UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Mineração de redes complexas k-partidas

Fabiana Rodrigues de Góes

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

ICMC USP SÃO CARLOS

Fabiana Rodrigues de Góes

Mineração de redes complexas k-partidas

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. Alneu de Andrade Lopes

USP – São Carlos
Fevereiro de 2023

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

d278m

de Góes, Fabiana Rodrigues
Mineração de redes complexas k-partidas / Fabiana Rodrigues de Góes; orientador Alneu de Andrade Lopes. -- São Carlos, 2023.
132 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, 2023.

1. Aprendizado de Máquina. 2. Redes k-partidas. 3. Propagação em grafos. 4. Aprendizado de Representação. I. de Andrade Lopes, Alneu, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Fabiana Rodrigues de Góes

Mining k-partite complex networks

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Alneu de Andrade Lopes

USP – São Carlos
February 2023

Dedico esse trabalho aos meus pais Góes (in memoriam) e Maria,
com todo meu amor e gratidão,
por serem os maiores incentivadores dos meus sonhos e projetos.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por toda a proteção e força na minha caminhada.

Aos meus amados pais Góes (in memoriam) e Maria por todo amor e apoio incondicional ao longo da minha vida. Obrigada por todo incentivo e esforço feito para que eu pudesse me dedicar exclusivamente aos estudos.

Ao meu namorado Lucas Ribas que esteve sempre ao meu lado ao longo destes anos de doutorado. Agradeço por todo amor, parceria e apoio incansável, que foram essenciais nesta jornada.

Aos meus amigos incríveis Daniela, Lílian, Rafael, Allisfrank, Gabriel, Diego, Kamila, Alessandra e Renata e Elizabeth por todo amor, pelas longas conversas acolhedoras e pela torcida.

Às minhas queridas amigas Marcela, Isabelle, Mariane e Francielle, por todo carinho, apoio e acolhimento. Aos meus amigos Wouter e Yuri por todos os momentos repletos de boas risadas.

A todos os meus colegas do LABIC pela convivência diária. Em especial, agradeço aos meus amigos Alan Valejo e Jorge Valverde pelas conversas, colaborações e apoio.

Ao meu orientador prof. Dr. Alneu de Andrade Lopes por ter me dado a oportunidade de desenvolver esta pesquisa, pela orientação e dedicação para fazer com que esse trabalho fosse possível.

Ao meu supervisor de estágio prof. Dr. Evangelos Milios pela orientação, pelos conhecimentos compartilhados e pelo acolhimento durante o meu período de pesquisa na *Dalhousie University*.

Ao Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São Paulo (Brasil) por fornecer a estrutura acadêmica que permitiu o desenvolvimento deste trabalho.

Por fim, gostaria de agradecer à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro. Ao programa ELAP (*Emerging Leaders in the Americas Program*) por financiar o meu estágio no Canadá.

“Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar.
Mas o mar seria menor se lhe faltasse uma gota.”
(Madre Teresa de Calcutá)

RESUMO

GÓES, F. R. **Mineração de redes complexas k-partidas**. 2023. 132 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, 2023.

Nos dias atuais, há uma grande quantidade de dados sendo produzida e disponibilizada diariamente. Como consequência, a organização e extração de informações úteis de forma manual a partir destes dados exige um grande esforço de especialistas. Deste modo, métodos computacionais de aprendizado de máquina e mineração de dados têm ganhado destaque, pois possibilitam a extração automática de conhecimento de grandes volumes de dados para resolver tarefas complexas em diversos contextos e aplicações. Em paralelo, Redes Complexas tornou-se uma importante área de pesquisa, principalmente, em razão da sua eficiência em modelar inúmeros sistemas da natureza e da sociedade. As redes k-partidas são casos particulares das redes heterogêneas, pois representam vértices de diferentes tipos que podem ser divididos em k conjuntos disjuntos. Esse tipo de rede é relevante para estudar diversos sistemas do mundo real, visto que modela os padrões intrínsecos das conexões entre diferentes tipos de objetos, o que não é naturalmente possível obter com as redes homogêneas. Os métodos de aprendizado de representação baseados em redes buscam aprender representações numéricas compactas que conservem as características intrínsecas e capturem informações latentes dos relacionamentos entre os vértices das redes. Técnicas do aprendizado de máquina mostram que diferentes visões de dados tendem a contribuir entre si, favorecendo o aprendizado. Pode-se, portanto, assumir uma rede k-partida como um conjunto de diferentes visões bipartidas, relacionadas entre si, que possibilitam a troca de informações. Assim, esta tese propõe abordagens baseadas na transferência de informações entre diferentes camadas de redes k-partidas, utilizando como base um método de propagação em redes bipartidas, para problemas de aprendizado não supervisionado de representação. A fim de demonstrar a importância da proposta, diferentes abordagens foram desenvolvidas para contextos reais que possuem dados que assumem uma estrutura k-partida, como recomendação em sistemas colaborativos de marcação e predição de associação entre lncRNAs e doenças. As análises experimentais mostram resultados promissores nas aplicações abordadas e fornecem indícios para a elaboração de trabalhos futuros. Sendo assim, os achados do trabalho poderão apoiar o desenvolvimento de novos métodos de aprendizado em redes k-partidas e novas abordagens para diversos tipos de dados e aplicações.

Palavras-chave: Aprendizado de Máquina, Redes k-partidas, Propagação em grafos, Aprendizado de Representação.

ABSTRACT

GÓES, F. R. **Mining k-partite complex networks**. 2023. 132 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, 2023.

Nowadays, there is a large amount of data being produced and made available daily. As a consequence, the organization and extraction of useful information manually from this data requires a great effort from specialists. Thus, computational methods of machine learning and data mining have gained prominence, as they enable the automatic extraction of knowledge from large volumes of data to solve complex tasks in different contexts and applications. In parallel, Complex Networks has become an important area of research, mainly due to its efficiency in modeling numerous systems of nature and society. K-partite networks are particular cases of heterogeneous networks, characterized by vertices of different types that can be separated into k disjoint sets. This type of network is relevant for studying different systems in the real world, since it models the intrinsic patterns of connections between different types of objects, which is not naturally possible to obtain through the homogeneous networks. Network-based representation learning methods seek to learn compact numerical representations that preserve the intrinsic characteristics and capture latent information on the relationships between the vertices of the networks. Machine learning techniques show that different views of data tend to contribute to each other, favoring learning. Therefore, we can assume a k-partite network as a set of different bipartite views, related to each other, that enable the exchange of information. Thus, this thesis proposes approaches based on the transfer of information between different layers of k-partite networks, using as a basis a propagation method in bipartite networks, for unsupervised representation learning problems. In order to demonstrate the importance of the proposal, different approaches were developed for real contexts that have data that assume a k-partite structure, as a recommendation in collaborative tagging systems and prediction of associations between lncRNAs and diseases. The experimental analyzes show promising results in the applications addressed and provide clues for the elaboration of future works. Thus, the findings of the thesis may support the development of new learning methods in k-partite networks and new approaches for different types of data and applications.

Keywords: Machine Learning, k-partite networks, Propagation in graphs, Representation Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
ANN	*Artificial Neural Network*
ARR	Aprendizado de Representação em Rede
AUC	*Area Under the ROC Curve*
BOW	*Bag-of-words*
FC	Filtragem colaborativa
FCTags	FC baseada no perfil de atribuição de *tags*
GAD	Grafo Acíclico Dirigido
GFHF	*Gaussian fields and Harmonic Functions*
HNE	*Heterogeneous Network Embedding*
IA	Inteligência Artificial
IID	Independente e Identicamente Distribuídos
IMC	*Inductive Matrix Completion*
KL	*Kullback-Leibler*
LDA	*Latent Dirichlet Allocation*
LE	*Laplacian Eigenmaps*
LLGC	*Learning with Local and Global Consistency*
LOO	*leave-one-out*
MDS	*Multi-Dimensional Scaling*
MeSH	*Medical Subject Headings*
MMPT	Mix of Most Popular Tags
ncRNAs	RNAs não-codificantes
NLM	*National Library of Medicine*
NMF	*Non-negative Matrix Factorization*
PBG	*Propagation in Bipartite Graph*
PCA	*Principal Components Analysis*
PMF	*Probabilistic Matrix Factorization*
PTGAP	**P**ropagation in **T**ripartite **G**raph for **A**ssociation **P**rediction
PTGRL	***P**ropagation in **T**ripartite **G**raph for **R**epresentation **L**earning*
ROC	*Receiver Operating Characteristic Curve*
TB	*Tag-based Model*

TCBHN *Transductive Classification based on Bipartite Heterogeneous Network*

TPBG *Transductive Propagation in Bipartite Graph*

TRLBG ***T**ag **R**ecommender based on **R**epresentation **L**earning in **B**ipartite **G**raph*

UCTM *User-Centric Tag Model*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

Muitos sistemas reais são caracterizados por possuírem uma estrutura complexa, pois são compostos por um grande número de componentes que interagem entre si por meio de padrões de conexões não triviais; suas propriedades globais não podem ser explicadas por meio de análises individuais dos seus componentes de forma isolada (NEWMAN, 2003). Consequentemente, a comunidade científica se empenhou para identificar estruturas eficientes para modelar, extrair e reconhecer padrões inerentes dos sistemas complexos (AMARAL; UZZI, 2007).

A Teoria das Redes Complexas emergiu como uma forma de compreender o comportamento e organização dos elementos que compõem os sistemas complexos (BOCCALETTI *et al.*, 2006; AMARAL; UZZI, 2007) através da representação natural desses sistemas como redes, estruturas compostas por um conjunto de vértices conectados entre si conforme as suas interações (COSTA *et al.*, 2010; NEWMAN, 2010). Essa é uma área de pesquisa multidisciplinar que desperta interesse em diversas áreas da ciência, uma vez que permite descrever e analisar uma ampla variedade de sistemas reais, como redes sociais, de comunicação, biológicas e químicas (SHI *et al.*, 2017). Assim, as redes complexas se estabeleceram como uma abstração poderosa para representar padrões de interações entre as partes de um sistema, pois a estrutura e o padrão particular das interações refletem diretamente o comportamento dos sistemas complexos (NEWMAN, 2010).

Os avanços tecnológicos ocorridos nas últimas décadas impulsionaram o aumento da geração, aquisição e armazenamento de dados de diferentes domínios. À medida que os conjuntos de dados aumentaram em tamanho e complexidade, a análise de dados caminhou para uma nova vertente (ZANIN *et al.*, 2016) que estabeleceu a necessidade de ferramentas computacionais mais sofisticadas para o processamento automatizado de dados (FACELI *et al.*, 2011). Por consequência, a mineração de dados e o aprendizado de máquina surgiram para impulsionar

a descoberta automática de conhecimento através da modelagem, análise e identificação de padrões em grandes conjuntos de dados usando uma variedade de técnicas computacionais.

As redes complexas e o aprendizado de máquina propiciam ferramentas para a análise e extração de informações de sistemas complexos. Portanto, essas duas áreas compartilham objetivos similares, como o fortalecimento do estudo de comportamento de sistemas reais de larga escala por meio de análises de dados mais adequadas e robustas (ZANIN *et al.*, 2016). No entanto, os métodos tradicionais de aprendizado de máquina não baseadas em redes assumem que os exemplos de dados são Independente e Identicamente Distribuídos (IID). Entretanto, uma variedade de dados reais possuem estrutura intrínseca de rede (relacional) ou podem ser representados em rede através de técnicas de modelagens como, por exemplo, conectar exemplos de dados que são mais similares (ou próximas) entre si (SILVA; ZHAO, 2016). Desse modo, nas últimas décadas, a análise de rede tornou-se um tópico de pesquisa em destaque na mineração de dados e aprendizado de máquina devido à demanda de técnicas que lidam com dados complexos, relacionais e não-lineares. Nesse contexto, o paradigma básico é encontrar padrões ocultos através da mineração das relações a partir de dados modelados em rede (ZANIN *et al.*, 2016; SHI *et al.*, 2017; BERTON; LOPES *et al.*, 2016).

Os algoritmos de aprendizado de máquina podem ser divididos em três categorias conforme a utilização da informação de rótulos, são elas: (i) aprendizado supervisionado, (ii) aprendizado não supervisionado e (iii) aprendizado semissupervisionado (SILVA; ZHAO, 2016). No aprendizado supervisionado, o intuito é aprender uma função que mapeia os dados em seus respectivos rótulos, denominada hipótese ou modelo, a partir de um conjunto de dados previamente rotulado. A ideia é de que o modelo gerado seja capaz de rotular novos exemplos de dados dos quais os rótulos são desconhecidos. Por outro lado, no aprendizado não supervisionado não há a necessidade de conhecimento prévio sobre rótulos existentes, pois o objetivo principal é identificar padrões que ocorrem naturalmente em um conjunto de dados sem a utilização de informações de rótulos. Entre as categorias anteriores encontra-se o aprendizado semissupervisionado que, além dos dados rotulados (em geral, em uma quantidade pequena), utiliza dados não rotulados a fim de agregar mais informações que possam contribuir no processo de aprendizado (CHAPELLE; SCHOLKOPF; ZIEN, 2006; SILVA; ZHAO, 2016).

1.2 Motivação

Algoritmos de aprendizado de máquina buscam aprender padrões específicos em um conjunto de dados, representados por meio de atributos ou características, para resolver tarefas relativamente complexas, como diagnóstico automático de doenças. Desta forma, o desempenho de algoritmos de aprendizado de máquina está ligado diretamente com a representação de dados utilizada (BENGIO; COURVILLE; VINCENT, 2013). Nesse contexto, o aprendizado de representação é uma subárea recente do aprendizado de máquina com ênfase no estudo de extração de

representações matematicamente e computacionalmente eficientes, para a geração de modelos classificadores ou outros preditores (BENGIO; COURVILLE; VINCENT, 2013; SUN *et al.*, 2020). Métodos tradicionais de extração de características podem não extrair com eficácia informações latentes a partir de dados complexos e não lineares (SUN *et al.*, 2020) e, como alternativa, os métodos baseados em redes complexas buscam aprender e gerar representações numéricas compactas que conservem as características intrínsecas e capturem informações latentes dos relacionamentos entre os exemplos de dados, que correspondem aos vértices das redes. Em geral, o processo de aprendizado de representação pode ser supervisionado, semissupervisionado e não supervisionado (GONG *et al.*, 2014; WANG; CUI; ZHU, 2016; ZHANG *et al.*, 2018; HUANG *et al.*, 2019; SUN *et al.*, 2020). Porém, as técnicas não supervisionadas são uma potencial alternativa para os casos de aplicações que não dispõem de uma quantidade significativa de dados rotulados.

Um aspecto importante das redes é a heterogeneidade de tipos de vértices, podendo ser caracterizadas como homogêneas ou heterogêneas. As abordagens baseadas em redes homogêneas se aplicam a casos em que os objetos e seus relacionamentos são tratados como sendo de um único tipo (SUN; HAN, 2013; SHI *et al.*, 2017). Entretanto, essa modelagem se torna limitada para casos em que a rede é formada naturalmente por objetos de diferentes tipos conectados através de interações que carregam informações semânticas importantes. Em contrapartida, as redes heterogêneas são representações poderosas e expressivas, que possibilitam a obtenção de informações mais ricas, pois modelam os padrões intrínsecos das conexões entre diferentes tipos de objetos.

As redes k-partidas são um caso especial de redes heterogêneas em que o conjunto total de vértices da rede é decomposto em k conjuntos disjuntos, de forma que os vértices que pertencem ao mesmo conjunto não possuem conexão entre si, uma vez que são permitidas apenas ligações entre vértices de tipos distintos. Sendo assim, essa propriedade tende a fornecer vantagens para a extração de padrões, visto que captura bem as relações heterogêneas formadas entre camadas de diferentes tipos de vértices. Além disso, as redes k-partidas são uma forma intuitiva e eficiente de modelar dados (FALEIROS; ROSSI; LOPES, 2017), pois muitas aplicações possuem naturalmente esta estrutura (PAVLOPOULOS *et al.*, 2018) e, sendo assim, se adaptam perfeitamente a problemas que envolvem descoberta de padrões nas relações entre dois, ou mais, tipos de objetos.

Nos últimos anos, diversas técnicas de aprendizado de máquina baseadas em redes heterogêneas vêm sendo desenvolvidas e mostrando avanço nos resultados para a solução de diferentes problemas, como: reposicionamento de fármacos (LUO *et al.*, 2017; SHAHREZA *et al.*, 2017; CHEN; ZHANG, 2018), classificação em redes bibliográficas (KONG *et al.*, 2012), classificação textual (ROSSI; LOPES; REZENDE, 2016; FALEIROS; ROSSI; LOPES, 2017) e extração de tópicos (FALEIROS; VALEJO; LOPES, 2020). Porém, apesar das redes heterogêneas fornecerem um novo paradigma de modelagem dos dados, também apresentam desafios para

muitas tarefas de mineração de dados (SHI *et al.*, 2017). Nesta modelagem, não existe uma única camada de elementos e relações, mas sim um conjunto de camadas que possuem padrões de conexões específicas entre si e, assim, torna-se mais complexo compreender se, e como, as relações entre diferentes camadas da rede auxiliam no processo de aprendizado.

No contexto de aprendizado semissupervisionado em redes bipartidas, foi proposto o método *Transductive Propagation in Bipartite Graph* (TPBG) Faleiros, Rossi e Lopes (2017) que consiste em um esquema de propagação iterativo em uma estrutura de rede bipartida para realizar aprendizado transdutivo semissupervisionado em coleções textuais e apresenta bons resultados de classificação em comparação com os métodos do estado da arte. Por outro lado, o método *Propagation in Bipartite Graph* (PBG) (FALEIROS; VALEJO; LOPES, 2020) compartilha da mesma base teórica do TPBG, mas para o aprendizado não supervisionado de padrões textuais, como extração de tópicos e agrupamento a partir de grafos bipartidos. Visto que os dois métodos foram projetados exclusivamente para redes bipartidas, uma pesquisa com potencial é a exploração de generalizações dos métodos para redes heterogêneas k-partidas, em que mais informações são incorporadas ao processo de aprendizado a partir da adição de mais camadas a estrutura bipartida. Além disso, técnicas de modelagem e mineração textual têm sido utilizadas em diferentes tipos de dados e aplicações, como classificação de dados de expressão gênica (BICEGO *et al.*, 2010; YALAMANCHILI; KHO; RAYMER, 2017), análise taxonômica e funcional de espécies (CHEN *et al.*, 2011; CHEN *et al.*, 2012), classificação de imagens (FOUMANI; NICKABADI, 2019) e sistemas de recomendação (KRESTEL; FANKHAUSER; NEJDL, 2009). Portanto, os métodos TPBG e PBG possuem potencial para serem explorados além do contexto de mineração de texto, como aplicações relacionadas aos sistemas de recomendação e predição de associações (conexões) entre entidades de sistemas biológicos.

1.3 Hipótese e Objetivos

Devido às lacunas destacadas anteriormente que limitam a capacidade do método PBG apenas para redes bipartidas, este trabalho tem como propósito o desenvolvimento de abordagens e análises experimentais que possibilitem a aplicação deste método em redes com mais de dois tipos de vértices, como as tripartidas. Além disso, devido à naturalidade das redes k-partidas em representar muitos sistemas reais, as novas abordagens foram desenvolvidas e aplicadas para solucionar problemas em diferentes aplicações.

Embora as contribuições principais desta tese sejam no contexto de aprendizado não supervisionado de representações, o desenvolvimento da pesquisa baseia-se em diferentes aspectos do aprendizado em redes complexas, como a teoria do aprendizado semissupervisionado. Uma das principais inspirações do estudo é o método *Co-training* (BLUM; MITCHELL, 1998), algoritmo semissupervisionado que utiliza múltiplas visões de um conjunto de amostras de dados e assume que cada visão fornece informações complementares sobre os dados e podem contribuir

entre si. Esse conceito fundamentou a ideia de que informações extraídas em diferentes camadas de uma estrutura *k*-partida, sejam elas classes ou vetores de informações latentes, poderiam colaborar entre si e auxiliar no aprendizado de informação realizado em cada camada.

