro Espectral do Grafo aliada a HNMF com a HNMF sem o filtro utilizando a rede VIS Conference.

VIZLATTES

Nesse capítulo apresentaremos uma aplicação de NMF em visualização de redes de colaboração científica. Utilizamos a NMF e os caminhos da rede para detectar comunidades e tópicos de interesse. Como pesquisadores e autores podem pesquisar em mais de uma área científica, torna-se interessante encontrar um método que permitisse que os autores (nós) pudessem estar em mais de um grupo diferente na simplificação.

Aplicamos nosso método em redes de colaboração científica geradas através de informações extraídas da base de dados do Lattes, que possui o currículo acadêmico completo de pesquisadores no Brasil.

Analisando essas redes buscamos responder as seguintes perguntas:

- Quais são as comunidades mais importantes de uma determinada rede?
- Quais são os principais tópicos de pesquisa?
- É possível identificar autores de comunidades diferentes pesquisando sobre o mesmo assunto?

Na busca por encontrar respostas para essas perguntas elaboramos o Vizlattes, um novo método para visualizar e analisar redes de colaboração científica a partir da base de dados Lattes. O VizLattes utiliza os caminhos definidos na rede e a fatoração de matrizes não negativas para identificar tópicos similares, preservando a capacidade de encontrar comunidades relacionadas. Dados os títulos dos trabalhos de cada autor, criamos uma coleção de documentos que a partir da qual extraímos padrões da rede. A NMF nos permite encontrar padrões nos documentos de texto e é usada para encontrar e agrupar tópicos semelhantes dentro da rede. Nossa proposta também utiliza os caminhos mínimos da rede. O agrupamento é feito no que chamamos de caminhos relevantes da rede, em vez de feito nos nós. Um nó pertence a mais de um caminho relevante. Então este nó pode aparecer em mais de um grupo após a simplificação. Nosso método baseia-se

Figura 19 – Pipeline do Vizlattes.

em uma metáfora visual que revela claramente as comunidades, seus tópicos subjacentes e autores mais proeminentes dentro de cada comunidade.

5.1 Metodologia

Nessa seção vamos apresentar a metodologia baseada na NMF para analisar e visualizar as redes de colaboração científica. O pipeline da metodologia pode ser visto na figura 19.

5.1.1 Redes de colaboração científica

Seja $\mathcal{G} = (V, E, C)$ uma rede, sendo V o conjunto de nós, E o conjunto de arestas e C um conjunto de documentos de texto, em que cada documento carrega informações vinculadas a um nó em V .

Dada a rede $\mathcal{G}$, realizamos o pré-processamento dos documentos de texto para criar uma matriz de termo-frequência da coleção, obtendo assim uma matriz X de tamanho n (número de nós) por m (número de palavras). Cada linha representa um documento (nó) e cada coluna representa uma palavra. O elemento a_{ij} da matriz X informa o número de vezes que a palavra j aparece no documento (nó) i . Cada documento é descrito por um vetor x_i de X .

Por exemplo, em nossos modelos de rede de colaboração do Lattes, cada nó representa um pesquisador e está associado a um texto formado pelos títulos de todos os seus trabalhos publicados. Uma aresta entre dois nós indica que os respectivos pesquisadores colaboraram no mesmo trabalho.

Utilizamos o sistema scriptLattes ([MENA-CHALCO; JUNIOR; MARCONDES, 2009](#)), que nos fornece o nome de todos os trabalhos de um pesquisador, para elaborar a rede de colaboração com todas as informações necessárias associadas aos autores e suas publicações.

5.1.2 Caminhos relevantes

As arestas também possuem um documento de texto gerado pelos seus dois nós, isto é, os títulos dos artigos dos dois autores que estão ligados pela aresta. Assim, um caminho da rede

conectando múltiplos pesquisadores também tem uma informação textual associada a ele, obtida dos seus nós.