A hipótese levantada neste trabalho é que a transferência de vetores de informações latentes, aprendidos por propagação em redes bipartidas em uma estrutura *k*-partida, pode contribuir com o aprendizado de representações mais ricas e discriminativas.

A fim de confirmar a hipótese, este trabalho tem como objetivo geral investigar a transferência de informações entre diferentes camadas de redes *k*-partidas utilizando como base a teoria do método de propagação em redes bipartidas PBG para problemas de aprendizado não supervisionado de representação em diferentes aplicações. Os objetivos específicos que norteiam esta pesquisa, são os seguintes:

- Investigar e estudar a formulação do método PBG a fim de obter fundamentos para o desenvolvimento de novas abordagens baseadas nesse método para a extração de representações em redes *k*-partidas;
- Testar diferentes formas de modelar a transferência de informação sobre as redes;
- Identificar aplicações em que os dados assumem estrutura *k*-partida e que possuam limitações e desafios que podem ser abordados pela proposta da pesquisa;
- Desenvolver abordagens baseadas em estrutura de redes *k*-partidas para o aprendizado de representações em diferentes aplicações;
- Realizar análises experimentais para avaliar e demonstrar a influência da transferência de informações entre as camadas das redes *k*-partidas utilizadas.

1.4 Contribuições

A principal contribuição teórica do trabalho foi a construção de modelos de transferência de vetores de informações entre camadas de redes *k*-partidas para o aprendizado não supervisionado. As modelagens de propagação propostas podem ser aplicadas em diferentes contextos de dados e facilmente aperfeiçoadas conforme novas necessidades. Com isso, foi apresentado um caminho de exploração das possibilidades de aprendizado em redes compostas por diferentes camadas, por meio de modelagens não ”caixa preta” que permitem a exploração simples e compreensíveis desse tipo de estrutura.

Além do desenvolvimento de novas técnicas, foi possível contribuir com dois diferentes domínios: biomédico e sistemas de recomendação. No contexto biomédico, o problema de predição de associação entre lncRNAs e doenças foi abordado, pois apesar da existência de evidência que mostram que essas moléculas têm importante relação com diversas doenças humanas, poucas

associações são conhecidas e validadas. Portando, existe a necessidade de ferramentas computacionais para a identificação de potenciais associações. Sendo assim, a abordagem PTGAP foi proposta para a predição de associação entre lncRNAs e doenças considerando uma rede tripartida lncRNA-doença-gene. Por outro lado, os sistemas de recomendação desempenham um papel importante nos sistemas digitais, visto que podem ajudar os usuários em vários processos de tomada de decisão por meio da recomendação de potenciais produtos ou serviço com base em perfis de consumo. Nesse sentido, foram desenvolvidas as propostas TRLBG para a recomendação de *tags* e PTGRL para a recomendação de itens em sistemas colaborativos de marcação, constituídos por uma estrutura de dados tripartida que descreve os relacionamentos entre usuários, *tags* e itens.

Além disso, análises experimentais foram realizadas com redes sintéticas para a observação dos comportamentos dos métodos conforme determinadas propriedades das redes. Para tanto, foi utilizada a ferramenta de geração de redes heterogêneas sintéticas, denominada BNOC (VALEJO *et al.*, 2019), desenvolvida em colaboração com outros membros do grupo de pesquisa. Desta forma, as abordagens propostas foram testadas em redes com diferentes características e comportamentos, sendo elas reais e sintéticas. Portanto, esta tese visou contribuir tanto para com o aspecto técnico da área de aprendizado em redes complexas, como no avanço do estado da arte de aplicações.

1.5 Organização da Tese

No Capítulo 2 é apresentada a fundamentação teórica da tese, contendo os principais aspectos relacionados a teoria de redes complexas e aprendizado de máquina, assim como a descrição de técnicas tradicionais da literatura. No Capítulo 3 apresentado o contexto de sistemas de recomendação para sistemas colaborativos de marcação. Nesse capítulo são abordadas as tarefas de recomendação de *tags* e *itens*, apresentadas as abordagens desenvolvidas e as análises experimentais. No Capítulo 4 é apresentado o problema de predição de associação entre LncRNA e doenças, trabalhos relacionados, abordagem proposta e análise experimental. Por fim, no Capítulo 5 são apresentadas as conclusões, limitações das abordagens propostas e os trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

O trabalho desta tese foi desenvolvido com base em duas áreas: redes complexas e aprendizado de máquina. Portanto, neste capítulo são discutidos os principais fundamentos teóricos e técnicas relacionados com a investigação desta tese. Na Seção 2.1, os conceitos e tipos de redes complexas utilizados nesta tese são apresentados. Na Seção 2.2 são apresentados os fundamentos teóricos de aprendizado de máquina e os principais métodos, especialmente dos paradigmas semissupervisionado e não supervisionado para aprendizado de representação.

2.1 Redes Complexas

A Teoria das Redes Complexas é uma área de pesquisa que tem como um de seus fundamentos a Teoria dos Grafos, um campo da matemática que estuda as relações entre um conjunto de objetos através de ferramentas matemáticas abstratas, denominadas grafos. A Teoria dos Grafos surgiu com o problema das sete pontes de Konigsberg resolvido por Euler em 1736 (EULER, 1741), através da representação formal de um grafo formado por vértices e arestas, como conhecido até hoje. Inicialmente, a Teoria dos Grafos focava no estudo de grafos regulares [1], a partir da década de 1950 com o trabalho de Erdos, Rényi *et al.* (1960) muitos sistemas passaram a ser estudados por meio do modelo de grafo aleatório (ALBERT; BARABÁSI, 2002).

Mais recentemente, houve o crescimento do interesse pelo estudo de sistemas complexos, compostos por uma grande quantidade de elementos que interagem entre si. Essas interações são caracterizadas por padrões não triviais, formando uma topologia complexa com propriedades e comportamentos emergentes (ALBERT; BARABÁSI, 2002; NEWMAN, 2003). O decorrer de pesquisas na área, fez cientistas questionarem se o paradigma de modelo de grafo aleatório seria suficiente para representar as propriedades de tais sistemas. Assim, houve o desenvolvimento de novos conceitos da área através de três importantes estudos sobre modelos de redes (COSTA *et*

[1] Um grafo é denominado regular se todos os vértices possuem o mesmo grau.

al., 2007): descoberta de redes pequeno mundo (WATTS; STROGATZ, 1998), descoberta de redes livre de escala (BARABÁSI; ALBERT, 1999) e identificação de estruturas de comunidade em redes (GIRVAN; NEWMAN, 2002)).

Nesse contexto, surgiu uma nova concepção sobre a modelagem de sistemas complexos, visto que os modelos propostos na Teoria dos grafos mostraram-se restritos diante das características dos sistemas complexos (BOCCALETTI *et al.*, 2006). Assim, agregando conceitos de áreas do conhecimento como a Física, Estatística, Ciência da Computação e Teoria dos Grafos, as Redes complexas surgiram como uma forma robusta de modelar e caracterizar as propriedades topológicas e o comportamento de sistemas complexos. A popularidade e a importância de redes complexas pode ser explicado pela sua característica multidisciplinar e eficiência em modelar inúmeros sistemas da natureza e da sociedade (COSTA *et al.*, 2007; BARABÁSI, 2016). Assim, a representação em redes têm sido utilizada para modelar uma ampla variedade de sistemas complexos em diferentes áreas. Newman (2010) define quatro principais categorias de redes complexas, como descrito a seguir:

Redes sociais: são redes em que os vértices correspondem a atores sociais e as relações representam alguma forma de interação entre eles em um determinado contexto social. Um exemplo são as redes de relacionamentos, obtidas a partir de redes sociais *online* que desempenham um papel importante na sociedade moderna. Neste contexto, os vértices representam usuários conectados por arestas que constituem as relações de amizade. Exemplos de redes sociais de relacionamento são: Facebook[2], Twitter[3], Instagram[4], entre outras.

Redes de informação: modelam objetos que correspondem a dados e suas relações que caracterizam troca de informação. Alguns exemplos deste tipo de redes são: redes de preferência, redes de coautoria, redes de citação e redes de páginas web (ou *World Wide Web*). As redes de preferência são compostas por objetos, que geralmente representam usuários, que estão conectados aos objetos (itens) de sua preferência, como livros, músicas ou filmes. Nessas aplicações um fator comum é a presença de informação sobre a intensidade da preferência, *e.g.* atribuição de uma nota a um item.

Redes tecnológicas: nessas redes os objetos representam equipamentos tecnológicos e as relações correspondem às conexões estabelecidas entre os objetos, caracterizadas por uma ligação física entre eles. Dentre os exemplos de redes tecnológicas estão as redes de energia elétrica, redes de telefonia e redes de computadores.

Redes biológicas: são constituídas por objetos que correspondem às entidades de um sistema biológico, tal como moléculas (*e.g* genes e proteínas), e relações que representam di-

[2] <https://www.facebook.com/>
[3] <https://www.twitter.com/>
[4] <https://www.instagram.com/>

ferentes tipos de interações entre essas entidades, como interações moleculares. Novas informações sobre funções e processos biológicos têm sido estudadas e reveladas por meio de modelagens de redes biológicas (PAVLOPOULOS *et al.*, 2018), como interação proteína-proteína (IPP), redes de regulação gênica, redes metabólicas, interação entre genes, interação entre microRNAs[5] (miRNA) e genes.

2.1.1 Conceitos Básicos de Redes

Inicialmente, quando se utiliza modelagem baseada em redes, é importante definir as propriedades básicas relacionadas com as arestas, vértices e formas de representação. Nesta seção são introduzidos estes aspectos, as notações e terminologias utilizadas para a definição de rede neste trabalho. É importante ressaltar que neste trabalho redes e grafos são considerados sinônimos.

As redes são descritas e representadas com base em conceitos da Teoria dos Grafos. Um grafo pode ser definido como $G(V,E)$, em que V representa o conjunto de vértices e E o conjunto de arestas ou relações, que por sua vez é definido por $\{(u,v) \mid u,v \in V\}$. Os tamanhos dos conjuntos V e E são denotados por $|V|$ e $|E|$, e representam, respectivamente, o número total de vértices e arestas do grafo. Um grafo pode possuir arestas com pesos ou forças de associação, neste caso o grafo é dito como ponderado e definido por $G(V,E,W)$, em que W corresponde ao conjunto de pesos das arestas. O peso de uma aresta entre um par de vértices u e v é denotado por $w(u,v)$.

Diz-se que a rede é direcionada (orientada ou dirigida) se os pares (u,v) no conjunto E possuem direção, caso contrário o grafo é denominado como não-direcionado. Assim, no que se refere as relações entre os vértices, uma rede não direcionada é dita simétrica, $\forall(u,v) \in E \Rightarrow (v,u) \in E$, enquanto uma rede direcionada não é necessariamente simétrica.

Os vértices dos grafos podem ser definidos de acordo com conceitos relacionados à conectividade. Dois vértices $x \in V$ e $u \in V$ são ditos adjacentes, ou vizinhos, se estão conectados por uma aresta $e \in E$. Em grafos não direcionados, se x for adjacente a y, então y deve ser adjacente a x. Por outro lado, em grafos direcionados, se x é adjacente a y, então y pode, ou não, ser adjacente a x. Assim, a vizinhança de um vértice $u \in V$ corresponde ao conjunto de vértices adjacentes a u, definida pela relação $\Gamma(u) = \{v \mid e_{u,v} \in E\}$. O grau de um vértice $u \in V$ é o número total de vértices adjacentes que ele possui, dado por $k_u = |\Gamma(u)|$. No caso dos grafos direcionados, existe o grau de saída, k^{out}, igual à quantidade arestas divergentes (que saem) de u, e o grau de entrada, k^{in}, correspondente ao número de arestas que incidem em u; o grau total é dado por, $k_u = k^{in} + k^{out}$.

Modelos de redes podem ser caracterizados a partir da distribuição do grau que afere se a rede possui estrutura de conexões aleatória, livre de escala, ou pequeno mundo. A distribuição

[5] MiRNAs são pequenas moléculas de RNA que têm como função a regulação da expressão gênica.

do grau é uma função de probabilidade que indica a probabilidade de ocorrência $P(k)$ de um determinado vértice v possuir grau k. Em termos de caracterização de redes, medidas estatísticas também podem ser obtidas a partir da distribuição do grau, tais como entropia e energia. Além disso, existe uma variedade de medidas utilizadas para fornecer uma descrição detalhada das estruturas das redes (COSTA *et al.*, 2010), podendo ser relacionadas com a estrutura topológica, centralidade, cálculos de caminhos e distâncias, conectividade, dinâmica entre outras (as descrições completas dessas medidas podem ser encontradas em (COSTA *et al.*, 2007)).

Estruturas para a Representação de Redes

Uma estrutura frequentemente utilizada para a representação de redes é a matriz de adjacências. Cada entrada de uma matriz de adjacência $A_{n\times n}$ indica se existe relação entre os vértices da rede, de forma que $a_{v,u} = 1$ quando os vértices estão conectados por uma aresta $e_{v,u}$, e zero caso contrário. No caso dos grafos ponderados, o peso das conexões podem ser considerados como os elementos da matriz de forma que $a_{v,u} = w_{v,u}$. Em um grafo não direcionado, a matriz de adjacência é simétrica, pois a entrada $a_{u,v}$ é igual à entrada $a_{v,u}$. Porém, se o grafo for direcionado, a matriz de adjacência pode não ser simétrica.

Em muitos casos a matriz de adjacência é esparsa, ou seja, a maioria dos seus elementos são iguais a zero, visto que a quantidade de conexões é muito menor do que o número de vértices da rede. Nestas situações é utilizada uma quantidade a mais de memória do que é necessário para armazenar as informações associadas as arestas do grafo.

Uma alternativa à matriz de adjacência é a estrutura de lista de adjacência, que consiste em um conjunto de listas A no qual cada vértice $u \in V$ do grafo possui uma lista que contém todos os vértices adjacentes a u. Para um grafo ponderado é necessário manter, para cada vértice u, uma lista extra que contenha os pesos das arestas. No caso de um grafo direcionado, duas listas de adjacência podem ser criadas, uma para armazenar as arestas de entrada e outra contendo apenas as arestas de saída de cada vértice u.

A matriz Laplaciana é outra maneira de representar as informações sobre a conectividade do grafo, mas não de forma direta como a matriz de adjacência. Na teoria espectral de grafos, a matriz Laplaciana é utilizada para encontrar propriedades importantes de um grafo. Dado um grafo com n vértices, sua matriz Laplaciana $L_{n\times n}$ é dada por

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2.1}$$

em que A é a matriz de adjacência do grafo e D é a matriz de grau que contém as informações sobre o grau de cada vértice $u \in V$ em sua diagonal, de forma que

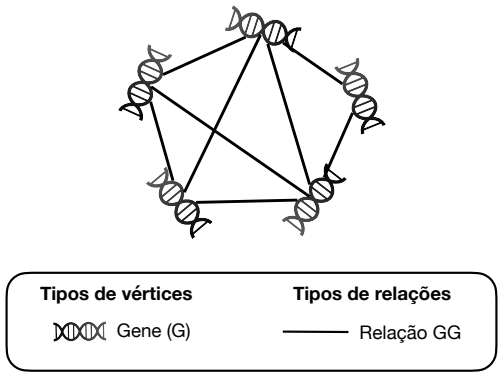
$$\mathbf{D}_{u,v} = \begin{cases} k_u, \text{ se } u = v \\ 0, \text{ caso contrário .} \end{cases}$$

2.1.2 Tipos de Redes

Um grafo $G = (V,E)$ é associado a uma função de mapeamento de tipo de vértice $f_v : V \mapsto T_v$ e uma função de mapeamento de tipo de aresta $f_e : E \mapsto T_e$, em que T_v e T_e denotam o conjunto de tipos de nós e tipos de arestas, respectivamente. Cada objeto $v_i \in V$ pertence a um tipo de objeto particular no conjunto de tipo de objeto $T_v : f_v(v_i) \in T_v$ e cada aresta $e_{i,j} \in E$ pertence a um tipo de relação particular no conjunto de tipo de relação $T_e : f_e(e_{i,j}) \in T_e$. Desta forma, as redes podem ser categorizadas conforme a heterogeneidade dos tipos de seus vértices e arestas.

Quando o conjunto de vértices V possui apenas um tipo de vértice e de relação, de forma que $|T_v| = |T_e| = 1$, a rede é denominada homogênea. Exemplos de redes homogêneas são as redes de citações, redes de interação entre proteínas e redes de coautoria. Muitas das análises tradicionais de redes assumem que o tipo de objetos ou conexões é único (SUN; HAN, 2013; SHI *et al.*, 2017), podendo ser modeladas de forma explicita, a partir de relações naturais, ou implícita, em que objetos similares são conectados entre si. Na Figura 1 é ilustrada uma rede homogênea de co-expressão gênica que, a partir de perfis de expressão, modelam a correlação entre os genes em um determinado estado de um sistema biológico.

Figura 1 – Exemplo de rede homogênea de co-expressão gênica.



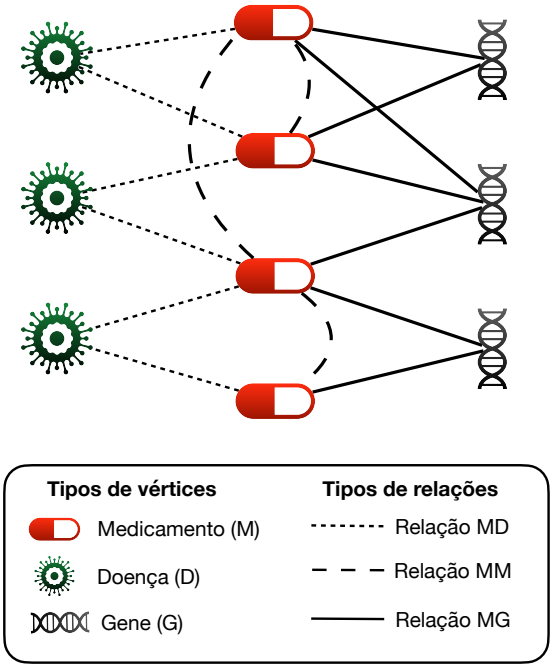
Fonte: Elaborada pelo autor.

Quando o conjunto de vértices V é composto por k tipos de vértices, $|T_v| > 1$ e/ou $|T_e| > 1$, a rede é denominada heterogênea. Redes multi-relacionais com um único tipo de vértice são um caso de redes heterogêneas como, por exemplo, redes sociais nas quais os usuários estão amplamente conectados uns com os outros por meio de diferentes tipos de relações como, por exemplo, conexões de amizade, troca de mensagens, conexões familiares e determinadas preferências. Por outro lado, as redes heterogêneas formadas por mais de um tipo de vértice incorporam relações derivadas das conexões entre vértices de categorias distintas. Shi *et al.* (2017) destacam duas vantagens importantes das redes heterogêneas: (i) capacidade de integrar várias fontes de dados que tenham relações entre si, pois modelam naturalmente diferentes tipos de relações entre diferentes tipos de objetos e (ii) possuem uma rica semântica de informações

proveniente da diversidade de tipos de objetos e suas conexões.

Na Figura 2 é ilustrado um exemplo de rede heterogênea utilizada no problema de reposicionamento de fármacos que contém três entidades principais: doença, medicamento e genes. O reposicionamento de fármacos consiste na utilização de fármacos já comercializados para o tratamento de novas doenças, visando ser uma alternativa mais rápida em relação ao desenvolvimento e validação de um novo medicamento. A rede é modelada de forma que cada medicamento possui conexões com um conjunto de doenças, genes e outros medicamentos, que mapeiam três tipos de relações: (i) medicamento-gene, (ii) medicamento-doença, e (iii) similaridade entre medicamentos. Normalmente, um fármaco tende a ter efeito por meio de interações com um ou mais genes e, por isso, durante o seu desenvolvimento é desejável que as relações entre medicamentos, genes e doenças sejam investigadas simultaneamente (SHAHREZA *et al.*, 2017). Neste sentido, é evidente a necessidade de uma representação integrativa de dados, tornando as redes heterogêneas uma alternativa potencial para investigar problemas com esses aspectos.

Figura 2 – Exemplo de rede heterogênea de reposicionamento de fármacos composta por vértices dos tipos medicamento, doença e gene.

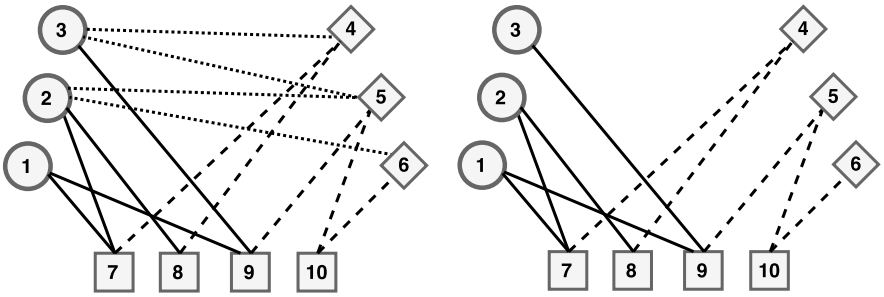


Fonte: Elaborada pelo autor.

Existem casos especiais de redes heterogêneas com características específicas, como as redes k-partidas em que o conjunto de vértices V é dividido em k conjuntos disjuntos, que podem ser interpretados como partições ou camadas da rede. A quantidade de vértices na camada i é dada por $|V_i|$. Nas Figuras 3(a) e 3(b) são ilustrados dois exemplos de redes heterogêneas k-partidas com $k = 3$. Um caso particular de rede k-partida muito utilizado é o grafo bipartido em que $k = 2$. Portanto, o conjunto total de vértices V de um grafo bipartido é particionado em

dois conjuntos disjuntos V_1 e V_2, de forma que $V_1 \cup V_2 = \emptyset$. Assim, toda aresta do grafo tem uma extremidade em V_1 e outra em V_2, como mostra a Figura 4(a).

Figura 3 – Exemplos de redes heterogêneas k-partidas.

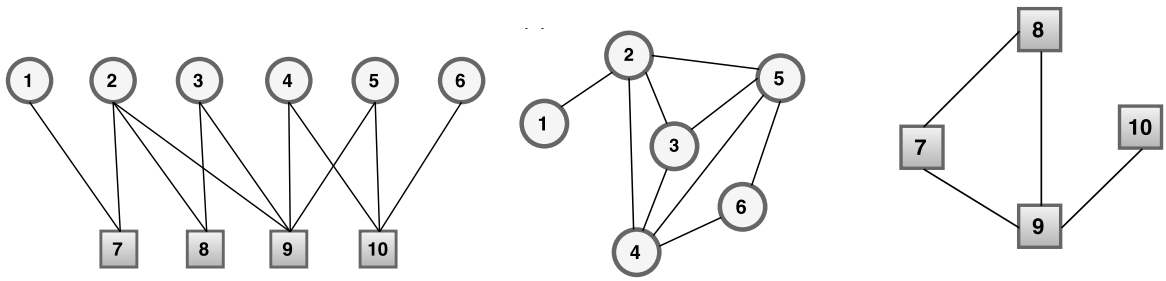


(a) (b)

Fonte: Elaborada pelo autor.

Uma rede bipartida pode ser projetada em relação a cada um de seus dois tipos de vértice (KITSAK; KRIOUKOV, 2011), denominada de projeção unipartida (do inglês, *one-mode projection*), em que dois vértices de um mesmo tipo (mesma camada) são conectados por uma aresta se possuírem pelo menos um vizinho em comum. Sendo assim, é possível obter duas redes, uma contendo os vértices do conjunto V_1 e outra contendo os vértices do conjunto V_2, como pode ser visto, respectivamente, na Figura 4(b) e Figura 4(c). Com isto, através da projeção unipartida é possível observar as relações entre os vértices de uma partição da rede, porém em muitos dos casos esta projeção pode ocasionar a perda significativa de informações (KITSAK; KRIOUKOV, 2011).

Figura 4 – Exemplo de rede bipartida e suas projeções. (a) Rede bipartida (b) Projeção unipartida da partição V_1 da rede bipartida (c) Projeção unipartida da partição V_2 da rede bipartida.



(a) (b) (c)

Fonte: Elaborada pelo autor.

2.1.3 Geração de Redes Sintéticas

O desenvolvimento de novos estudos e novas técnicas baseadas em redes complexas tem como desafio a necessidade de conjuntos de dados de referência que possuam características

topológicas bem determinadas para que novas abordagens e experimentos possam ser testados em contextos controlados, garantindo que a avaliação contemple diferentes propriedades. No entanto, a obtenção de uma considerável gama de redes reais que representem diferentes variações em suas estruturas topológicas, não é factível. Esta circunstância motiva geração e utilização de redes sintéticas. Desta forma, encontra-se na literatura trabalhos que propõem ferramentas para a geração de redes sintéticas a fim de auxiliar o desenvolvimento e análise de técnicas, como detecção de comunidades em redes.

Estrutura de comunidades é um conceito importante na análise de redes, uma vez que contribui com o entendimento de como os elementos de um sistema estão organizados (NEWMAN, 2010). Comunidade em redes pode ser definida como um conjunto de vértices que possuem uma alta densidade de conexão entre deles e baixa densidade de conexão com vértices de outros grupos (NEWMAN, 2003; COSTA *et al.*, 2007). Um exemplo da presença de estrutura de comunidades no mundo real são as redes sociais, caracterizadas por grupos de pessoas fortemente conectadas entre si que compartilham características em comum. Neste tipo de rede também ocorre a presença de elementos que interagem com diferentes grupos, caracterizando as comunidades como disjuntas ou sobrepostas (REES; GALLAGHER, 2012). No primeiro caso, os vértices pertencem exclusivamente a uma comunidade, enquanto as comunidades sobrepostas possuem vértices em comum.

Ao longo dos anos, diversos modelos teóricos foram desenvolvidos para sintetizar redes com determinadas características específicas (e.g., distribuição de grau), tais como os modelos de grafo aleatório (ERDOS; RÉNYI *et al.*, 1960), redes *small-world* (WATTS; STROGATZ, 1998) e redes livres de escala (BARABÁSI; ALBERT, 1999). A partir desses trabalhos, estudos passaram a recorrer a modelos sintéticos para simular e analisar o desempenho de algoritmos de detecção de comunidades (GIRVAN; NEWMAN, 2002; NEWMAN; GIRVAN, 2004). Com isso, ferramentas mais especializadas foram propostas para a geração de redes complexas com maior diversidade de propriedades presentes em redes reais (LANCICHINETTI; FORTUNATO; RADICCHI, 2008; MOUSSIADES; VAKALI, 2009; LARGERON *et al.*, 2015).

Até onde sabemos, dentre as ferramentas propostas para a geração de redes sintéticas, nenhuma das soluções aborda a geração de redes k-partidas com estruturas de comunidades sobrepostas. Essa lacuna foi a principal motivação para o desenvolvimento da ferramenta BNOC[6,7] (VALEJO *et al.*, 2019), desenvolvida em colaboração com a autora desta tese. Considerando que a principal contribuição desta tese é a proposta de diferentes estratégias para integração de informações entre diferentes camadas de redes k-partidas, é de grande importância a realização de análises experimentais em torno de redes com diferentes características topológicas e estruturais. Nesse sentido, a ferramenta BNOC foi utilizada para este fim.

[6] Inicialmente, a BNOC foi desenvolvida para redes bipartidas, mas recebeu uma extensão para redes k-partidas e heterogêneas, denominada HNOC.

[7] Disponível em: https://github.com/alanvalejo/bnoc

Através da ferramenta BNOC é possível gerar redes k-partidas que simulam diversas propriedades presentes em redes reais, por meio de parâmetros que garantem flexibilidade no controle da escala das estruturas de comunidades e demais propriedades topológicas das redes. Esses parâmetros são utilizados em uma distribuição binomial negativa responsável pela construção das redes. Os múltiplos parâmetros podem ser manipulados para criar redes com diferentes variações de características básicas, tais como quantidade de vértices, densidade de arestas, quantidade de camadas (redes bipartidas, tripartidas, etc.) e grau de ruído nos padrões de conexão. Além disso, é possível simular diferentes características em relação à estrutura de comunidades, como quantidade, tamanho, distribuição de arestas intra e intercomunidades e grau de ocorrência de sobreposição. A Tabela 1 mostra os parâmetros disponíveis na ferramenta.

Tabela 1 – Descrição dos parâmetros principais do BNOC para geração de redes sintéticas.

Parâmetro	Domínio	Descrição
-v, –vertices	$[1, \|V\|] \subseteq \mathbb{Z}$	Número de vértices de cada camada (tipo)
-e, –scheme	$\mathbb{N}$	Padrão de conexão entre as camadas
-d, –dispersion	$\mathbb{R}_+$	Dispersão da distribuição binomial negativa
-n, –noise	$(0, 1] \subseteq \mathbb{R}$	Ruído aplicado em cada camada
-b, –balanced	$\{0, 1\}$	Determina uma distribuição uniforme de comunidades
-u, –unweighted	$\{0, 1\}$	Define a rede como não-valorada
-c, –communities	$[1, \|V\|] \subseteq \mathbb{Z}$	Número de comunidades para cada camada
-x, –x	$(0, \|V\|] \subseteq \mathbb{N}$	Número de vértices sobrepostos para cada camada
-z, –z	$(0, \|c\|] \subseteq \mathbb{N}$	Número de comunidades sobrepostas para cada camada
-p, –probabilities	$(0, \|V\|] \subseteq \mathbb{R}$	Probabilidades dos vértices pertencerem às comunidades

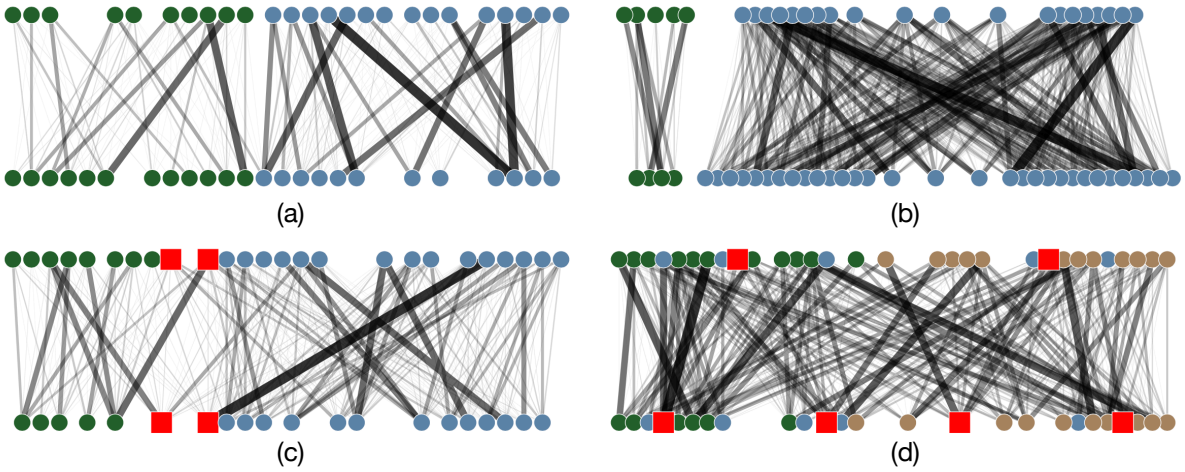
Fonte: Adaptada de Valejo *et al.* (2019).

A Figura 5 mostra várias redes construídas a partir de diferentes combinações de parâmetros. A rede da Figura 5(a) possui uma estrutura topológica esparsa e distribuição de comunidades balanceada, enquanto a rede da Figura 5(b) é mais densa e desbalanceada. A rede da Figura 5(c) possui vértices sobrepostos que participam das duas comunidades de cada camada. Por fim, a Figura 5(d) ilustra uma rede com topologia densa, presença de ruído e uma maior quantidade de vértices sobrepostos.

Neste trabalho, o potencial de utilização do BNOC foi demonstrado através da avaliação de dois algoritmos de detecção de comunidades sobrepostas, HLC (AHN; BAGROW; LEHMANN, 2010) e OSLOM (LANCICHINETTI *et al.*, 2011), em redes com diferentes propriedades. Os resultados indicam que cada método tende a ser mais sensível a diferentes propriedades. Por exemplo, o HLC é sensível ao número de vértices, comunidades sobrepostas e nível de ruído e, por outro lado, o OSLOM é sensível à densidade de arestas e a quantidade de vértices sobrepostos. Ademais, o tempo de execução do HLC são afetados pelo tamanho da rede e pela densidade das comunidades, enquanto os tempos de execução do OSLOM são afetados pela intensidade de ruído e pelo tamanho da rede.

Além de analisar métodos de detecção de comunidades, é importante ressaltar que

Figura 5 – Redes bipartidas com diferentes propriedades topológicas geradas pela ferramenta BNOC. A intensidade das arestas refletem os valores dos pesos; os círculos coloridos representam os vértices não sobrepostos e as suas respectivas comunidades conforme indicam as cores e os quadrados vermelhos representam vértices sobrepostos. Os valores dos parâmetros são informados na parte inferior de cada rede.



Fonte: Valejo *et al.* (2019).

é possível utilizar a ferramenta para avaliar o comportamento de métodos de outras tarefas, como classificação de vértices, extração de representação, predição de *links*, etc. O estudo de desempenho, conforme diferentes propriedades de redes, fornecem indícios de pontos que merecem atenção para o desenvolvimento e aplicação desses algoritmos.

2.2 Aprendizado de Máquina

A área de Aprendizado de Máquina (AM) evoluiu a partir da Inteligência Artificial (IA) devido à necessidade da modelagem computacional dos processos de aprendizagem, para que aquisições de conhecimento, reconhecimento de padrões e previsões sobre dados se tornassem processos automatizados (MITCHELL, 1997; JORDAN; MITCHELL, 2015). O aprendizado de máquina pode ser definido como uma área de estudo para o desenvolvimento de algoritmos que permitem que sistemas computacionais aprendam a partir de dados, sem serem explicitamente programados para realizar uma tarefa específica (MITCHELL, 1997; FACELI *et al.*, 2011; SILVA; ZHAO, 2016). Em muitos domínios, os exemplos de dados podem ser previamente rotulados manualmente por especialistas para que esta informação seja utilizada no processo de aprendizado; os dados com rótulo ou classe são denominados como dados rotulados e os demais como não rotulados. Assim, as abordagens de aprendizado de máquina podem ser categorizadas conforme as tarefas executadas nos dados e a utilização das informações de rótulos disponíveis.

Os algoritmos de AM podem ser especificados conforme o processo de inferência realizado sobre um conjunto de dados, em aprendizado transdutivo e aprendizado indutivo. No aprendizado indutivo, um conjunto de dados, denominado treinamento, $X = \{x_1, ..., x_n\}$ é

utilizado para a indução de uma função, ou modelo, $f : X \rightarrow Y$ capaz de realizar a predição de rótulos de novos exemplos de dados que não pertencem a X. Assim, o modelo de classificação induzido é então utilizado para rotular novos dados, conhecidos como conjunto de dados de teste. Por outro lado, o aprendizado transdutivo utiliza simultaneamente todo o conjunto de dados para estimar os rótulos dos exemplos não rotulados, sem a necessidade de indução de um modelo. Formalmente, o aprendizado transdutivo visa gerar uma função da forma $f : X^{L+U} \rightarrow Y^{L+U}$, a partir de dados rotulados X_L e dados não rotulados X_U, para predizer os rótulos de X_U. Assim, o aprendizado transdutivo pode ser definido como o processo de estimar os valores de uma função a partir dos pontos de dados de interesse (VAPNIK, 1998).

O aprendizado supervisionado realiza a indução de uma função $f : X \rightarrow Y$ a partir de um conjunto de dados $X = \{x_1, ..., x_n\}$ que possui um conjunto de rótulos $Y = \{y_1, ..., y_n\}$ associado, ou seja, constrói um modelo que mapeia o exemplo x_i a um rótulo de classe y_i. A partir do modelo gerado, as informações de rótulos de novos dados não rotulados poderão ser preditas no futuro. Conforme o domínio dos valores dos rótulos, esta categoria de aprendizado pode ser categorizada como (i) classificação quando os dados possuem uma saída qualitativa e (ii) regressão quando os dados possuem uma saída quantitativa, em que para cada exemplo de dado é predito um valor quantitativo.

No aprendizado não supervisionado é realizada a identificação da organização dos dados em estruturas intrínsecas a partir de um conjunto de dados $X = \{x_1, ..., x_n\}$ não rotulados. Desta forma, o objetivo é encontrar um padrão de estrutura em X sem nenhum conhecimento prévio de rótulos sobre os dados. De acordo com Silva e Zhao (2016), agrupamento e redução de dimensionalidade estão entre as principais tarefas do aprendizado não supervisionado. A tarefa de agrupamento consiste na criação de grupos de dados que compartilham caraterísticas e padrões semelhantes, de forma que dados em um mesmo grupo sejam mais similares entre si. Para tal, uma função é adotada para medir a similaridade entre os dados. Na redução de dimensionalidade, o objetivo é mapear os dados de um espaço de alta dimensão em um espaço de dimensão inferior, de modo a extrair uma representação reduzida que comprima as informações e relações entre os dados e aprender características representativas.

Para a geração de modelos de aprendizado, é necessário que uma quantidade suficiente de exemplos rotulados seja disponível. Porém, devido o processo de obtenção de dados rotulados ser geralmente muito custoso (ABNEY, 2007), a abordagem semissupervisionada tornou-se uma alternativa em que tanto os dados rotulados como não rotulados são considerados no processo de aprendizado. Espera-se com este tipo de aprendizado que sejam obtidas melhores performances em comparação com os métodos totalmente supervisionados, devido à inclusão das informações de dados não rotulados.

Representação de Dados

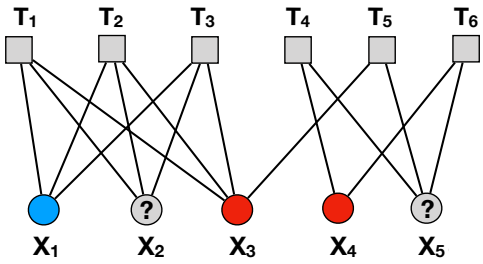
Para a utilização de algoritmos de aprendizado de máquina, independente do tipo ou categoria, os dados devem estar mapeados em um formato apropriado que permita a manipulação computacional. A qualidade da representação, ou modelagem, do conjunto de dados reflete diretamente no desempenho dos algoritmos de AM (AMANCIO *et al.*, 2014), sendo uma questão fundamental que norteia o processo de aplicação de algoritmos de AM. Desta forma, os algoritmos de AM podem ser categorizados segundo a forma de representação dos dados adotada, podendo ser baseada no modelo espaço-vetorial e baseada em grafos.

Na representação baseada no modelo espaço vetorial, os exemplos de um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$ são compostos por um conjunto de atributos $T = \{t_1, t_2, ..., t_m\}$. Assim, cada exemplo x_i corresponde a um vetor m-dimensional $\vec{x_i} = (t_1, ..., t_m)$. Em problemas de AM supervisionado ou semissupervisionado, cada exemplo de X assume a forma $(\vec{x_i}, y_i)$, em que y_i é o valor do atributo classe Y de x_i. Assim, a junção dos vetores de atributos resultam em uma matriz atributo-valor $X_{n,m}$, caracterizada por ser uma coleção de exemplos independentes e identicamente distribuídas, como mostra a Figura 6(a).