Cada caminho na rede representa a ideia de áreas de pesquisa entre os autores. Uma vez que autores que colaboraram entre si possivelmente trabalham em áreas de pesquisa similares. Nesse contexto, podemos definir o conceito de caminho relevante como o caminho mínimo que conecta pesquisadores em áreas de pesquisa similares.

Para determinar os caminhos mínimos precisamos calcular a distância entre dois nós. Utilizamos a medida de distância gerada a partir da similaridade do cosseno

$$d_{\cos}(x_i, x_j) = 1 - \cos(x_i, x_j)$$

para determinar a distância entre dois autores. Todos os caminhos mínimos são determinados pelo algoritmo de Floyd-Warshall (FLOYD, 1962). Depois filtramos os caminhos mínimos baseados no seu tamanho. Os caminhos relevantes serão os caminhos mínimos de tamanho menor ou igual a uma constante k , escolhida pelo usuário.

Dados os caminhos relevantes, criamos uma nova matriz de termo-frequência P de tamanho $m \times p$, onde p é o número de caminhos relevantes restantes após a filtragem com base no tamanho. Cada termo p_{ij} registra a frequência que a palavra i aparece nos documentos associados ao caminho j (ver figura 20).

Figura 20 – Representação dos caminhos relevantes através da matriz de termo-frequência.

Na aplicação da rede de colaboração, nossa suposição é que a decomposição pela NMF dos caminhos revelará grupos de tópicos de pesquisa semelhantes ou relacionados.

Uma função de transferência é aplicada para equilibrar os pesos da matriz de frequência de palavras P e medir a importância da palavra para um caminho. Utilizamos essa função, pois a palavra pode aparecer muitas vezes em um caminho, mas aparecer em somente um ou dois nós do caminho. Então ela não é realmente tão relevante para o caminho como um todo. Além disso,

podem existir outros caminhos nos quais essa palavra aparece com mais frequência. Então ela é mais relevante para esses caminhos em que aparecem mais. A função de transferência é dada por uma função $f : [0, 1] \times [0, 1] \rightarrow [0, 1]$ criada usando uma superfície de Bezier (figura 21) com os seguintes pontos de controle:

$$P_{1,1} = [0, 0, 0]; \quad P_{1,2} = [0, 0.5, 0.25]; \quad P_{1,3} = [0, 1, 0.5];$$

$$P_{2,1} = [0.5, 0, 0.4]; \quad P_{2,2} = [0.5, 0.5, 0.7]; \quad P_{2,3} = [0.5, 1, 0.8];$$

$$P_{3,1} = [1, 0, 0.6]; \quad P_{3,2} = [1, 0.5, 0.9]; \quad P_{3,3} = [1, 1, 1];$$

Dado um caminho c e uma palavra t , as entradas x e y para a função de transferência $f(x, y)$ são definidas da seguinte forma:

- $x = N_p/N_t$ em que N_p é a frequência da palavra t no caminho c e N_t é a frequência mais alta de t em um caminho qualquer.
- $y = N_{nt}/N_{nc}$, onde N_{nt} é o número de nós no caminho c que inclui a palavra t e N_{nc} é o número total de nós em c .

Figura 21 – Função de transferência de peso da frequência das palavras.

Ou seja, se a palavra t aparece mais frequentemente em um caminho c_1 que no caminho c , ela não será tão importante para o caminho c . Da mesma forma, se uma palavra só aparece em um nó do caminho, ela também não será tão importante para o caminho. Ela será mais importante se aparecer em todos os nós do caminho.

Uma nova matriz Y com tamanho $m \times p$ é obtida, onde cada termo y_{ij} dá a saída da função de transferência para uma palavra i e um caminho j .

5.1.3 NMF e métricas

Após encontrar a matriz Y que nos dá a relevância de cada palavra para cada caminho, utilizamos a NMF para decompor a matriz Y nas matrizes W e H como na seção 3.1.1. Assim,

W revela os tópicos extraídos dos dados e H nos dá a relação entre os caminhos e esses tópicos.