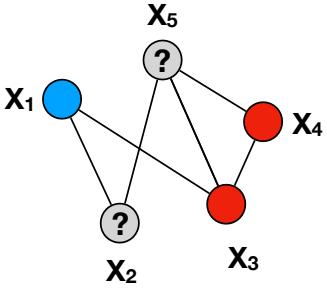
Figura 6 – Diferentes formas de representação de dados.

	T_1	T_2	T_3	T_4	T_5	T_6	Y
X_1	3	1	2	0	0	0	1
X_2	2	0	2	0	0	0	?
X_3	0	1	0	2	1	4	2
X_4	0	0	0	3	1	2	2
X_5	0	0	0	2	1	3	?

(a) Matriz atributo-valor



(b) Rede bipartida



(c) Rede homogênea

Fonte: Elaborada pelo autor.

Existem muitas aplicações em que os dados possuem alta dimensionalidade, como classificação textual (UYSAL; GUNAL, 2012), classificação de dados de expressão gênica (CHAKRABORTY; MAULIK, 2014) e classificação de imagens (CUI; PRASAD, 2013). Para amenizar este problema, técnicas de seleção de atributos são empregadas para a obtenção de um

subconjunto de atributos que contenha as informações mais significativas para o problema em questão e consiga representar eficientemente o conjunto de atributos original para o processo de aprendizado. Ademais, técnicas de extração de atributos (características) também são uma alternativa para a redução de dimensionalidade dos dados. Neste contexto, o objetivo é encontrar um mapeamento de baixa dimensão que represente as características latentes e intrínsecas dos dados de alta dimensão.

A representação de um conjunto de dados também podem ser feita utilizando redes. Esta abordagem possui como vantagem o uso da informação de relacionamento entre os objetos e pode prover informações adicionais em relação ao formato atributo-valor. Desta forma, o aprendizado baseado em redes é uma alternativa interessante ao aprendizado tradicional baseado no modelo espaço vetorial (GUILLORY; BILMES, 2009). Neste contexto, os conjuntos de dados multidimensionais podem ser mapeados para a estrutura de redes, na qual cada vértice corresponde a um exemplo do conjunto de dados (AGGARWAL, 2015, p. 34) (BERTON; LOPES *et al.*, 2016; BERTON *et al.*, 2017). As relações entre os vértices podem ser naturais ou geradas a partir de informações intrínsecas do conjunto de dados. Portanto, a modelagem em rede pode ser utilizada para representar e modelar dados de diferentes domínios e aplicações.

Em geral, as redes heterogêneas tendem a ser representações naturais, ou explícitas, de conjuntos de dados que contém diferentes tipos de objetos e relações, como redes bipartidas documento-termo que modelam as ocorrências das palavras nos documentos através de informações explícitas de uma coleção textual, como ilustrado na Figura 6(b). Em contrapartida, as redes homogêneas também podem representar naturalmente algum conjunto de dados, porém são bons exemplos de redes geradas por intermédio de informações implícitas dos dados. A modelagem da rede é um passo tão importante quanto realizar escolha entre os métodos de aprendizado existentes (ZHU, 2005) e, para tal, existe uma série de estudos e técnicas específicas (JEBARA; WANG; CHANG, 2009; BERTINI *et al.*, 2011; BERTON; LOPES *et al.*, 2016; BERTON *et al.*, 2017). Usualmente, a modelagem das relações são extraídas através de uma medida da similaridade aplicada entre os exemplos de um conjunto de dados. A Figura 6(c) mostra um exemplo de rede homogênea gerada a partir de uma representação atributo-valor. Abaixo são descritos dois métodos populares para construção de redes:

Rede ε-vizinhança: dois pontos de dados x_i e x_j são conectados por uma aresta se $\left|(x_i, x_j)\right| < \varepsilon$, em que ε é um limiar pré-definido e $|.|$ corresponde a uma função de distância, como a distância Euclidiana. A densidade da rede é determinada por ε, de forma que valores baixos resultam em redes esparsas e valores altos formam redes densas em que os vértices possuem alto grau de conectividade.

Rede k-vizinhos mais próximos: pode ser criada através de duas abordagens, *kNN* simétrica que cria uma aresta entre dois pontos de dados x_i e x_j se x_i estiver entre os k vizinhos mais próximos de x_j ou x_j estiver entre os n vizinhos mais próximos de x_i e *kNN* mútua em que

x_i e x_j são conectados somente se um par de vértice forem mutuamente os vizinhos mais próximos, ou seja, $x_i \in kvizinhos(x_j) \wedge x_j \in kvizinhos(x_i)$.

2.3 Aprendizado Semissupervisionado

A obtenção de dados rotulados é muito mais custosa em relação à coleta de dados sem informações de rótulos (SILVA; ZHAO, 2016), devido à dependência de um grupo de especialistas para analisar e definir um processo de rotulação. Como mencionado anteriormente, as tarefas de aprendizado não supervisionado e supervisionado se diferenciam principalmente na utilização de rótulos para o processo de aprendizado. O aprendizado semissupervisionado fundamenta-se na hipótese de que associar os dados não rotulados aos rotulados possibilitam uma melhor identificação de padrões. Assim, esse tipo de aprendizado baseia-se na combinação das teorias do aprendizado supervisionado e não supervisionado para que algoritmos utilizem não somente as informações de rótulos no processo de inferência, mas também considerar informações de distribuição dos dados não rotulados para inferir conhecimento e aprender padrões sobre o conjunto de dados em sua totalidade.

Formalmente, o processo de aprendizado semissupervisionado baseia-se em um conjunto de dados rotulados $X_l = \{x_1, ..., x_l\}$, associados a um conjunto rótulos $Y_l = \{y_1, ..., y_l\}$, e não rotulados $X_u = \{x_{l+1}, ..., x_{l+u}\}$. Assim, o total $X = X_l + X_u$ de exemplos no conjunto de dados é composto de L e U dados rotulados e não rotulados, respectivamente. Quando $X = X_l$ o processo de aprendizado é visto como um problema supervisionado. Alguns pressupostos sobre a consistência dos dados devem ser seguidos para que a combinação dos dados rotulados e não rotulados contribua de fato com o processo de aprendizado, de forma que as técnicas de aprendizado semissupervisionado tenham bom desempenho. Segundo Chapelle, Scholkopf e Zien (2006), o aprendizado semissupervisionado se baseia em três pressupostos principais:

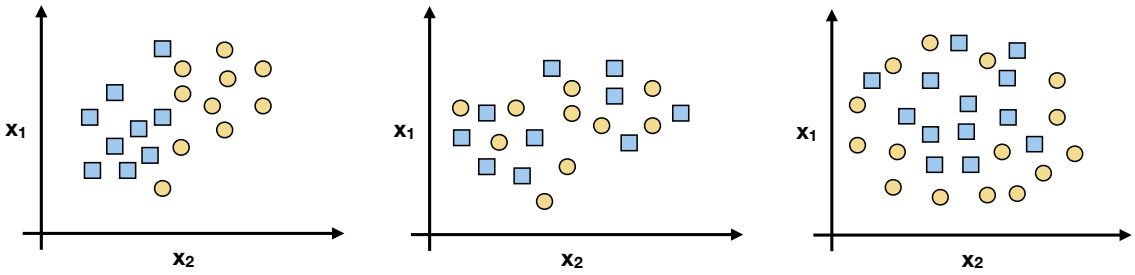
Suavidade: define que amostras de dados que estão em regiões muito densas do espaço de atributos são, provavelmente, membros da mesma classe. Assim, os rótulos dos exemplos devem variar de uma forma suave em áreas de densidade alta. Em outras palavras, se dois pontos x_1 e x_2 estiverem próximos, o mesmo deve ocorrer com as suas saídas correspondentes y_1 e y_2. A Figura 7 (a) e (b) mostra duas distribuições de dados, em que a primeira cumpre com este pressuposto e a outra não.

Formação de grupos (do inglês, *cluster assumption*): assume que amostras de dados localizados no mesmo grupo, provavelmente pertencem a mesma classe. Entretanto, esta definição não implica que os grupos são formados estritamente por dados da mesma classe, mas que amostras de classes diferentes tendem a não pertencer ao mesmo grupo (CHAPELLE; SCHOLKOPF; ZIEN, 2006). Este pressuposto afirma que os limites de decisão encontram-se em regiões de baixa densidade, definidas pela formação de grupos em regiões de alta

densidade. A Figura 7 (a) e (c) mostra duas distribuições de dados, na qual a primeira apresenta formação de grupos entre exemplos da mesma classe e a segunda não.

***Manifold*:** assume que um conjunto de amostras de dados em um espaço de alta dimensão pode ser, aproximadamente, reduzido a um espaço de menor dimensão, denominado estrutura *manifold*. Essa premissa é geralmente empregada para amenizar o problema de alta dimensionalidade (SILVA; ZHAO, 2016), dado que se os dados estiverem em um *manifold* o algoritmo de aprendizado pode operar nesse espaço correspondente de menor dimensão.

Figura 7 – Exemplos de pressupostos do aprendizado semissupervisionado.



(a) (b) (c)

Fonte: Elaborada pelo autor.

2.3.1 Aprendizado Semissupervisionado via modelo espaço vetorial

O *Self-training* (YAROWSKY, 1995) é um algoritmo semissupervisonado que, de forma iterativa, incrementa uma pequena quantidade de dados rotulados disponível inicialmente utilizando informações de dados não rotulados (HAFFARI; SARKAR, 2012; TRIGUERO; GARCÍA; HERRERA, 2015), como mostra o Algoritmo 1. Para tal, primeiramente, um classificador é treinado a partir de uma pequena quantidade de amostras rotuladas D_l. Em seguida, o modelo de classificação gerado é usado para classificar os exemplos de dados não rotulados em D_u e, então, as rotulados com mais confiança são selecionadas e movidos para o conjunto de dados rotulados D_l para que o modelo seja re-treinado considerando as novas amostras rotuladas. A confiabilidade de predição de uma amostra é determinada através de uma medida de confiança. Todo o processo é então repetido até que o conjunto não rotulado D_u esteja vazio ou um determinado critério de parada seja atendido como, por exemplo, um número máximo de iterações definido.

O *Self-training* é uma estratégia simples para o caso de aplicações que possuem pouca informação de dados rotulados. No entanto, este algoritmo tem como desvantagem uma possível degeneração do modelo caso haja a inserção de dados rotulados incorretamente nas primeiras iterações (AGGARWAL, 2015).

Algoritmo 1 – Algoritmo *Self-training*

1: **enquanto** U não é vazio **faça**
2: Treinar h com dados de L;
3: Rotular amostras de U com h;
4: selecionar um subconjunto de amostas U' de U classificadas com maior confiança;
5: $L \leftarrow L + U'$;
6: $U \leftarrow U - U'$;
7: **fim enquanto**

O *Co-training* (BLUM; MITCHELL, 1998) é um algoritmo semissupervisionado tradicional que utiliza diferentes visões de um conjunto de amostras de dados para treinar classificadores para cada visão e, iterativamente, aumentar a quantidade de dados rotulados disponível por meio da rotulação de dados não rotulados. Para tal, a técnica assume que cada visão fornece informações complementares sobre os dados, para que um classificador rotule as amostras de dados com base em informações desconhecidas pela visão do outro classificador.

Formalmente, o método *Co-training* funciona com base em duas visões $X^{(1)}$ e $X^{(2)}$ que representam o conjunto de dados X, de forma que $X = X^{(1)} \cup X^{(2)}$, e que cada exemplo x é dado como um par $(x^{(1)}, x^{(2)})$. O conjunto de dados X é composto por amostras rotuladas e não rotuladas $X = X_l + X_u$, em que o conjunto de dados rotulados é definido como $X_l = \left(\vec{x}_i^{(1)}, \vec{x}_i^{(2)}, y\right)_i^l$ e o conjunto de dados não rotulados é dado por $X_u = \left(\vec{x}_i^{(1)}, \vec{x}_i^{(2)}\right)_{i=l+1}^{l+u}$. Então, dois classificadores C_1 e C_2 são treinados sobre as funções objetivo respectivas $f_1 : X^{(1)} \rightarrow Y$ e $f_2 : X^{(2)} \rightarrow Y$. Inicialmente, os classificadores C_1 e C_2 são gerados a partir de uma pequena quantidade de dados rotulados X_l que, iterativamente, aumenta conforme os exemplos não rotulados em X_u, classificados com maior confiança, são incorporados como amostras rotuladas. O método *Co-training* considera que para um dado exemplo $x = (x^{(1)}, x^{(2)})$ as funções objetivo geradas em diferentes visões devem concordar na predição do rótulo, de forma que $f_1(x^{(1)}) = f_2(x^{(2)}) = y$.

Para o seu correto funcionamento, a técnica *Co-Training* baseia-se em duas premissas, (i) suficiência, assume que cada visão seja suficiente para treinar classificadores que realizem classificações satisfatórias e (ii) independência, define que as visões devem ser condicionalmente independentes dada a classe, ou seja, os atributos da visão X_1 e X_2 não dependem ou tem correlação direta entre si. Portanto, os classificadores treinados em diferentes visões possuem vieses distintos, que permitem que um colabore com o aprendizado do outro. No entanto, em um contexto real, é difícil encontrar conjuntos de dados que satisfaçam completamente estas duas premissas (WANG; ZHOU, 2007), principalmente a de independência. Consequentemente, diversos trabalhos investigaram a atenuação das premissas e mostram através de resultados promissores que o paradigma *Co-Training* pode ter bom desempenho mesmo quando as premissas não são rigorosamente atendidas (GOLDMAN; ZHOU, 2000; BALCAN; BLUM; YANG, 2005; ZHOU; LI, 2005).

Com base na formalização descrita acima, o processo do método *Co-training* é definido

no Algoritmo 2. A partir de um conjunto de dados rotulados e não rotulados $X = L \cup U$, o primeiro passo é selecionar um subconjunto de dados não rotulados U' de tamanho menor do que U. Em cada iteração, um classificador deve ser treinado com base em cada visão utilizando os dados em L. Em seguida, cada classificador gerado deve escolher p exemplos positivos e n exemplos negativos classificados com maior confiança para serem removidos do subconjunto U' e utilizados para atualizar o conjunto L. Por fim, os exemplos extraídos de U' devem ser repostos com $2n+2p$ exemplos não rotulados escolhidos aleatoriamente a partir de U.

Algoritmo 2 – Algoritmo *Co-training*

Entrada: $L \leftarrow$ amostras rotuladas, $U \leftarrow$ amostras não rotuladas, *loops* $\leftarrow$ quantidade de interações, $p \leftarrow$ Número de exemplos positivos a serem escolhidos, $n \leftarrow$ Número de exemplos positivos a serem escolhidos

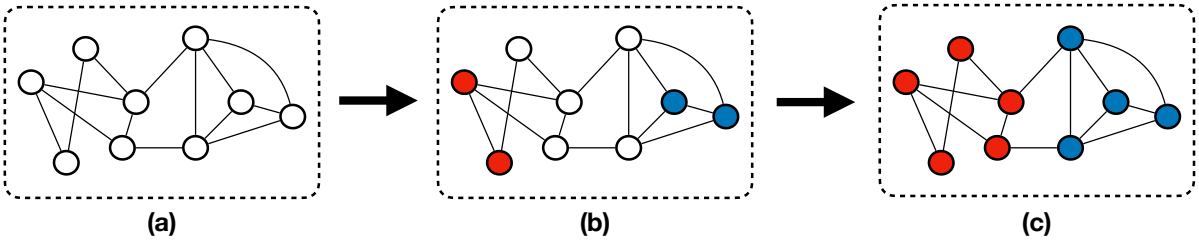
1: Obter duas visões L^1 e L^2 de L;
2: Criar um conjunto aleatório U' a partir de U;
3: **para** $k \leftarrow 1$ **até** loops **faça**
4: Treinar h_1 com L^1;
5: Treinar h_2 com L^2;
6: Rotular U' com h_1;
7: Selecionar $p+n$ exemplos rotulados com a maior confiança para atualizar L;
8: Rotular U' com h_2;
9: Selecionar $p+n$ exemplos rotulados com a maior confiança para atualizar L;
10: Escolher aleatoriamente $2p+2n$ exemplos do conjunto U para repor U';
11: **fim para**

2.3.2 Aprendizado Semissupervisionado em redes

O aprendizado semissupervisionado em rede consiste em um processo de atribuição de rótulos para cada vértice não rotulado através da estrutura da rede. O crescente interesse pelo uso de redes como representação de dados pode ser justificado pela capacidade da estrutura topológica dos relacionamentos de dados funcionar como uma informação valiosa para a inferência de classes. As conexões da rede refletem, ou indicam, a intensidade de similaridade entre os vértices e, assim, objetos relacionados tendem a pertencer a mesma classe (ZHU, 2005). Para tal, as técnicas baseadas em rede exploram as estruturas de vizinhanças da rede para analisar e prever rótulos para exemplos não rotulados (SILVA; ZHAO, 2016).

Em geral, os métodos de AM baseados em redes são transdutivos, pois todos os vértices da rede, rotulados ou não, são utilizados pelos algoritmos durante a fase de aprendizagem, sem a necessidade de projetar uma função de generalização global (SILVA; ZHAO, 2016). Assim, a classificação transdutiva em redes realiza a classificação de objetos não rotulados considerando informações de classe dos objetos vizinhos (ZHU, 2005), como ilustrado na Figura 8. Os algoritmos que realizam aprendizado transdutivo em redes podem sem categorizados como (ROSSI; LOPES; REZENDE, 2017): (i) classificação coletiva e (ii) regularização.

Figura 8 – Propagação de rótulos em uma rede homogênea.



Fonte: Elaborada pelo autor.

A classificação transdutiva baseada em rede objetiva minimizar uma função que satisfaz simultaneamente duas premissas: (i) os valores das informações de classes entre vértices vizinhos e fortemente relacionados devem ser próximos e (ii) a diferença entre as classes atribuídas aos vértices rotulados durante o processo de classificação e as suas classes reais, deve ser pequena.

A primeira premissa corresponde a propriedade de suavidade (do inglês, *smoothness*) da distribuição dos rótulos no grafo, que se refere a condição de que se dois objetos são próximos, as suas classes também devem ser próximas. Essas duas premissas podem ser sintetizadas em um *framework* de regularização, no qual o primeiro termo é a função de regularização e o segundo termo é a função de perda (ZHU, 2005). Desta forma, os algoritmos para aprendizado transdutivo baseado em grafos objetivam minimizar o seguinte *framework* de regularização:

$$Q(G) = \frac{1}{2} \sum_{e_{i,j} \in E} W_{i,j} \Omega(R_j, R_i) + \mu \sum_{v_i \in V^l} \Omega'(R_i, Y_i), \tag{2.2}$$

no qual R_i é um vetor l dimensional associado a cada vértice $v_i \in V$, sendo l o número de classes, em que o valor presente em cada dimensão corresponde ao quanto o vértice está associado a uma determinada classe. No segundo termo, Y_i corresponde ao vetor de classes, de dimensão l, associado a um objeto rotulado v_i. A k-ésima dimensão do vetor $Y_{i,k}$ conterá a informação de classe do objeto v_i, de forma que $Y_{i,k} = 1$ e $Y_{i,r} = 0$ para todo $r \neq k$. Na Equação 2.2 o primeiro termo corresponde a primeira premissa e o segundo corresponde a segunda premissa, controlada pelo parâmetro μ. As funções Ω e Ω' são medidas de similaridade ou distância entre os vértices e $W_{i,j}$ corresponde ao peso da aresta entre dois vértices v_i e v_j. Vários métodos baseados em grafos são semelhantes entre si, diferindo na escolha particular da função de perda e do regularizador (ZHU, 2005; SILVA; ZHAO, 2016). Os métodos baseados em regularização podem ser categorizados conforme o tipo de rede (homogênea ou heterogênea) para o qual eles foram modelados, pois é através da estrutura da rede que as informações de rótulo são atribuídas aos vértices não rotulados.