Usando as duas matrizes W e H , criamos uma metáfora visual capaz de mostrar uma visão geral da relevância dos tópicos e comunidades identificadas na rede de colaboração. Para gerar essa metáfora visual iremos precisar primeiro interpretar a NMF e gerar algumas métricas.

A NMF é uma ferramenta para a descoberta automatizada de tópicos em grandes coleções de documentos. Basicamente, identifica e produz um conjunto de tópicos nos dados, bem como uma representação de instâncias de dados em termos dos tópicos como já explicado na seção 3.1.1. Também podemos interpretar cada linha na matriz H como base para a distribuição de caminhos que contribuem para um determinado tópico. A partir desta informação podemos extrair a importância relativa dos caminhos para cada tópico, onde cada coluna de H representa um caminho.

Com base nesse conceito, empregamos o agrupamento k -means nas colunas de H para agrupar os caminhos nas redes associadas aos tópicos extraídos, resultando em um agrupamento da rede em um conjunto de subredes C que chamaremos de comunidades. Caminhos que possuam uma distribuição nos tópicos parecidas foram agrupados na mesma comunidade.

Embora cada comunidade da rede esteja associada a um tópico importante, a NMF oferece suporte a uma representação fuzzy. A fim de capturar e ilustrar melhor os múltiplos tópicos em uma comunidade, calculamos a relevância de cada tópico l em uma comunidade de rede c da seguinte forma:

$$r_{lc} = \sum_{j \in \mathcal{P}_c} h_{lj}, l = 1, \dots, k \quad (5.1)$$

em que $\mathcal{P}_c$ são os índices dos caminhos dentro de uma comunidade $c \in C$ (ver figura 22).

Figura 22 – Elementos que serão somados para calcular a relevância do tópico l em uma comunidade c .

Para avaliar a importância relativa dos autores em uma comunidade específica da rede, consideramos os caminhos nessa comunidade que incluem esse autor. A relevância de cada autor

i em uma comunidade de rede c relativa a um tópico l é dada por:

$$r_{lc_i} = \sum_{j \in \mathcal{P}_{c_i}} h_{lj}, l = 1, \dots, k \quad (5.2)$$

em que $\mathcal{P}_{c_i}$ são os índices dos caminhos que incluem o autor i dentro de uma comunidade, $c \in C$, da rede (ver figura 23).

Figura 23 – Elementos que serão somados para calcular a relevância do um autor i em uma comunidade c relativa a um tópico l .

Dadas essas medidas, o aplicativo VizLattes pode mostrar uma visão geral resumida de uma determinada rede de colaboração científica, em termos das comunidades de coautoria formadas, seus tópicos descritivos e a relativa expressividade de comunidades de tópicos e autores distintos na rede de colaboração geral.

Apesar de estarmos usando o termo “importância” para as métricas relacionando o autor no tópico, não estamos querendo dizer que tal pessoa é o melhor pesquisador e nem estamos fazendo uma lista classificando os mesmos. O que ela nos fornece é uma similaridade entre autor e tópico. Nos diz que os caminhos referentes a esse autor são melhor expressos por determinado tópico.

5.1.4 VizLattes

Agora vamos apresentar nossa ferramenta visual para explorar os resultados da decomposição da NMF.

A metodologia utilizando NMF e as métricas foi capaz de identificar algumas comunidades e tópicos na rede original de colaboração. A nossa metáfora de visualização, VizLattes, representa cada comunidade como uma barra vertical (figura 24). O tamanho da barra representa proporcionalmente o tamanho de sua comunidade. O tamanho de uma comunidade é medido pelo número de pesquisadores.

Cada barra é dividida em regiões retangulares menores de cores diferentes, representando a distribuição de tópicos dentro da comunidade. Cada tópico é identificado por uma cor diferente e sua importância relativa para a comunidade é representada pela área total de sua região retangular.

Figura 24 – Construção do método de visualização. Cada barra representa uma comunidade. As cores representam diferentes tópicos.

As importâncias relativas dos tópicos dentro da comunidade são calculadas usando as métricas descritas anteriormente.

Por exemplo, na figura 25 a sexta comunidade (olhando da esquerda para a direita) possui o maior número de autores entre as comunidades encontradas. O tópico principal para a comunidade é o tópico de palavras representado pela cor laranja. Ou seja, os seus autores estão em maior parte pesquisando sobre esse tópico. A primeira comunidade tem o tópico representado pela cor verde clara como seu tópico principal. Da mesma forma, a visualização nos mostra o principal tópico de cada comunidade de autores.

Além disso, o Vizlattes permite que o usuário interaja com a visualização (ver figura 25). A interação com a visualização é feita passando o mouse sobre cada barra e suas regiões retangulares, o que resulta em três tipos de informações mostradas:

- Os autores mais representativos em uma comunidade para o tópico dado pela área colorida sob o mouse são mostrados em uma caixa a direita da barra vertical.
- As palavras mais representativas do tópico sob o mouse são mostradas no lado esquerdo da visualização.
- Os autores mais representativos para o tópico sob o mouse, considerando a rede como um todo, são mostrados na área inferior da visualização.

Figura 25 – Método de visualização do Vizlattes.

5.2 Experimentos

Para validar o método proposto, realizamos três experimentos em dados, descritos a seguir:

Primeiro experimento

Para uma validação inicial do método e sua implementação, criamos um conjunto de dados que consiste em seis tópicos nos quais cada tópico é caracterizado por dez palavras. Criamos documentos textuais concatenando as palavras características de cada tópico, e, em seguida, fizemos uma associação específica desses seis tópicos com cinco nós.

As arestas entre os nós foram estabelecidos para que os nós do mesmo tópico sejam densamente ligados, enquanto nós associados a tópicos distintos estão esparsamente ligados. Esperamos que a aplicação do nosso método a essa rede artificial revele adequadamente os seis tópicos com suas respectivas palavras e nós proeminentes.

Como o número de tópicos é conhecido, a NMF foi configurada para agrupar os dados em seis tópicos. Como esperado, a VizLattes foi capaz de transmitir as palavras mais importantes em cada tópico e os nós mais representativos associados a cada tópico.

Segundo experimento

O segundo experimento foi novamente conduzido em um conjunto de dados artificial com os mesmos seis tópicos do primeiro experimento em que cada tópico é caracterizado por dez palavras.

A diferença é a maneira como construímos os documentos textuais de cada nó. Dado um nó, queremos que esse nó pertença a um tópico. O documento textual associado deste nó foi criado contendo mais palavras desse tópico e algumas poucas palavras de outros tópicos.

Novamente, bons resultados foram obtidos e o VizLattes foi capaz de capturar as comunidades e palavras relevantes para cada tópico.

Terceiro experimento

O terceiro experimento foi realizado em um conjunto de dados real, usando o sistema scriptLattes para gerar uma rede de colaboração de docentes do ICMC-USP com artigos publicados até 2015. Os resultados são mostrados na Tabela 3 e na figura 25.

Tabela 3 – Esta tabela resume os tópicos encontrados pela decomposição e respectivos autores proeminentes extraídos com o método proposto da rede de colaboração do corpo docente do ICMC em 2015.

Tópicos								
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
trees	bayesian	assisted	mutation	schemes	evolving	web	maps	safety-critical
classifica- tion	estimators	crossings	mutants	3d	retrieval	content	notes	aerial
multiclas- ses	regression	multidim- ensional	softwares	flows	indexing	materials	topology	unmanned
multi- objective	surviving	biosensing	coverage	free	caracteri- zação	accessible	genericity	vehicles
radial	measuring	spectros- copy	programs	smoothing	querying	interactive	manifolds	embeddings
algoritmos	poisson	highly	contexto	surfaces	university	videos	planes	lines
Autores mais relevantes em cada tópico								
A.C.P.L.F. Carvalho	J.A. Achcar	M.C.F. Oliveira	J.C. Maldonado	A. Cas- telo Filho	A.J.M. Traina	R.P.M. Fortes	M.A.S. Ruas	K.R.L.J.C. Branco
R.F. Mello	F. Louzada	S.M. Aluísio	M.C.F. Oliveira	J.A. Cu- minato	C. Traina Jr.	R. Goularte	C. Biasi	P.C. Masiero
A.C.B. Delbem	F.A. Rodrigues	L.G. Nolato	E.F. Barbosa	L.G. Nolato	M.C.F. Oliveira	R.T.V. Braga	E.M. Zuffi	M.C.F. Oliveira
F.M.B. Toledo	M.C. Andrade Filho	P.C. Masiero	A.S. Simão	A.N. Carvalho	J.E.S. Batista Neto	R.F. Mello	R.N.A. Santos	R.T.V. Braga