É possível destacar dois métodos, referência na literatura, que executam a classificação transdutiva em grafos homogêneos: (i) *Learning with Local and Global Consistency* (LLGC) (ZHOU *et al.*, 2004) e (ii) *Gaussian fields and Harmonic Functions* (GFHF) (ZHU; GHAH-

RAMANI; LAFFERTY, 2003). Basicamente, estes métodos funcionam como o esquema de propagação de rótulos em um grafo homogêneo, em que os rótulos dados aos vértices do grafo são iterativamente propagados, minimizando o *framework* de regularização da Equação 2.2.

No que se refere a grafos heterogêneos, o método *GNetMine* (JI *et al.*, 2010) é uma generalização do LLGC que considera informações de relações entre diferentes tipos de objetos. Em grafos heterogêneos bipartidos, a propagação de informação de classes é feita entre as duas camadas de vértices. Considerando essa técnica, algoritmos que realizam classificação transdutiva em grafos heterogêneos bipartido foram propostos: (i) *Tag-based Model* (TB) (YIN *et al.*, 2009), (ii) *Transductive Classification based on Bipartite Heterogeneous Network* (TCBHN)(ROSSI; REZENDE; LOPES, 2015) e (iii) *Transductive Propagation in Bipartite Graph* (TPBG) (FALEIROS; ROSSI; LOPES, 2017). A formulação do método de propagação em grafos bipartidos para classificação transdutiva TPBG é descrita com detalhes a seguir, dado que constitui a base teórica para a proposta desta tese.

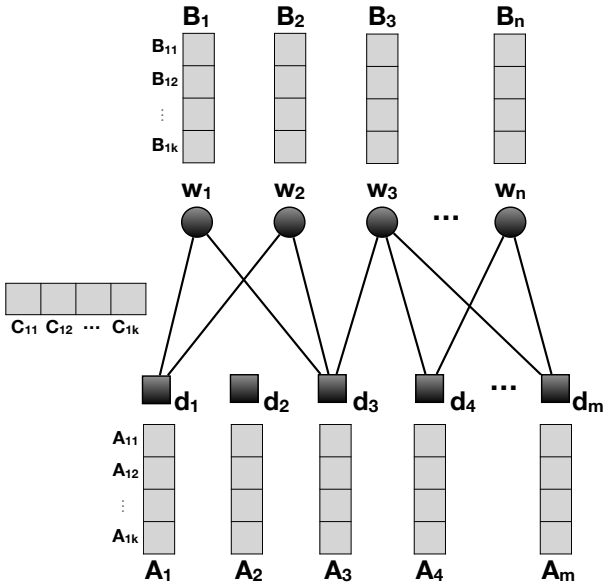
2.3.3 Método de Propagação Transdutiva em Grafos Bipartidos

O aprendizado semissupervisionado é uma opção adequada para muitos problemas de classificação textual, visto que em geral as coleções textuais têm uma pequena fração de documentos rotulados (FALEIROS; ROSSI; LOPES, 2017; ROSSI; LOPES; REZENDE, 2016). Baseado nisto e na vantagem da representação de conjunto de dados textuais em rede, o método de propagação de rótulos em grafos bipartidos *Transductive Propagation in Bipartite Graph* (TPBG) (FALEIROS; ROSSI; LOPES, 2017) foi desenvolvido. Esse método possui sua modelagem baseada nas propriedades do *framework* de regularização definido pela Equação 2.2 e emprega a divergência *Kullback-Leibler* (KL) como medida de similaridade. Observa-se que as definições e formulações matemáticas foram obtidas a partir do trabalho de Faleiros (2016).

O método utiliza como base um grafo bipartido composto por vértices do tipo D e W, no qual uma das partições possui dados rotulados e não rotulados, *i.e* $D = D_l \cup D_u$. Além disto, os vértices $d_j \in D_u$, $w_i \in W$ e as arestas $e_{i,j} \in E$ são associados, respectivamente, a vetores que contêm informação de classes A_j, B_i, $C_{j,i}$. Os vértices rotulados $d_j \in D_l$ são associados a vetores Y_j que contém informações de classe conhecidas *a priori*. Sendo o TPBG um método de aprendizado transdutivo, o objetivo é encontrar uma função $F : D_l \cup D_u \rightarrow Y$. Existem casos nos quais as conexões de um vértice w_i com vértices do tipo d podem ter significados diferentes que remetem a classes distintas. Desta forma, com o intuito de garantir que esse comportamento seja contornado, um vetor com informações de classes $C_{j,i}$ é associado para cada conexão $d_j - w_i$, garantindo que w_i possa propagar diferentes informações de classes para a sua vizinhança e tornando o processo de propagação mais consistente. A Figura 9 ilustra o exemplo de um grafo bipartido e os vetores utilizados no processo de propagação.

O TPBG é fundamentado em dois pressupostos: (i) a minimização da divergência entre as distribuições de probabilidades associadas às dimensões dos vetores dos vértices e arestas do

Figura 9 – Exemplo dos elementos do TPBG para um grafo bipartido.



Fonte: Elaborada pelo autor.

grafo favorece a descoberta dos padrões de classes dos vértices não rotulados D_u; (ii) quanto maior o peso da conexão $f_{j,i}$ entre dois vértices d_j e w_i, maior deve ser a concordância das informações de classes dos vetores $(A_j \odot B_i)$ e $C_{j,i}$ (FALEIROS; ROSSI; LOPES, 2017). Assim, define-se a seguinte função objetivo considerando a divergência KL:

$$Q_G(A,B,C) = \sum_{e_{j,i} \in E} \left(f_{j,i} C_{e_{j,i}} \log \frac{A_j \odot B_i}{C_{e_{j,i}}} \right) + \sum_{d_j \in D} R(A_j, \alpha) + \sum_{d_j \in D^l} Y_j \log \frac{A_j}{Y_j} \tag{2.3}$$

em que $R(A_j, \alpha)$ e α correspondem ao termo de regularização dos vértices d_j e ao parâmetro responsável pela distribuição de informações de classes de A_j, respectivamente. O termo regulador $R(A_j, \alpha)$ é definido como:

$$R(A_j, \alpha) = (\alpha - A_j) \log A_j + A_j (\log A_j - 1). \tag{2.4}$$

Conforme definem as premissas da função de regularização da Equação 2.2, o termo $\sum_{d_j \in D^l} Y_j \log \frac{A_j}{Y_j}$ assegura que as informações de classes preditas e reais são próximas.

Para calcular os vetores de classes, deve-se resolver o problema de otimização dado por:

$$Q(G) = \arg\max_{A^*,B^*,C^*} \sum_{c_k \in \mathbb{C}} [Q_G(A,B,C)]\, k, \tag{2.5}$$

em que a otimização da função Q é determinada pelos vetores A^*, B^*, C^*. Para realizar a otimização, é empregado o método de gradiente descendente, sendo o seu valor máximo em relação a A_j, B_i e $C_{e_{j,i}}$ dado pela direção do gradiente.

O processo de transdução segue três passos nos quais é realizada a maximização da função objetivo em relação aos vetores $C_{e_{j,i}}$, A_j e B_i que geram três equações de atualização que são a base para o algoritmo TPBG. Inicialmente, é maximizada a Equação 2.3 em relação ao vetor $C_{e_{j,i}}$ e como resultado tem-se:

$$C_{e_{j,i}} = \frac{A_j \odot B_i}{\sum_{c_k \in \mathbb{C}} (A_j \odot B_i) k}. \tag{2.6}$$

Em seguida, a Equação 2.3 é maximizada em relação a A_j, resultando em:

$$A_j = \alpha + \sum_{w_i \in W_{d_j}} f_{j,i} C_{e_{j,i}}. \tag{2.7}$$

Por fim, a Equação 2.3 é maximizada em relação a B_i, gerando:

$$\hat{B}_{i,k} = \sum_{j=1}^{D} \sum_{k}^{K} f_{j,i} C_{e_{j,i}}. \tag{2.8}$$

Aplicando uma normalização no valor de $\hat{B}_{i,k}$ sobre todos os vértices $w \in W$, tem-se

$$B_{i,k} = \frac{\hat{B}_{i,k}}{\sum_{p \in W} \hat{B}_{p,k}} \tag{2.9}$$

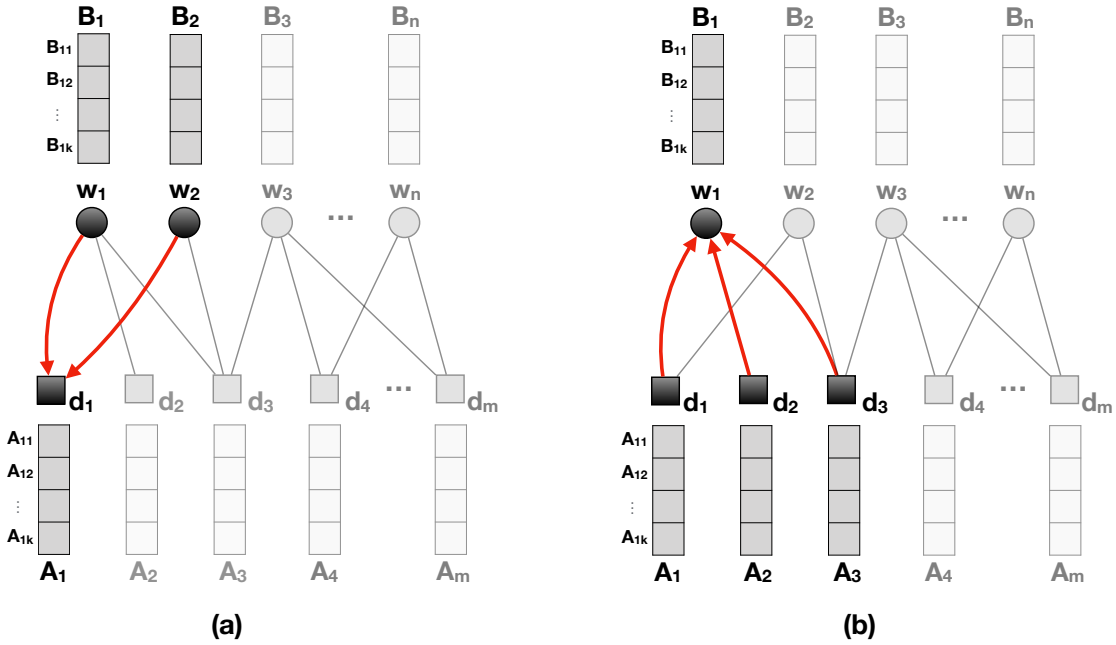
O Algoritmo TPBG

As Equações 2.6, 2.7 e 2.8 são fundamentais para o algoritmo TPBG, visto que são utilizadas no processo de atualização dos vetores de classes A_j, B_i e $C_{e_{j,i}}$. O algoritmo TPBG propaga iterativamente as informações de classe dos vértices por intermédio das suas vizinhanças, compostas por vértices das camadas opostas. Por exemplo, dado um vértice $d_j \in D$, o conteúdo do seu vetor de classe é propagado para a sua vizinhança em W. Com isso, o algoritmo possui duas categorias de iteração para atualização: (i) local, no qual as informações são propagadas nas vizinhanças dos vértices, como ilustrado na Figura 10 (a); (ii) global, em que as informações de classe são propagadas em toda a estrutura do grafo, conforme exemplificado na Figura 10 (b).

Para descrição do algoritmo TPBG, considere as seguintes entradas: grafo bipartido $G = (D, W, E)$, o conjunto de vértices rotulados $D_l \in D$ e o parâmetro α. A inicialização dos vetores de classe é a etapa inicial do processo. Os vetores de classes correspondentes aos vértices d já rotulados são inciados de forma que $A_j = Y_j$, enquanto os vetores dos vértices $d_j \in D_u$ e $w_i \in W$ são inicializados de forma aleatória, de forma que a soma das informações dos vetores é igual a 1. Os processos de propagação local e global são realizadas para que cada vértice d_j e w_i receba as informações de sua vizinhança, como descrito no Algoritmo 3.

Na propagação local, um vértice d_j recebe informação de todos as arestas $e_{j,i}$ incidentes nele através da criação de um vetor l-dimensional $C_{e_{j,i}}$ normalizado como $\sum_{c_k \in C} C_{e_{j,i},k} = 1$. Tal

Figura 10 – Exemplo dos processos de propagação do TPBG. (a) propagação local no grafo bipartido e (b) propagação global.



Fonte: Elaborada pelo autor.

Algoritmo 3 – Algoritmo TPBG

procedimento TPBG(G, D^l, α) ▷ G grafo bipartido, D^l conjunto de vértices do tipo d rotulados, α parâmetro de concentracão
  Inicia vetor A_j para cada vértice $d_j \in D$;
  Inicia vetor B_i para cada vértice $w_i \in W$;
  enquanto não convergir **faça**
    para todo $d_j \in D$ **faça**
      repita
        $A_j \leftarrow$ `localPropag`(G, d_j, A_j, B, D^l);
      até A_j não converge;
    fim para
    $B \leftarrow$ `globalPropag`(G, A, B);
  fim enquanto
  para todo $d_j \in D^u : \{Y_{j,k} = 1 | k = argmax_{\hat{k}=1}^{l} A_{j,\hat{k}}\}$;
  retorna Y ▷ rótulos atribuídos a cada vértice D^u
fim procedimento

vetor é resultando do produto Hadamard entre os vetores A_j e B_i. Esse vetor é utilizado para atualizar o vetor A_j, como descrito na Equação 2.7. Caso contrário, para todos os vértices d_j do conjunto D_l, tem-se que $A_j = Y_j$.

Para cada vértice d_j, o processo de propagação local é repetido enquanto houver mudanças nos valores de A_j ou até atingir número máximo de iterações previamente determinado, como descrito no Algoritmo 4.

Algoritmo 4 – Propagação Local do algoritmo TPBG

procedimento LOCALPROPAG(G, d_j, A_j, B, D^l)
  para todo aresta $e_{j,i}$ incidente em d_j **faça**
    $C_{e_{j,i}} = \frac{A_j \odot B_i}{\sum_{c_k \in \mathbb{C}} (A_j \odot B_i)_k}$;
  fim para
  se $d_j \in D^l$ **então**
    $A_j \leftarrow Y_j$;
  senão
    $A_j = \alpha + \sum_{w_i \in W_{d_j}} f_{j,i} C_{e_{j,i}}$;
  fim se
  retorna A_j;
fim procedimento

Após a execução da propagação local para os vértices d_j, a propagação global é executada para todas as arestas $e_{j,i}$ incidentes em cada vértice $w_i \in W$, criando um vetor l-dimensional $C_{ej,i}$ normalizado como $\sum_{c_k \in C} C_{e_{j,i},k} = 1$. Esse vetor também é obtido por meio do produto Hadamard de A_j e B_i, e é utilizado para atualizar o vetor B_i, como na Equação 2.8. Finalmente, é realizado o processo de normalização do vetor B_i para todos os vértices $w_p \in W$. Esse processo de propagação global é descrito no Algoritmo 5.

Algoritmo 5 – Propagação Global do algoritmo TPBG

procedimento GLOBALPROPAG(G, A, B)
  para todo vértice $w_i \in W$ **faça**
    para todo aresta $e_{j,i}$ incidente em w_i **faça**
      $C_{e_{j,i}} = \frac{A_j \odot B_i}{\sum_k (A_j \odot B_i)_k}$;
    fim para
    $B_i \leftarrow \sum_{d_j \in D} f_{j,i} C_{e_{i,j}}$;
  fim para
  para todo vértice $w_i \in W$ **faça**
    para todo $c_k \in \mathbb{C}$ **faça**
      $B_{i,k} = \frac{\hat{B}_{i,k}}{\sum_{p \in W} \hat{B}_{p,k}}$;
    fim para
  fim para
  retorna B;
fim procedimento

2.4 Aprendizado Não Supervisionado

O aprendizado não supervisionado é um processo conduzido pela própria distribuição de dados para inferir conhecimentos desejados sem a utilização de informações prévias relacionadas à classe ou categoria. Este conceito é utilizado, principalmente, na tarefa de agrupamento de dados que consiste na criação de grupos significativos de exemplos de dados que compartilham

caraterísticas e padrões semelhantes. Outra vertente do aprendizado não supervisionado é a redução de dimensionalidade e extração de características, em que é gerada uma representação com dimensão reduzida que condensa o máximo de informações possível, visando representar eficientemente as informações originais. Atualmente, devido ao surgimento de conjuntos de dados de aplicações em grande escala, torna-se cada vez mais importante e desafiador gerar representações que condensem as informações e conservam as características intrínsecas e latentes dos dados Sun *et al.* (2020). Deste modo, esta vertente tornou-se uma questão importante para a área. Na última década, diversos algoritmos tradicionais de extração de características e aprendizado de representação, foram propostos no contexto do aprendizado de máquina (SUN *et al.*, 2020).

Uma solução clássica consiste na engenharia de atributos manual em que são construídos *pipelines* de pré-processamento e transformações de dados que resultam em uma representação dos dados que pode auxiliar na melhora da generalização dos algoritmos de aprendizado de máquina (BENGIO; COURVILLE; VINCENT, 2013). Muitas vezes este processo depende de conhecimento especializado sobre problema e os atributos obtidos são geralmente projetados para tarefas específicas e não se generalizam em diferentes abordagens (GROVER; LESKOVEC, 2016). Nos últimos anos, houve um crescente interesse em Redes Neurais devido à sua capacidade de aprender múltiplas transformações não lineares a partir de dados. Neste sentido, emergiu um novo paradigma de extração de características, denominado de aprendizado de representação (do inglês *representation learning*), visando o desenvolvimento de modelos que sejam capazes não apenas de mapear os dados para características discriminativas, mas também de aprender automaticamente novas representações.

2.4.1 Aprendizado Não supervisionado de representação em espaço-vetorial

Algoritmos tradicionais usados em análise exploratória de dados, realizam a extração de atributos através de técnicas de álgebra linear para projetar um conjunto de dados de alta dimensão para uma representação de baixa dimensão. *Principal Components Analysis* (PCA) (PEARSON, 1901) é uma técnica matemática que realiza uma transformação ortogonal para descrever um conjunto de observações de variáveis em termos de suas dimensões inerentes, denominadas componentes principais. De forma similar, o método *Multi-Dimensional Scaling* (MDS) extrai a partir de uma matriz de distância, ou similaridade, uma representação de menor dimensão que conserva o máximo possível das distâncias originais entre os pontos de dados.

A modelagem de tópicos é advinda da necessidade de extração de informação a partir de coleções textuais, caracterizadas por serem dados não estruturados. Normalmente, um conjunto de textos pode ser estruturado em uma matriz documento-termo $M_{d,t}$, denominada *Bag-of-words* (BOW) (AGGARWAL; ZHAI, 2012), que contabiliza o número de ocorrências de cada termo em cada documento. Um aspecto característico da representação BOW é a alta dimensionalidade

e esparsidade, uma vez que cada dimensão corresponde a um termo presente na coleção textual. Assim, a redução de dimensionalidade por meio de métodos de extração de tópicos pode ser aplicada a fim de encontrar um espaço semântico correspondente com a representação BOW original, em que cada dimensão corresponda a um tópico (CRAIN *et al.*, 2012).

Modelos de tópicos são modelos estatísticos que analisam automaticamente coleções textuais de modo a encontrar estruturas semânticas latentes presentes nos dados. Um tópico pode ser definido um conjunto de termos que ocorrem frequentemente em documentos relacionados. Assim, os métodos de extração de tópicos visam identificar padrões latentes que caracterizem as relações entre documentos e termos e permitam anotar, ou agrupar, os documentos em determinados temas (BLEI, 2012). Dado um conjunto de k tópicos para a representação de uma coleção de textos, o processo de extração de tópicos transforma uma matriz documento-termo M, formada pelos conjuntos de D documentos e T termos, em: (i) uma matriz documento-tópico de tamanho $|D| \times k$, em que cada linha armazena a pertinência de cada documento aos tópicos (ii) uma matriz termo-tópico de tamanho $|T| \times k$, na qual as linhas representam relevância de cada termo aos tópicos. Do conjunto de métodos tradicionais de detecção de tópicos destacam-se *Non-negative Matrix Factorization* (NMF) e *Latent Dirichlet Allocation* (LDA). Em geral, as técnicas para extração de tópicos não necessitam da presença de informação de rótulos. Entretanto, existem versões supervisionadas dos métodos de extração de tópicos que consideram os rótulos dos dados.