5.3 Discussões e limitações

Apesar da ideia ser bem interessante, ela possui um grande problema com relação ao custo computacional dela. A partir do momento que deixamos de trabalhar com nós e arestas para trabalhar com caminhos, tem-se um grande aumento no número de dados. Já que o número de caminhos é bem maior que o número de arestas. O que torna o Vizlattes inviável para redes de larga escala.

CONCLUSÃO

Esta tese apresentou estudos sobre a utilização de técnicas de fatoração de matrizes não negativas, emparelhamento de grafos e processamento de sinais em grafos para auxiliar a simplificação de uma rede com dados multivariados.

Utilizando a fatoração de matrizes não-negativas e emparelhamento de grafos apresentamos um novo método, HNMF, para criar um agrupamento hierárquico de uma rede com informações multivariadas definidas nos nós. Com as duas teorias, a HNMF apresenta uma base matemática sólida e bem definida que nos permite utilizar análise de dados multivariados combinada com teoria dos grafos. A HNMF teve bons resultados em todas comparações feitas, seja com outras métricas de similaridade ou com outras técnicas de agrupamento de rede e com os agrupamentos dados pela formulação clássica da NMF. Demonstrando então, ser um bom método para simplificar uma rede com foco em visualização.

A teoria de filtragem espectral de grafos se mostrou uma importante ferramenta como uma forma de pré-processamento dos dados vinculados aos nós de uma rede. A metodologia proposta de filtragem dos dados multidimensionais apresentou bons resultados quando utilizados para simplificar a rede. Tanto na questão do comportamento do agrupamento gerado, quanto na qualidade da simplificação. Dessa forma, a filtragem nos permitiu uma outra maneira de analisar a rede e sua simplificação.

REFERÊNCIAS

- ABELLO, J.; HAM, F. van; KRISHNAN, N. Ask-graphview: A large scale graph visualization system. **IEEE TVCG**, 2006. Citado na página 32.
- AUBER, D.; CHIRICOTA, Y.; JOURDAN, F.; MELANÇON, G. Multiscale visualization of small world networks. In: **INFOVIS**. [S.l.: s.n.], 2003. Citado na página 32.
- BLANCH, R.; DAUTRICHE, R.; BISSON, G. Dendrogramix: A hybrid tree-matrix visualization technique to support interactive exploration of dendrograms. In: **2015 IEEE PacificVIS**. [S.l.: s.n.], 2015. p. 31–38. ISSN 2165-8765. Citado na página 32.
- BONDY, J.-A.; MURTY, U. S. R. **Graph theory**. New York, London: Springer, 2007. (Graduate texts in mathematics). ISBN 978-1-8462-8969-9. Citado na página 24.
- BOTE, V. P. G.; ZAPICO-ALONSO, F.; ESPINOSA-CALVO, M. E.; CRISÓSTOMO, R. G.; ANEGÓN, F. de M. Binary pathfinder: An improvement to the pathfinder algorithm. **Inf. Process. Manage.**, v. 42, n. 6, p. 1484–1490, 2006. Citado na página 27.
- CAI, D.; HE, X.; HAN, J.; HUANG, T. S. Graph regularized nonnegative matrix factorization for data representation. **IEEE PAMI**, IEEE Computer Society, Los Alamitos, CA, USA, 2011. Citado nas páginas 24, 31, 37 e 38.
- CHAN, P. K.; SCHLAG, M. D. F.; ZIEN, J. Y. Spectral k-way ratio-cut partitioning and clustering. **IEEE Trans. on CAD of Integrated Circuits and Systems**, v. 13, n. 9, p. 1088–1096, 1994. Citado nas páginas 16, 30, 45, 48, 49, 50 e 51.
- CHOO, J.; LEE, C.; REDDY, C. K.; PARK, H. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. **IEEE transactions on visualization and computer graphics**, IEEE, v. 19, n. 12, p. 1992–2001, 2013. Citado na página 36.
- COL, A. D.; VALDIVIA, P.; PETRONETTO, F.; DIAS, F.; SILVA, C. T.; NONATO, L. G. Wavelet-based visual analysis for data exploration. **Computing in Science and Engineering**, v. 19, n. 5, p. 85–91, 2017. Citado na página 53.
- CORREA, C. D.; CRNOVRSANIN, T.; MA, K.-L. Visual reasoning about social networks using centrality sensitivity. **IEEE TVCG**, 2012. Citado na página 28.
- CORREA, C. D.; CRNOVRSANIN, T.; MA, K. liu; KEETON, K. **The Derivatives of Centrality and their Applications in Visualizing Social Networks**. [S.l.], 2009. Citado nas páginas 15, 28 e 29.
- CORTEZ, P.; CERDEIRA, A.; ALMEIDA, F.; MATOS, T.; REIS, J. Modeling wine preferences by data mining from physicochemical properties. **Decis. Support Syst.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 47, n. 4, p. 547–553, nov. 2009. ISSN 0167-9236. Citado na página 60.

DEARHOLT, D.; SCHVANEVELDT, R. Properties of pathfinder networks. In: SCHVANEVELDT, R. (Ed.). **Pathfinder associative networks: Studies in knowledge organization**. Norwood, NJ: Ablex Publishing Corporation, 1990. p. 1–30. Citado na página 27.

DHEERU, D.; TANISKIDOU, E. K. **UCI Machine Learning Repository**. 2017. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 59.

DIAS, M. D.; MANSOUR, M. R.; DIAS, F.; PETRONETTO, F.; SILVA, C. T.; NONATO, L. G. A hierarchical network simplification via non-negative matrix factorization. In: **2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2017. v. 00, p. 119–126. ISSN 2377-5416. Citado nas páginas 23, 24 e 25.

DIAS, M. D.; VALDIVIA, P.; PETRONETTO, F.; NONATO, L. G. Graph spectral filtering for network simplification. In: **31th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2018. v. 00. Citado na página 25.

DIAS, M. D. S. S.; MANSOUR, M.; ROMANETTO, L.; OLIVEIRA, M. C.; NONATO, L. Vizlattes: a tool for relevance analysis from scientific co-authorship networks. In: **Workshop on visual analytics, information visualization and scientific visualization**. [S.l.: s.n.], 2015. Citado na página 25.

DINKLA, K.; WESTENBERG, M. A.; WIJK, J. van. Compressed adjacency matrices: Untangling gene regulatory networks. **IEEE TVCG**, 2012. Citado na página 29.

DWYER, T.; MEARS, C.; MORGAN, K.; NIVEN, T.; MARRIOTT, K.; WALLACE, M. Improved optimal and approximate power graph compression for clearer visualisation of dense graphs. **2014 IEEE PacificVIS**, IEEE Computer Society, Los Alamitos, CA, USA, 2014. Citado na página 29.

DWYER, T.; RICHE, N. H.; MARRIOTT, K.; MEARS, C. Edge compression techniques for visualization of dense directed graphs. **IEEE TVCG**, v. 19, n. 12, p. 2596–2605, 2013. Citado nas páginas 15 e 29.

FENG, T.; LI, S. Z.; SHUM, H.-Y.; ZHANG, H. Local non-negative matrix factorization as a visual representation. In: IEEE. **Development and Learning, 2002. Proceedings. The 2nd International Conference on**. [S.l.], 2002. p. 178–183. Citado na página 36.