Nos últimos anos houve um crescente interesse por extração de representação baseada em Redes Neurais Artificias, do inglês *Artificial Neural Network* (ANN). As técnicas de aprendizado profundo constituem uma categoria, em que as ANN são compostas por diversas camadas ocultas. Para extrair relacionamentos complexos intrínsecos dos dados, estas técnicas baseiam-se no aprendizado de vários níveis, ou camadas, de representações a partir dos dados originais, formando uma representação hierárquica, denominada arquitetura profunda (DENG; YU, 2014). Basicamente, a primeira camada corresponde aos dados de entrada e a saída da camada final produz a representação dos dados originais em um espaço de baixa dimensão. Entre a entrada e a saída final, a rede é formada por várias camadas intermediárias. Neste caso, cada nível intermediário usa como entrada a representação produzida pelo nível anterior e produz novas representações como saída, utilizadas pelos níveis subsequentes. Em muitas redes neurais profundas, os primeiros níveis geram representações mais abstratas e gerais, enquanto à medida que os níveis finais são calculados, representações mais específicas do problema em questão são obtidas. Assim, as saídas das camadas intermediárias também podem ser vistas como representações dos dados de entrada.

Autoencoders é uma técnica baseada em ANNs para o aprendizado de representações eficientes e compactas de um conjunto de dados X através de uma arquitetura composta por uma parte que compacta os dados de entrada em uma nova representação, denominada codificador (do inglês, *encoder*), e outra parte que reconstrói a representação gerada para uma saída que

corresponde aos dados de entrada, denominada decodificador (do inglês, *decoder*). O codificador pode ser definido como uma função f_θ que permitirá o cálculo simples e eficiente de um vetor de características $h = f_\theta(x)$ para uma entrada x. Assim, para cada exemplo x_t de um conjunto de dados $X = \{x_1, ..., x_t\}$ pode ser definido $h = f_\theta(x_t)$, em que h_t é o vetor de características ou representação de x_t. O decodificador é dado por uma função g_θ que mapeia do espaço de recursos para o espaço de entrada, produzindo uma reconstrução $r = g_\theta(h)$. A arquitetura do *Autoencoder* é considerada profunda quando a quantidade de camadas ocultas é maior que um (DENG; YU, 2014).

2.4.2 Aprendizado Não Supervisionado de representação em redes

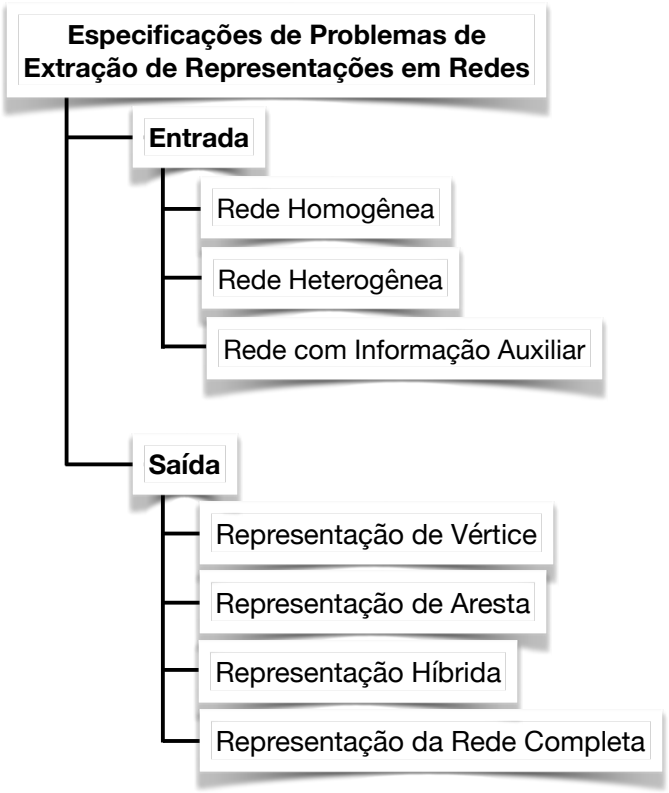
O reconhecimento de padrões em redes complexas emergiu como uma alternativa importante para diferentes aplicações, pois fornece meios para a compreensão sobre o que está por trás dos relacionamentos entre os dados. Esta questão é muito relevante para a realização de várias tarefas como, classificação, recomendação, identificação de módulos ou grupos. Neste contexto, uma alternativa básica é realizar a caracterização das redes para codificar em um vetor de atributos informações sobre a estrutura de vizinhança e características topológicas globais e locais de cada exemplo de um conjunto de dados (HAMILTON; YING; LESKOVEC, 2017)

Métodos tradicionais de extração de representação têm como limitação não capturar com êxito a estrutura e o padrão de relacionamento de dados complexos e não lineares (SUN *et al.*, 2020). No início dos anos 2000, os pesquisadores iniciaram o desenvolvimento de algoritmos de extração de representação em grafos para a redução de dimensionalidade de dados estruturados no formato atributo-valor (GOYAL; FERRARA, 2018). Nos últimos anos, houve o fortalecimento da pesquisa de métodos não supervisionados para o aprendizado de representação em redes complexas, visando usufruir das relações naturais que compõem conjunto de dados complexos não lineares (CAI; ZHENG; CHANG, 2018; SUN *et al.*, 2020). No cenário de Aprendizado de Representação em Rede (ARR), o objetivo é aprender uma representação vetorial eficiente que extraia informações latentes e preserve propriedades da topologia original da rede (SUN *et al.*, 2020; DONG *et al.*, 2020).

O ARR não é uma tarefa trivial, visto que cada configuração específica de rede necessita ser analisada de uma forma particular (CAI; ZHENG; CHANG, 2018). Cai, Zheng e Chang (2018) agrupa essas particularidades em dois níveis: o tipo da rede utilizada como entrada e tipo de saída desejada. A Figura 11 sintetiza as abordagens dos problemas de aprendizado de representação em redes. Dentre as possibilidades de entrada destacam-se as redes homogêneas, heterogêneas e redes compostas por informações adicionais associadas aos seus vértices como, por exemplo, vetores de atributos e meta-dados (CAI; ZHENG; CHANG, 2018; SUN *et al.*, 2020; YANG *et al.*, 2020). Diferentes tipos de redes carregam propriedades específicas que precisam ser consideradas no processo de aprendizado, pois devem ser preservadas e encapsuladas na representação gerada. Em relação à saída gerada, diferentes aspectos das redes podem

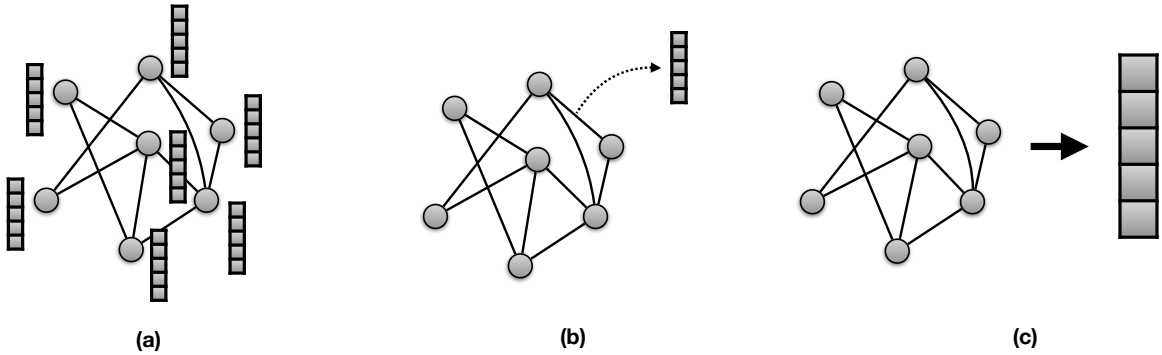
ser convertidos em uma representação de baixa dimensão, conforme o objetivo do problema específico estudado. Portanto, no que se refere aos tipos de saída pode-se destacar as extrações de representações para de vértices, arestas e grafo inteiro, conforme ilustrado na Figura 12.

Figura 11 – Diagrama de especificações de entradas e saídas de problemas de extração de representação em redes.



Fonte: Adaptada de Cai, Zheng e Chang (2018).

Figura 12 – Formas de extração de representação em redes homogêneas: (a) representações para os vértices de um grafo, (b) representações para as arestas de um grafo e (c) representações para um grafo inteiro.



Fonte: Elaborada pelo autor.

Nesta tese é abordado apenas o contexto da extração de representação para os vértices de uma rede, sem considerar informações adicionais, como atributos de vértices e meta-dados.

Portanto, nesta seção são abordados métodos que se encaixam nesse contexto. Formalmente, o objetivo do aprendizado de representação para os vértices de uma rede é aprender uma função de mapeamento $f : v_i \mapsto r_v \in \mathbb{R}^d$ em que r_v é o vetor de representação aprendido, de tamanho d, para o vértice v_i. Geralmente, medidas de proximidade são adotadas para capturar as propriedades do grafo que são embutidas em uma nova representação (CAI; ZHENG; CHANG, 2018).

A proximidade de primeira ordem: corresponde a proximidade local entre pares de vértices conectados diretamente por arestas. Tomando como exemplo uma rede ponderada, a proximidade de primeira ordem entre os vértices é dada pelos pesos das arestas que os conectam. Assim, se dois vértices estão ligados por uma aresta, eles devem estar próximos um do outro no espaço latente de baixa dimensão (CHEN *et al.*, 2020).

A proximidade de segunda ordem: captura a similaridade das estruturas de vizinhança dos vértices. Desta forma, vértices com vizinhanças semelhantes tendem a ficar próximos no espaço de representação de baixa dimensionalidade (SUN *et al.*, 2020), mesmo que não haja aresta direta entre os vértices.

A proximidade de ordem superior: caracteriza a proximidade das vizinhanças não imediatas a dois vértices. Assim, pode-ser perceber que as proximidades de ordem superior e de segunda ordem capturam a semelhança entre um par de vértices indiretamente conectados, que compartilham estruturas de vizinhanças semelhantes (ZHANG *et al.*, 2018).

O *Isomap* (TENENBAUM; SILVA; LANGFORD, 2000) é um método clássico baseado em MDS, que visa preservar a geometria intrínseca dos dados. A partir da construção de um grafo KNN, são calculados os caminhos mais curtos entre os vértices, considerados como as distâncias geodésicas, utilizados pelo algoritmo MDS que obtém uma nova representação de baixa dimensão. De outro modo, *Laplacian Eigenmaps* (LE) (BELKIN; NIYOGI, 2003) preserva as informações locais das vizinhanças dos vértices, de forma que vértices muito similares estejam próximos uns dos outros no espaço de baixa dimensão (CAI; ZHENG; CHANG, 2018). A partir da matriz laplaciana obtida pela geração de um grafo ε-vizinhança ou KNN o método mapeia os dados para uma nova representação através do cálculo de autovetores do grafo.

O conceito de caminhada aleatória é utilizado por vários métodos de ARR e pode ser definido como um processo que gera trajetórias aleatórias em um grafo. Desta forma, partindo de cada vértice v contido em um grafo G são realizadas n caminhadas aleatórias de tamanho l através das vizinhanças. Dado um vértice v_i no passo t, um vértice v_j na sua vizinhança é visitado no passo $t+1$, com base em uma função de probabilidade $p(i \to j)$. As caminhadas aleatórias geradas em um grafo podem ser utilizadas para a extração de representação dos vértices de uma rede, visto que contém informações significativas sobre a estrutura local e global das vizinhanças presentes em um grafo.

O método *DeepWalk* (PEROZZI; AL-RFOU; SKIENA, 2014) utiliza as caminhadas aleatórias extraídas de um grafo como entrada de uma ANN para produzir as novas representações dos vértices. Este processo pode ser entendido como uma versão da técnica *Word2vec* (MIKOLOV *et al.*, 2013) para redes, em que os passeios aleatórios gerados a partir de cada vértice $v_i \in V$ são utilizados para criar sequências de vértices, como se fossem sentenças em um *corpus*. Analogamente, o método *node2vec* (GROVER; LESKOVEC, 2016) foi proposto. A caminhada aleatória empregada pelo *DeepWalk* é um caso especial do *node2vec*, em que os parâmetros p e q são iguais a 1, e ao considerar a variação dos valores desses dois parâmetros o *node2vec* aprende características mais robustas em comparação ao *DeepWalk*, pois o método pode ajustar a preferência em enfatizar as informações locais ou globais da estrutura do grafo (SUN *et al.*, 2020).

Estes métodos produzem representações pouco discriminativas para vértices de redes heterogêneas, uma vez que não consideram em suas formulações a existência de diferentes tipos de vértices e suas relações. Assim, o método *metapath2vec* (DONG; CHAWLA; SWAMI, 2017) utiliza caminhadas aleatórias baseada em meta-caminhos e redes heterogêneas de modo a capturar as informações vizinhanças que consideram a semântica das relações entre vértices de diferentes tipos e, em seguida, emprega um modelo heterogêneo de *skip-gram* para extrair as representações dos vértices. O método *Heterogeneous Network Embedding* (HNE) (CHANG *et al.*, 2015) utiliza uma arquitetura profunda com múltiplas camadas não lineares para capturar as interações complexas em uma rede heterogênea.

2.4.3 Método de Propagação Não Supervisionado em Grafos Bipartidos

O método *Propagation in Bipartite Graph* (PBG), é um algoritmo não supervisionado para problemas de extração de tópicos e agrupamento. O PBG não atribui um simples rótulo para cada vértice do grafo, como o método propagação transdutiva TPBG, ao invés disso são propagados vetores k-dimensionais A_j, B_i e $C_{e_{j,i}}$ que guardam informações latentes da estrutura de uma rede bipartida, em que o valor k é definido *a priori*. Através de um processo iterativo, o vetor com as informações latentes relacionadas a um vértice é propagado para os vetores dos seus vértices vizinhos. Este processo é composto pelas propagações globais e locais, assim como na versão transdutiva.

O algoritmo PBG utiliza a divergência KL como medida de similaridade entre as informações latentes dos vértices e arestas. O raciocínio que fundamenta o algoritmo é que quanto maior o peso de conexão $f_{j,i}$ entre dois vértices, maior deve ser a concordância entre a informação latentes A_j, B_i e $C_{e_{j,i}}$ para todos os vértices $d_j \in D$, $w_i \in W$ e arestas $e_{j,i} \in E$. Deste modo, é

definida a seguinte função de maximização:

$$Q_G(A,B,C)_k = \sum_{e_{j,i} \in E} \left(f_{j,i} C_{e_{j,i},k} \log \frac{[A_j \odot B_i]_k}{C_{e_{j,i},k}} \right) + \sum_{d_j \in D} R(A_j, \alpha) \tag{2.10}$$

em que $R(A_{j,k}, \alpha)$ é o termo regularizador submetido a cada vértice d_j, como definido na Equação 2.4 e a constante α controla a concentração das informações latentes no vetor A_j.

Os vetores de informações latentes para todo o conjunto de vértices do grafo bipartido podem ser obtidos otimizando esta equação para cada par de vértices ligados por uma aresta, resultando na seguinte função de custo para um grafo G:

$$Q(G) = \arg\max_{A^*,B^*,C^*} \sum_{k \in \mathbb{K}} [Q_G(A,B,C)]_k \,, \tag{2.11}$$

A inferência das informações latente de $Q(G)$ é realizada pelo método gradiente descendente. O máximo de $Q(G)$ em relação à A_j,B_i e $C_{e_{j,i}}$, para todos os vértices $d_j \in D$, $w_i \in W$ e pares de arestas $e_{j,i} \in E$, é determinado configurando o gradiente para zero. A maximização da Equação 2.10 em relação aos vetores $C_{e_{j,i}}$, A_j e B_i geram, respectivamente, três equações de atualização Equação 2.6, 2.7 e 2.8 que são a base para o algoritmo PBG.

O Algoritmo PBG

O algoritmo PBG tem como entrada um grafo bipartido $G = (D,W,E)$, tamanho dos vetores K e o parâmetro de concentração α dos valores em A_j. Os vetores A_j e B_i com informações latentes associados, respectivamente, aos vértices do tipo d e w são inicializados aleatoriamente de forma que $\sum_{k=1}^{K} A_{j,k} = 1$ e $\sum_{k=1}^{K} B_{i,k} = 1$. Os processos de propagação local e global são executados para que, respectivamente, cada vértice d_j e w_i receba as informações de sua vizinhança, como mostra o Algoritmo 6. Observa-se que, nesta tese, foi acrescentado o parâmetro β, que controla a concentração dos valores em B_i, na Equação 2.8. Portanto, o parâmetro β também foi considerado nos experimentos desta tese.

O processo de propagação local é realizado em cada aresta $e_{j,i}$ incidente em um vértice d_j, que recebe um vetor K-dimensional $C_{e_{j,i}}$ produzido pelo produto Hadamard entre os vetores A_j e B_i, normalizado de maneira que $\sum_{k}^{K} C_{e_{j,i},k} = 1$. Os vetores $C_{e_{j,i}}$ relacionados de todas as arestas $e_{j,i}$ conectando ao vértice d_j são utilizados para a atualização do vetor A_j. Esse esquema de propagação deve ser repetido para cada vértice d_j ao passo que os valores do vetor A_j se alteram, como mostrado no Algoritmo 7. Diferentemente da versão transdutiva semissupervisionada, informações de classe pré-definidas não são utilizadas no durante processo de propagação local.

O procedimento de propagação global é efetuado em todo vértice $w_i \in W$ através de cada aresta $e_{j,i}$ que indice em w_i partindo dos seus vértices vizinhos em $d \in D$, conforme definido no Algoritmo 5. Esse processo gera um vetor K-dimensional $C_{e_{j,i}}$ dado pelo produto Hadamard

Algoritmo 6 – Algoritmo PBG

procedimento PBG(G, k, α) ▷ G grafo bipartido, dimensões de A_j e B_i, α parâmetro de concentração
    Inicia vetor A_j para cada vértice $d_j \in D$;
    Inicia vetor B_i para cada vértice $w_i \in W$;
    enquanto não convergir **faça**
        para todo $d_j \in D$ **faça**
            repita
                $A_j \leftarrow$ `localPropag`(G, d_j, A_j, B);
            até A_j não converge;
        fim para
        $B \leftarrow$ `globalPropag`(G, A, B);
    fim enquanto
    retorna A e B
fim procedimento

Algoritmo 7 – Propagação Local do algoritmo PBG

procedimento LOCALPROPAG(G, d_j, A_j, B)
    para todo aresta $e_{j,i}$ incidente em d_j **faça**
        $C_{e_{j,i}} = \frac{A_j \odot B_i}{\sum_k^K (A_j \odot B_i)_k}$;
    fim para
    $A_j = \alpha + \sum_{w_i \in W_{d_j}} f_{j,i} C_{e_{j,i}}$;
    retorna A_j;
fim procedimento

de A_j e B_i, normalizado de modo que $\sum_k^K C_{e_{j,i},k} = 1$. Os vetores $C_{e_{j,i}}$ associados de todas as arestas $e_{j,i}$ que incidem ao vértice w_i são utilizados para a atualização do seu vetor B_i. Após este processo, os vetores B_i são normalizados sobre todos os vértices $w \in W$.