FLOYD, R. W. Algorithm 97: Shortest path. **Commun. ACM**, ACM, New York, NY, USA, v. 5, n. 6, p. 345–, jun. 1962. ISSN 0001-0782. Citado na página 67.

GANSNER, E. R.; KOREN, Y.; NORTH, S. C. Topological fisheye views for visualizing large graphs. **IEEE TVCG**, 2005. Citado nas páginas 15, 31, 47 e 48.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **National Academy of Sciences**, 2002. Citado nas páginas 27, 28 e 42.

HADANY, R.; HAREL, D. A multi-scale algorithm for drawing graphs nicely. In: _____. **Graph-Theoretic Concepts in Computer Science: 25th International Workshop, WG'99 Ascona, Switzerland, June 17–19, 1999 Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999. Citado na página 31.

- HAM, F. van; WATTENBERG, M.; VIEGAS, F. B. Mapping text with phrase nets. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 15, n. 6, p. 1169–1176, nov. 2009. ISSN 1077-2626. Citado na página 29.
- HAUGUEL, S.; ZHAI, C.; HAN, J. Parallel pathfinder algorithms for mining structures from graphs. In: WANG, W.; KARGUPTA, H.; RANKA, S.; YU, P. S.; WU, X. (Ed.). **ICDM**. [S.l.]: IEEE Computer Society, 2009. p. 812–817. Citado na página 27.
- ISENBERG, P.; HEIMERL, F.; KOCH, S.; ISENBERG, T.; XU, P.; STOLPER, C.; SEDLMAIR, M.; CHEN, J.; MÖLLER, T.; STASKO, J. **Visualization Publication Dataset**. 2015. Citado na página 43.
- JIA, Y.; HOBEROCK, J.; GARLAND, M.; HART, J. On the visualization of social and other scale-free networks. **IEEE TVCG**, IEEE Computer Society, v. 14, n. 6, 2008. ISSN 1077-2626. Citado na página 28.
- KARPIS, G.; KUMAR, V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. **SIAM Journal on Scientific Computing**, v. 20, n. 1, p. 359–392, 1998. Citado na página 30.
- KUANG, D.; PARK, H.; DING, C. H. Q. Symmetric nonnegative matrix factorization for graph clustering. In: **SDM**. [S.l.]: SIAM / Omnipress, 2012. ISBN 978-1-61197-282-5. Citado nas páginas 24, 31, 38 e 46.
- LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, v. 401, n. 6755, 1999. Citado nas páginas 24, 31, 36 e 37.
- LIAO, Q.; SHI, L.; TONG, H.; HU, Y.; ZHAO, Y.; LIN, C. Hierarchical focus+context heterogeneous network visualization. **2014 IEEE PacificVIS**, IEEE Computer Society, v. 0, p. 89–96, 2014. Citado na página 32.
- MENA-CHALCO, J. P.; JUNIOR, C.; MARCONDES, R. Scriptlattes: an open-source knowledge extraction system from the lattex platform. **Journal of the Brazilian Computer Society**, SciELO Brasil, v. 15, n. 4, p. 31–39, 2009. Citado na página 66.
- NAVLAKEHA, S.; RASTOGI, R.; SHRIVASTAVA, N. Graph summarization with bounded error. In: WANG, J. T.-L. (Ed.). **SIGMOD**. [S.l.]: ACM, 2008. Citado na página 30.
- NEWMAN, M. E. J. Fast algorithm for detecting community structure in networks. **Physical Review E**, v. 69, September 2003. Citado nas páginas 16, 30, 45, 48, 49, 50 e 51.
- NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v. 69, n. 026113, 2004. Citado nas páginas 15, 27 e 28.
- NOCAJ, A.; ORTMANN, M.; BRANDES, U. Untangling hairballs - from 3 to 14 degrees of separation. In: **Graph Drawing - 22nd International Symposium, GD 2014, Würzburg, Germany, September 24-26, 2014, Revised Selected Papers**. [S.l.: s.n.], 2014. p. 101–112. Citado na página 27.
- _____. Adaptive disentanglement based on local clustering in small-world network visualization. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 6, p. 1662–1671, June 2016. ISSN 1077-2626. Citado na página 27.

- QUIRIN, A.; CORDÓN, O.; SANTAMARÍA, J.; VARGAS-QUESADA, B.; ANEGÓN, F. de M. A new variant of the pathfinder algorithm to generate large visual science maps in cubic time. **Inf. Process. Manage.**, v. 44, n. 4, p. 1611–1623, 2008. Citado na página 27.
- RIECK, B.; FUGACCI, U.; LUKASCZYK, J.; LEITTE, H. Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks. **IEEE Transactions on Visualization and Computer Graphics**, 2017. ISSN 10772626. Citado na página 32.
- SHI, L.; TONG, H.; TANG, J.; LIN, C. Vegas: Visual influence graph summarization on citation networks. **IEEE Trans. Knowl. Data Eng.**, v. 27, n. 12, p. 3417–3431, 2015. Citado na página 31.
- SHUMAN, D. I.; NARANG, S. K.; FROSSARD, P.; ORTEGA, A.; VANDERGHEYNST, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. **IEEE Signal Processing Magazine**, Ieee-Inst Electrical Electronics Engineers Inc, Piscataway, v. 30, n. 3, p. 16. 83–98, 2013. Citado nas páginas 24, 53 e 54.
- SUN, J.; PAPADIMITRIOU, S.; LIN, C.-Y.; CAO, N.; LIU, S.; QIAN, W. Multivis: Content-based social network exploration through multi-way visual analysis. In: **SDM**. [S.l.]: SIAM, 2009. p. 1064–1075. Citado na página 32.
- TIAN, Y.; HANKINS, R. A.; PATEL, J. M. Efficient aggregation for graph summarization. In: **SIGMOD**. [S.l.]: ACM, 2008. p. 567–580. Citado nas páginas 30 e 45.
- VALDIVIA, P.; DIAS, F.; PETRONETTO, F.; SILVA, C. T.; NONATO, L. G. Wavelet-based visualization of time-varying data on graphs. In: **VAST**. [S.l.]: IEEE Computer Society, 2015. p. 1–8. Citado na página 53.
- VEHLOW, C.; BECK, F.; WEISKOPF, D. The state of the art in visualizing group structures in graphs. In: **EuroVis - STARS**. [S.l.: s.n.], 2015. Citado na página 23.
- _____. Visualizing dynamic hierarchies in graph sequences. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, p. 2343–2357, 2016. Citado na página 32.
- WALSHAW, C. A multilevel algorithm for force-directed graph drawing. In: **Proceedings of the 8th International Symposium on Graph Drawing**. London, UK, UK: Springer-Verlag, 2001. (GD '00), p. 171–182. ISBN 3-540-41554-8. Citado na página 31.
- WANG, R.-S.; ZHANG, S.; WANG, Y.; ZHANG, X.-S.; CHEN, L. Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. **Neurocomputing**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 2008. ISSN 0925-2312. Citado nas páginas 31, 45 e 46.
- WATTENBERG, M. Visual exploration of multivariate graphs. In: **ACM. Proceedings of the SIGCHI conference on Human Factors in computing systems**. [S.l.], 2006. p. 811–819. Citado na página 32.
- YOGHOORDJIAN, V.; DWYER, T.; GANGE, G.; KIEFFER, S.; KLEIN, K.; MARRIOTT, K. High-quality ultra-compact grid layout of grouped networks. **IEEE TVCG**, v. 22, n. 1, p. 339–348, 2016. Citado na página 29.
- ZHANG, N.; TIAN, Y.; PATEL, J. M. Discovery-driven graph summarization. In: **ICDE**. [S.l.]: IEEE Computer Society, 2010. Citado nas páginas 30 e 45.

ZHOU, H.; XU, P.; YUAN, X.; QU, H. Edge bundling in information visualization. **Tsinghua Science and Technology**, TUP, 2013. Citado na página [32](#).