CAPÍTULO

3

PROPAGAÇÃO EM REDES K-PARTIDAS PARA SISTEMAS DE RECOMENDAÇÃO

Atualmente, os sistemas de recomendação desempenham uma importante função em diferentes sistemas digitais, pois têm a capacidade de prever a preferência dos usuários com base em perfil de consumo. Neste capítulo, são apresentadas propostas para os problemas de recomendação de *tags* e itens em sistemas colaborativos que possuem a informação de atribuição de *tags* em itens por parte dos usuários. Nas Seções 3.1 e 3.2 são introduzidos os conceitos de sistemas de recomendação e sistemas colaborativos de marcação que embasam o problema abordado e são identificados os desafios mais predominantes. Em seguida, nas Seções 3.3 e 3.4 são apresentadas duas abordagens desenvolvidas nesta tese baseadas na propagação em grafos k-partidos para recomendação de *tags* e itens, respectivamente. Juntamente, são apresentados os trabalhos relacionados, a avaliação e discussão do desempenho das propostas em relação aos métodos de recomendação referências do estado da arte no contexto abordado.

3.1 Sistemas de Recomendação

O rápido crescimento da importância da Web e a intensificação na introdução de novos serviços digitais devido à expansão da Internet sobrecarregam os usuários com uma grande variedade de serviços, produtos e meios de acesso a diferentes tipos de conteúdos (RICCI; ROKACH; SHAPIRA, 2011; AGGARWAL *et al.*, 2016; CHEN *et al.*, 2018). Como consequência, tem-se a necessidade do desenvolvimento de serviços personalizados que prevejam conteúdos interessantes para diferentes perfis de usuários. Aliado a isso, a facilidade com que a Web permite que os usuários forneçam *feedbacks* sobre suas preferências impulsiona o desenvolvimento dos sistemas de recomendação (AGGARWAL *et al.*, 2016).

Os sistemas de recomendação podem ser definidos como técnicas que, através de informações comportamentais e *feedbacks*, fornecem aos usuários sugestões de itens que possam ser

interessantes (RICCI; ROKACH; SHAPIRA, 2011; CHEN *et al.*, 2018). Portanto, os algoritmos de recomendação têm como função estimar um grau de interesse que um dado usuário alvo pode ter por itens com os quais ele ainda não tem associação. Nesta tese, o termo "item" é usado para denotar o que um determinado sistema recomenda aos seus usuários, como livros, vídeos, músicas, conteúdos, produtos em geral etc. As recomendações geradas são fundamentais para auxiliar diferentes processos de tomada de decisão que envolvem usuários em diferentes situações como, por exemplo, quais filmes assistir, quais artigos ler e quais produtos comprar (RICCI; ROKACH; SHAPIRA, 2011).

O interesse de um usuário em um item pode ser representado de diversas formas e assumir diferentes valores, dependendo do sistema em questão. Por exemplo, os clientes de um determinado *site* podem ser incentivados a darem *feedbacks* dos produtos comprados através de notas. Assim, as avaliações que constituem o perfil de preferência dos usuários, podem ser obtidas de forma implícita ou explícita. Na forma implícita, o interesse de um usuário por itens é capturado pelo seu comportamento no sistema (AGGARWAL *et al.*, 2016; CHEN *et al.*, 2018), como o clique na página de um produto ou até mesmo a compra. No caso da forma explícita, o usuário informa ao sistema a sua avaliação sobre os itens (produtos, filmes, músicas, artigos, *posts* etc). Segundo Aggarwal *et al.* (2016), as avaliações podem ser:

- **Numéricas:** geradas pela avaliação de um item através de uma nota (exemplo, nota de 0 a 5) para expressar a satisfação e interesse do usuário;
- **Ordinais:** obtidas através da escolha de um termo que indique a opinião do usuário sobre um item, como {"ruim", "bom", "excelente"};
- **Binárias:** quando um usuário decide se tem interesse ou não pelo item;
- **Unárias:** popular nas redes sociais, como o *Facebook*, são geradas por uma ação do usuário sobre um item, como o clique no botão "Curtir".

Assim, os dados de preferência dos usuários podem ser representados como uma matriz usuário-item $P_{n\times q}$, em que n denota o número de usuários, q o número de itens e $p_{u,i}$ registra a informação (valor) da preferência do usuário u por um item i. Em geral, as avaliações são convertidas para valores numéricos. Formalmente, a partir de um conjunto de usuários U e itens I, o problema central de um sistema de recomendação é a definição de uma função $r : U \times I \rightarrow R$ que estima o quão útil um item i pode ser para o usuário u, em que R é um conjunto de itens ordenado de acordo com o valor da utilidade potencial de cada item (ADOMAVICIUS; TUZHILIN, 2005). Usualmente, o resultado da recomendação é definido como uma lista L, denominada "top$-k$", que contém os k itens com maiores chances de serem interessantes para um usuário alvo u. Para tanto, são utilizadas as informações de preferências dos usuários conhecidas previamente, dado por um conjunto de pares (u,i). Portanto, o objetivo é encontrar para cada usuário $u \in U$ um item $i \in I$ que maximiza a utilidade da recomendação.

Os modelos tradicionais de sistemas de recomendação podem ser categorizados de acordo com o tipo de informação utilizada e como a recomendação é gerada: filtragem colaborativa, Baseado em Conteúdo e Híbridos. Em suma, os sistemas de filtragem colaborativa utilizam somente o histórico de preferências dos usuários e, por outro lado, os sistemas de filtragem baseados em conteúdo fundamentam-se nas informações dos atributos das entidades (itens e usuários). Abaixo são sintetizadas as características de cada categoria de sistema de recomendação:

Filtragem colaborativa (FC): é baseada nas informações fornecidas pelos usuários para que sejam recomendados itens considerados mais comuns entre os padrões de preferências dos usuários. Portanto, a FC utiliza os níveis de similaridade entre pares de usuários ou itens para gerar as recomendações. Esta categoria será detalhada abaixo, pois é utilizada como base de uma das abordagens propostas neste capítulo descrita na Seção 3.3 e 3.4.

Sistemas de recomendação baseados em conteúdo: relacionam os conteúdos das preferências e escolhas passadas de um usuário alvo com as informações dos possíveis itens a serem recomendados para fornecer recomendações. Assim, os itens recomendados são os que possuem conteúdo semelhante ao dos itens previamente associados a um usuário alvo (LÜ *et al.*, 2012). Esta abordagem considera apenas o perfil de preferências do usuário alvo e, portanto, não utiliza as informações sobre relações de preferências entre usuários para determinar o conteúdo a ser recomendado. No entanto, depende de que os itens possuam uma quantidade de informação significativa que possa ser obtida automaticamente (ZHANG; ZHOU; ZHANG, 2011). Em geral, este tipo de recomendação possui uma arquitetura baseada em três etapas (LOPS; GEMMIS; SEMERARO, 2011): (i) extração e estruturação de informações relevantes em atributos para representar o conteúdo dos itens; (ii) generalização sobre as preferências do usuário que possibilite a inferência de um perfil, ou modelo, de interesses do usuário; (iii) recomendação de itens relevantes com base no perfil do usuário.

Sistemas de recomendação Híbridos: combinam os diferentes aspectos da filtragem colaborativa e sistemas baseados em conteúdo com o intuito da criação de técnicas que podem ter um desempenho mais robusto em aplicações com diferentes particularidades (AGGARWAL *et al.*, 2016). A combinação pode ser feita de diferentes formas, como a combinação dos resultados de recomendação gerados separadamente pelas técnicas e o desenvolvimento de métodos que unificam as características de ambas as técnicas. Essa combinação é geralmente aplicada para solucionar o problema de *cold start* (CHEN *et al.*, 2018).

3.1.1 Filtragem Colaborativa

A técnica de filtragem colaborativa (FC) utiliza as observações dos interesses de diferentes usuários pelos itens. A concepção principal dos métodos de FC é baseada no fator

colaborativo das classificações fornecidas por vários usuários, pois as informações dos históricos de preferências tendem a ser altamente correlacionadas entre vários usuários e itens (AGGARWAL *et al.*, 2016). Assim, a FC faz recomendações para um usuário alvo utilizando as informações do histórico de preferência sem depender da análise do conteúdo e as características dos itens. Como ilustrado na Figura 13, em geral, a FC é composta por três etapas principais: modelagem dos dados, inferência e geração da recomendação final.

Os métodos de filtragem colaborativa podem ser categorizados em duas classes principais (AGGARWAL *et al.*, 2016; CHEN *et al.*, 2018): baseados em memória e baseados em modelo. Os algoritmos de FC baseados em memória realizam a identificação de novas associações entre usuário-item com base em suas vizinhanças. Para isso, medidas de similaridades são aplicadas na matriz de preferência dos usuários para que os usuários e itens mais próximos sejam identificados. Nesse sentido, a FC baseada em memória pode ser especificada em duas categorias conforme as informações são exploradas: baseada em usuário (do inglês, *user-based*) e baseada em item (do inglês, *item-based*).

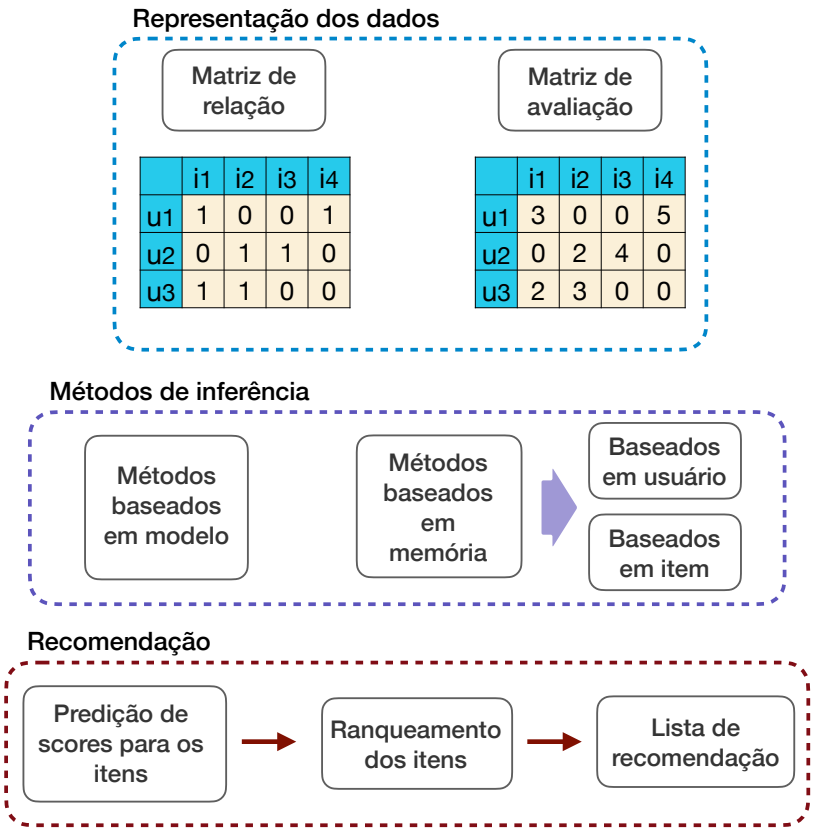
- **Baseada no usuário:** utiliza o conceito de que usuários com perfis de preferências similares tendem a ter interesses semelhantes (CHEN *et al.*, 2018). Assim, a recomendação para um usuário alvo é baseada na preferência dos usuários semelhantes, ou seja, os que compõem a sua vizinhança. Para tanto, é calculada a similaridade entre o perfil de preferência de um usuário alvo e os demais usuários para que os vizinhos com mais semelhança sejam selecionados. A partir disso, o histórico de preferência da vizinhança é utilizado para identificar potenciais itens para o usuário alvo.

- **Baseada no item:** baseia-se na ideia de que itens similares aos que estão presentes no perfil de preferência de um usuário alvo podem ser relevantes e complementar a experiência do usuário no sistema. Para isso, é calculada a similaridade entre cada item previamente avaliado pelo usuário alvo e os demais para gerar um conjunto S de itens mais semelhantes, segundo o histórico de preferência dos outros usuários pelos itens. Assim, os itens com mais avaliações em comum dadas pelos usuários tendem a possuir maior similaridade. Então, o *score* de recomendação para cada item i em S é calculado conforme o nível de preferência do usuário alvo pelos itens semelhantes a i.

As medidas de similaridade podem ser aplicadas entre as linhas da matriz usuário-item P que correspondem aos perfis de avaliação dos usuários para todos os itens e entre as colunas da matriz P que representam avaliações de todos os usuários para os itens. A filtragem colaborativa baseada em memória é muito utilizada pelos sistemas de recomendação devido a sua simplicidade e efetividade (AGGARWAL *et al.*, 2016). Um dos fatores que impactam no sucesso desta técnica é que as medidas de similaridade necessitam apenas de uma única fonte de informação, que nesse caso é o conjunto de avaliações feitas pelos usuários. Entre as métricas tradicionais mais utilizadas estão: similaridade de cosseno, correlação de *Pearson* e distância Euclidiana.

As abordagens de FC baseadas em modelo utilizam métodos de aprendizado de máquina supervisionados ou não supervisionados com os dados de preferências dos usuários por itens para o treinamento de modelos e predizem a propensão de um usuário alvo ter interesse por um item ainda desconhecido por ele (YANG *et al.*, 2014b; CHEN *et al.*, 2018). A FC baseada em modelo tende a garantir uma melhor escalabilidade do que a FC baseada em memória que necessita do cálculo de similaridade de cada par de usuário ou itens e se torna uma deficiência para sistemas que operam em larga escala (SARWAR *et al.*, 2001). Segundo Bobadilla *et al.* (2013), entre as técnicas mais utilizadas estão: redes Bayesianas (PARK; HONG; CHO, 2007), fatoração de matrizes (KOREN; BELL; VOLINSKY, 2009) e Redes Neurais (HE *et al.*, 2017).

Figura 13 – Fluxograma de filtragem colaborativa.



Fonte: Elaborada pelo autor.

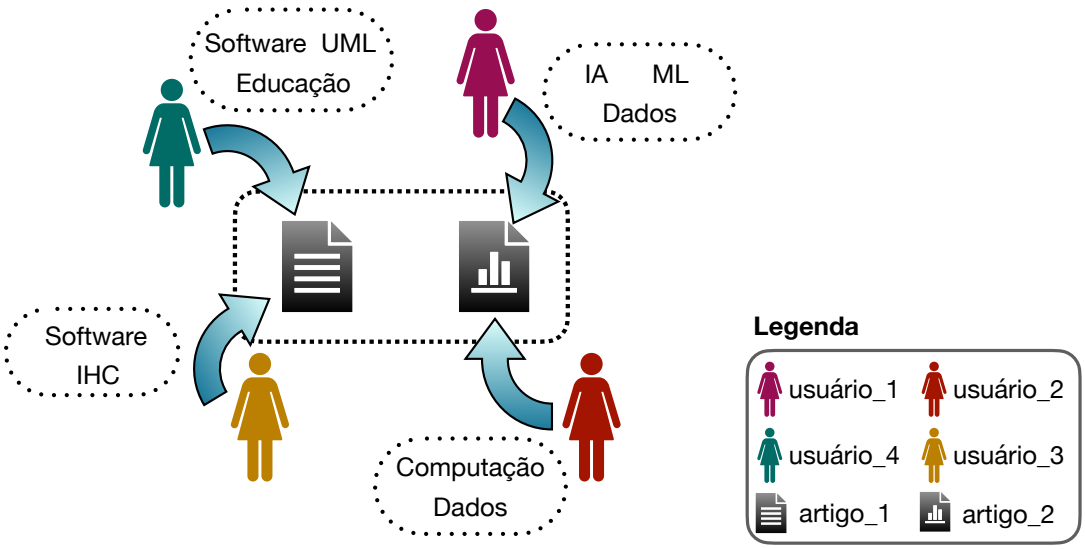
Muitas aplicações reais tendem a gerar conjunto de dados de alta dimensionalidade, pois possuem grande quantidade tanto de itens como de usuários. Além disso, em geral, a fração de usuários que avaliam os itens é muito pequena, o que gera uma matriz de relação usuário-item extremamente esparsa (AGGARWAL *et al.*, 2016). Estas características constituem-se com um dos maiores desafios dos métodos de filtragem colaborativa, pois causam impacto negativo direto sobre a qualidade das recomendações geradas. Em razão disso, os algoritmos de recomendação podem ter baixa precisão ou cobertura nos casos em que, respectivamente, produzem recomendações ruins ou deixam de recomendar itens ideais para os usuários (COSTA; MANZATO; CAMPELLO, 2019).

3.2 Sistemas Colaborativos de Marcação

Sistemas colaborativos de marcação são caracterizados por permitirem que os usuários atribuam termos ou palavras-chave, denominadas de *tags*, aos itens desejados (GOLDER; HUBERMAN, 2006; LIPCZAK *et al.*, 2009). As *tags* são uma alternativa para permitir que o usuário possa filtrar e categorizar informações, fornecendo uma navegação facilitada. Além disso, os conteúdos de marcação fornecidos voluntariamente pelos usuários também podem ser utilizados como informações relevantes para processos de busca, recuperação de informação e recomendação. A atribuição de *tags* pode ocorrer para diferentes tipos de objetos/itens de um sistema, como imagens, músicas, postagens em blogs e páginas da Web e assim por diante, sempre com base no conteúdo e característica do item.

O processo de marcação (do inglês, *tagging*) de um item é realizado de forma espontânea pelos usuários do sistema, sendo assim, subjetivo em relação à perspectiva que cada usuário tem sobre os itens. Desta forma, um grupo de usuários pode caracterizar um determinado item através de diferentes conjuntos de *tags*, como ilustrado na Figura 14, em que o $usuario_1$ marca o $artigo_1$ com as *tags* "computação"e "dados", enquanto o $usuario_2$ marca o mesmo artigo com as *tags* "IA", "ML"e "dados". À medida que a atribuição de *tags* aumenta, as relações semânticas entre usuários e itens são aprimoradas e surgem padrões de categorização (ZHANG; ZHOU; ZHANG, 2011). À vista disso, a marcação de *tags* torna-se uma opção de recurso adicional para gerar mais informações que, se devidamente explorado, pode auxiliar os sistemas de recomendação a melhorar as recomendações aos usuários (MARINHO *et al.*, 2011).

Figura 14 – Exemplo do processo de marcação de *tags* em um item pelo usuário.

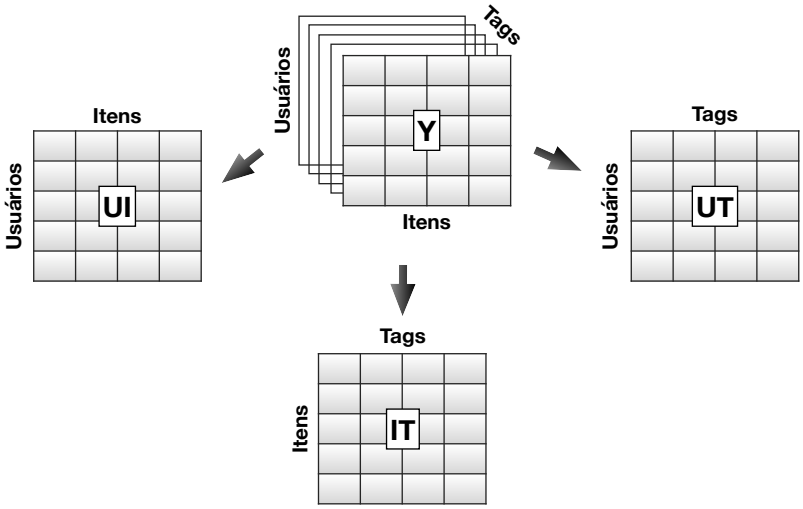


Fonte: Elaborada pelo autor.

As atribuições coletivas de *tags* em um sistema gera uma estrutura de dados denominada de folksonomia (do inglês *folksonomy*), que descreve os relacionamentos entre usuários, *tags* e itens. Formalmente, uma folksonomia é definida como uma tupla $F(U,T,I,Y)$ em que U, T e I

são conjuntos finitos não vazios, formados pelos elementos usuários, *tags* e itens, respectivamente. Já Y é uma relação ternária entre os elementos e $Y \subseteq U \times T \times I$ que descreve as atribuições de *tags*. A partir disto é possível extrair o conjunto de tags T_u usado por um dado usuário u e o conjunto de tags T_j atribuídas ao recurso r por diferentes usuários. Assim, o conjunto de as atribuições de *tags* feitas por um usuário i a um item j pode ser definido como uma tripla $Z_{u_i,r_j} = (u_i, r_j, T_{u_i,r_j})$, em que T_{u_i,r_j} é um conjunto de *tags*. O conjunto de todas as atribuições de *tags* feitas por um usuário é denominada de *personomy* e a coleção de todas as *personomies* de um sistema constitui a estrutura folksonomia (HOTHO *et al.*, 2006b; JÄSCHKE *et al.*, 2008). A Figura 15 ilustra a estrutura tridimensional da folksonomia e as possíveis decomposições em matrizes bidimensionais.

Figura 15 – Estrutura de folksonomia em matriz e as possíveis decomposições.



Fonte: Elaborada pelo autor.

O conjunto de triplas Y pode ser descrito como um grafo tripartido não direcionado $G = (V, E)$, em que $V = U \in T \in R$ são os conjuntos de vértices e cada atribuição de tag a um item por um usuário gera um conjunto de arestas que representam as relações entre esses três elementos do grafo. Desta forma, as definições presentes na Teoria dos Grafos tem potencial para serem utilizadas para analisar as relações entre os elementos da folksonomia. Um exemplo interessante é que duas *tags* podem ser consideradas como altamente relacionadas/similares quando possuem em comum uma vizinhança de vértices do tipo item. Dependendo da análise, o grafo tripartido da folksonomia pode ser divido em três grafos bipartidos, que modelam associações agregadas entre usuários e itens (UI), usuários e *tags* (UT) e itens e *tags* (IT). A função responsável por gerar os pesos das arestas do grafo pode ser definida com base em diferentes estratégias como, grau de conexão entre dois vértices, valor das avaliações realizadas pelos usuários e informação binária que registra se um usuário marcou ou anotou um item com uma determinada *tag*.

Os padrões de conexão entre os vértices do grafo tripartido podem ser intuitivamente

utilizados para a inferência sobre os comportamentos dos relacionamentos entre os elementos, pertencentes ao mesmo tipo ou entre tipos diferentes. Por exemplo, podemos inferir que dois itens são semelhantes quando possuem alto grau de conexões em comum com usuários e *tags* ou que dois usuários são semelhantes se possuírem perfis similares de conexão com *tags*. Assim, a Teoria dos Grafos fornece métricas de análise que se encaixam perfeitamente a modelagem dos dados dos sistemas colaborativos de marcação, como análise de centralidade e proximidade.

Os sistemas colaborativos de marcação possibilitam que os conteúdos sejam estruturados de uma forma interessante para a descoberta de conhecimento sobre conteúdos relevantes para os usuários. Uma vez que esses sistemas são compostos por dados produzidos pelas experiências e comportamento dos próprios usuários, a recomendação personalizada de *tags* visa facilitar o processo de atribuição de *tags* por parte dos usuários, com o intuito de que sejam utilizadas *tags* o mais coerente possível com o perfil de cada usuário (HAMOUDA; WANAS, 2011; BELÉM; ALMEIDA; GONÇALVES, 2017), auxiliando diretamente na melhora do desempenho dos serviços oferecidos pelos sistemas aos usuários. Da mesma forma, os padrões de marcação de *tags* em itens podem contribuir com a recomendação de itens mais pertinentes aos perfis dos usuários. Assim, os perfis de atribuição de *tags* dos usuários são inseridos e considerados como ponto central para as análises e desenvolvimento de métodos de recomendação em sistemas colaborativos de marcação.

A estrutura de dados folksonomia é usada em processos de mineração de dados e aprendizado de máquina para fornecer melhores experiências para os usuários em funções nos sistemas, tais como busca, organização e recomendação de itens e *tags*, pois o processo de marcação de *tags* agrega informações oportunas aos dados de preferências dos usuários por itens. Nesse contexto, métodos de aprendizado em redes complexas são técnicas promissoras de descoberta de padrões que fornecem recomendações personalizadas, processando eficientemente a abundância de informações.

Os dados de sistemas colaborativos baseado em atribuições de *tags* podem ser vistos do ponto de vista de modelagem textual. Um conjunto de itens pode ser entendido como os documentos de uma coleção textual, descritos por um conjunto de *tags* que são palavras-chave atribuídas pelos usuários (KRESTEL; FANKHAUSER; NEJDL, 2009). Nesse sentido, a extração de tópicos se torna uma técnica interessante, pois cada tópico é composto por palavras que representam um determinado conceito, e os documentos são representados como combinação desses conceitos (FALEIROS; VALEJO; LOPES, 2020). Assim, é possível modelar os dados originados do processo de marcação de *tags* como um problema não supervisionado de extração de tópicos em que são descobertos os tópicos latentes que representam os objetos do conjunto de dados, como itens, *tags* e usuários.

3.3 Recomendação de *Tags* com Integração de Grafos Bipartidos

Além de contribuir para melhorar a experiência do usuário no sistema, a recomendação de *tags* visa auxiliar no desempenho do sistema na busca, descrição e organização dos itens, pois são componentes chaves que caracterizam as relações entre usuários e itens (MISHNE, 2006; TANG *et al.*, 2019). Nesse contexto, o processo de recomendação deve ser modelado para identificar *tags* relevantes tanto para um usuário alvo u como para um determinado item i associado a ele. Portanto, o produto final do processo é uma lista das k principais *tags* mais adequadas para um par usuário-item. Assim, são necessárias abordagens específicas que se diferenciam das utilizadas para a recomendação de itens tradicional (BELÉM; ALMEIDA; GONÇALVES, 2017). Esta característica mais complexa influencia diretamente na importância da eficácia dos métodos, pois recomendações inconsistentes além de prejudicar a satisfação do usuário também interferem no desempenho de vários serviços oferecidos pelos sistemas baseados nas informações de *tags* (BELÉM; ALMEIDA; GONÇALVES, 2017). Considerando estes desafios, nesta seção é apresentada a proposta de uma abordagem de recomendação de *tags* baseada em aprendizado de representação em grafos bipartidos.

3.3.1 Trabalhos relacionados

Diferentes técnicas de mineração de dados e aprendizado de máquina têm sido aplicadas ao problema de recomendação de *tags*, tais como métodos supervisionados, não supervisionados, de extração de representação e baseados em conceitos de Recuperação da Informação. Além disso, observam-se diferentes abordagens em relação à modelagem da representação dos dados de atribuição, podendo-se destacar os grafos.

Hotho *et al.* (2006b), Hotho *et al.* (2006a) propuseram o algoritmo *FolkRank*, referência no problema de recomendação de *tags* em folksonomia, principalmente dentre os métodos baseados em grafos. A fundamentação do método é o algoritmo de busca na *web PageRank* (BRIN; PAGE, 1998) que define a importância de uma página *web* conforme a quantidade, e a relevância, de outras páginas que apontam (*hiperlinks*) para ela. Os autores do *FolkRank* identificaram que este mesmo princípio pode ser utilizado para representar o problema de recomendação em *folksonomia*. Para tanto, foi proposta uma adaptação do *PageRank* capaz de lidar com a característica da estrutura folksonomia, que é um grafo não direcionado composto por relações ternárias. A partir do *FolkRank* outras estratégias análogas foram propostas. Ramezani (2011) aplicaram o algoritmo *PageRank* em uma estrutura de *folksonomia* modelada como um grafo direcionado ponderado, com o intuito de capturar de uma melhor forma a navegação do usuário pelos itens e *tags*. Rawashdeh *et al.* (2013) propuseram uma abordagem baseada em grafo tripartido não direcionado e na medida *Katz* (KATZ, 1953), medida de centralidade baseada no cálculo dos caminhos entre cada par de vértices de um grafo, para fornecer aos usuários

recomendações de *tags*.

Marinho e Schmidt-Thieme (2008) aborda o problema de recomendação de *tags* com o embasamento da FC, inserindo a condição de que apenas os usuários que marcaram o item desejado i para o usuário alvo u poderão ser considerados para compor a vizinhança de u e assumindo que apenas as *tags* utilizadas pelos usuários da vizinhança de u para a marcação de i serão consideradas para a recomendação. Hamouda e Wanas (2011) adaptaram a filtragem colaborativa para contornar as limitações impostas pelos itens sem associações de *tags* e usuários sem histórico de associações e marcação com itens. Além disso, propuseram a construção de uma matriz de co-ocorrência que indica as principais *tags* para serem sugeridas, baseada nas atribuições de *tags* feitas por um usuário alvo e pelos usuários que atribuíram *tags* para os mesmos itens que o usuário alvo.

Mishne (2006) utiliza um mecanismo de recuperação de informação para encontrar documentos similares a um documento escrito por um usuário e recomenda as *tags* mais utilizadas pelos documentos similares, sendo que as já utilizadas recebem um peso maior. Byde, Wan e Cayzer (2007) propuseram uma recomendação de *tags* baseada em duas métricas de similaridade, uma fundamentada no perfil de atribuição de tags e outra na análise de conteúdo dos documentos representando pelas das palavras. Lipczak *et al.* (2009), Lipczak e Milios (2011) analisaram e propuseram a combinação das informações de *tags* associadas aos itens e aos usuários com palavras extraídas dos títulos dos itens [1] que co-ocorrem com as *tags* que compõem a folksonomia base, assim expandindo o conjunto de *tags* candidatas a recomendação.

Garg e Weber (2008) desenvolveram um método hibrido que combina recomendações de *tags* geradas pelo método *Naive Bayes* a partir do histórico de anotação de *tags* de um usuário alvo (informação local) com uma matriz de co-ocorrência em que cada entrada (i, j) conta em quantos itens as *tags* i e j co-ocorrem no conjunto de anotações total. Sigurbjörnsson e Zwol (2008) analisaram o comportamento de atribuição de *tags* do *site Flickr*[2], que permite o armazenamento e organização de fotos e vídeos, e propuseram diferentes estratégias de recomendação de *tags* baseadas em métricas de co-ocorrência, uma simétrica baseada no coeficiente *Jaccard* e outra não simétrica em que a taxa de co-ocorrência entre duas *tags* é normalizada pela frequência de uma delas. Wu *et al.* (2009) adotam as mesmas medidas propostas por Sigurbjörnsson e Zwol (2008) e incorporam outras informações de correlação geradas a partir das associações entre as *tags* e imagens. Em geral, essas estratégias calculam as taxas de co-ocorrências apenas entre pares de *tags*, levando a uma possível perda de relacionamentos de co-ocorrência mais complexos e importantes (BELÉM; ALMEIDA; GONÇALVES, 2017). Neste sentido, Menezes *et al.* (2010) desenvolveram um modelo que explora a co-ocorrência de *tags* a partir da extração de regras de associação.

Nos últimos anos, uma série de trabalhos foram desenvolvidos utilizando redes neurais

[1] Objetos do sistema de gerenciamento de favoritos e compartilhamento de publicações BibSonomy
[2] https://www.flickr.com/

profundas. Tang *et al.* (2019) utilizam redes neurais recorrentes com mecanismo de atenção para explorar três características que impactam a recomendação de *tag* baseada em conteúdo textual: modelagem sequencial de texto, correlação entre *tags* e sobreposição de tags nos conteúdos textuais. Liu *et al.* (2020) modelam as interações entre usuários, vídeos, *tags* e mídias como uma rede heterogênea e utilizam *graph neural network* para a extração de *embeddings* utilizadas para a recomendação das *tag*. Lei *et al.* (2020) propuseram a integração de mecanismo de atenção com *capsule networks* (SABOUR; FROSST; HINTON, 2017) para a recomendação de *tags* para documentos, mais especificamente artigos, a partir dos conteúdos textuais e histórico de atribuição de *tags*. Sun *et al.* (2021) desenvolveram um processo de recomendação de *tags* baseado em *hierarchical attention networks* para a construção de modelos baseados em diferentes tipos de informação.

3.3.2 Proposta

Inspirada pelas características do problema de recomendação de *tags* e abordagens utilizadas pelos trabalhos da literatura, esta proposta consiste na criação de uma abordagem de recomendação de *tags* baseada no método não supervisionado de propagação em redes bipartidas PBG (FALEIROS; VALEJO; LOPES, 2020), um algoritmo iterativo baseado em modelagem bipartida de dados textuais, em que os vértices representam documentos e palavras, e arestas correspondem as ocorrências das palavras nos documentos. Deste modo, as propriedades do PBG se ajustam ao problema em questão e têm potencial para o aprendizado de informações latentes dos dados de sistemas colaborativos de marcação. O intuito dessa proposta é (i) mostrar que a modelagem dos dados de atribuição de *tags* como grafos bipartidos é compatível com as propriedades do método de extração de tópicos PBG, (ii) demonstrar que as informações latentes extraídas para cada grafo são correspondentes, podendo ser integradas para fornecer a recomendação de *tags* e (iii) contribuir com uma nova abordagem para recomendação de *tag* em sistemas colaborativos de marcação.

A abordagem proposta, denominada ***T**ag **R**ecommender based on **R**epresentation **L**earning in **B**ipartite **G**raph* (TRLBG), é ilustrada na Figura 16. O primeiro passo do processo de recomendação de *tags* da abordagem proposta é a modelagem dos dados de atribuição em matrizes. A partir dos perfis de atribuição é possível modelar dois grafos bipartidos que representam os padrões de relações que as *tags* possuem com os itens e usuários, fazendo uma analogia com a modelagem textual. Desta forma, a abordagem é baseada nos seguintes grafos bipartidos:

- usuário-tag: os usuários $u \in U$ correspondem aos documentos e as *tags* $t \in T$ representam as palavras. Nesse caso, o grafo carrega a informação do perfil de preferência por *tags* que os usuários possuem para descrever os itens de suas coleções.

- item-tag: cada item $i \in I$ pode ser visto como um documento descrito por um conjunto de *tags* $t \in T$. Nesse cenário, o grafo representa a descrição dos itens pelas tags atribuídas

pelos usuários.

Um ponto importante da modelagem é a função de atribuição de pesos das arestas, uma vez que é uma informação utilizada no mecanismo de inferência do método de aprendizado e influencia diretamente no resultado do processo de propagação. Portanto, primeiramente, é realizada a modelagem das matrizes UT e IT correspondentes aos grafos bipartidos não direcionados $G1$ e $G2$, como mostra a Figura 16 (a). As entradas das matrizes possuem os valores que expressam a relação entre usuário-tag e item-tag, podendo ser gerados por diferentes esquemas de atribuição de pesos para as arestas do grafo. Neste estudo, as seguintes formas de ponderação foram utilizadas:

- **Binária:** considera se ouve ou não a atribuição de uma *tag*. Assim, os pesos das arestas (u,t) do grafo $G1$ recebem 1 quando ocorre a utilização da tag t pelo usuário u e 0 caso contrário. Da mesma forma, os pesos das arestas (i,t) do grafo $G2$ recebem 1 quando um item i recebe a atribuição da tag t e 0 caso contrário.
- **Baseada em frequência:** considera a frequência da ocorrência da atribuição de uma *tag*. Desta forma, os pesos das arestas (u,t) do grafo $G1$ equivalem à quantidade de itens i que o usuário u marcou com a *tag* t e 0 caso não tenha marcações. Do mesmo modo, os pesos das arestas (i,t) do grafo $G2$ correspondem ao número de usuários u que marcaram o item i com a *tag* t.

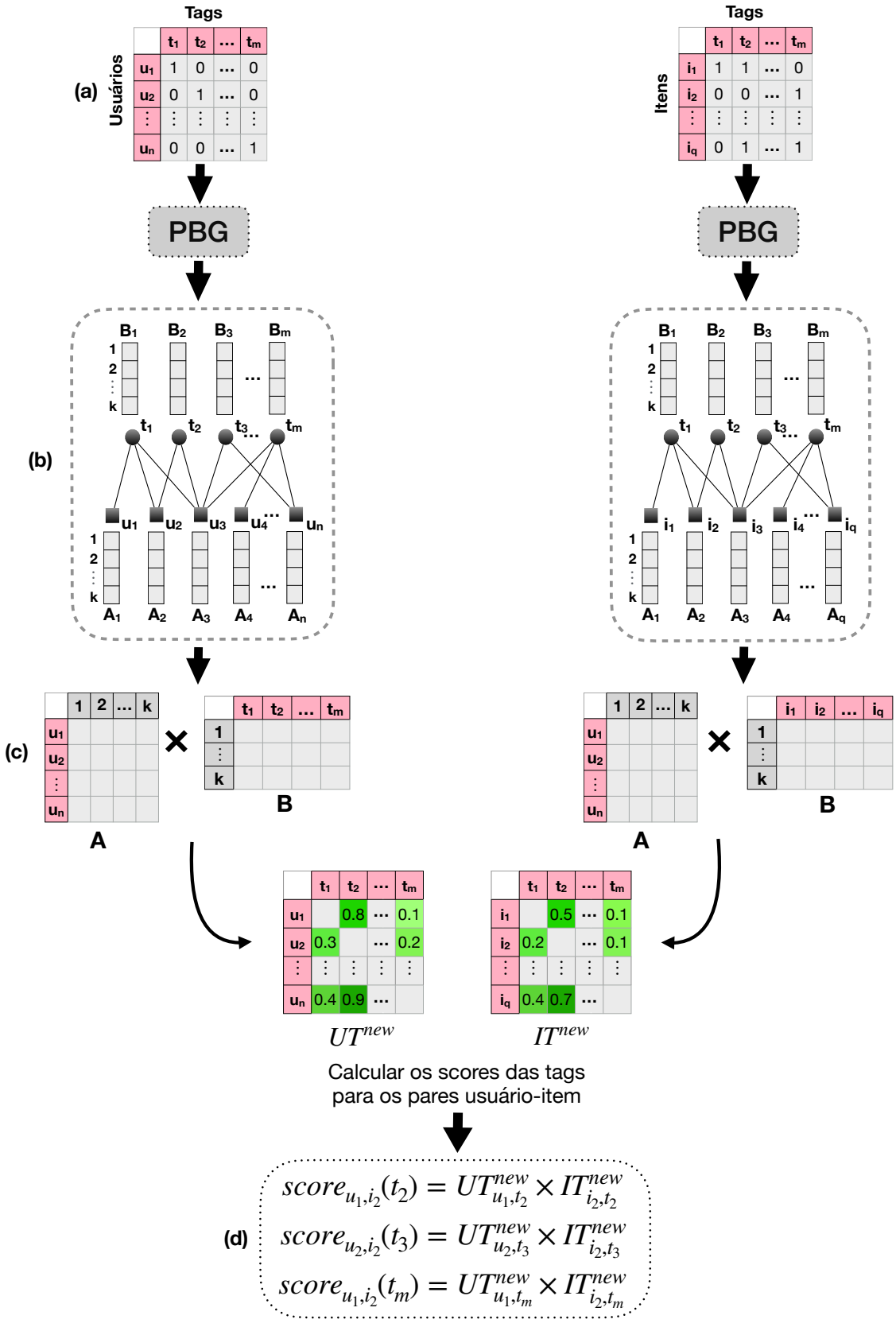
A partir da modelagem das matrizes, é iniciado o processo de aprendizado, no qual o método PBG é aplicado individualmente em cada grafo bipartido para a extração de representação, conforme detalhado na Seção 2.4.3. Após o processo de propagação ser executado nos grafos, são obtidas novas representações para cada vértice dos grafos, que sintetizam as informações latentes das associações entre os vértices. Para cada grafo, são geradas duas matrizes $A_{n\times k}$ (ou $A_{q\times k}$ para o caso do grafo item-tag) e $B_{m\times k}$ compostas pelos vetores de tamanho k computados para cada vértice das redes, como mostra a Figura 16 (b). Para cada rede bipartida, essas matrizes são utilizadas para que sejam obtidas as pontuações que indicam o grau de pertinência da existência de associação entre os vértices, como ilustra a Figura 16 (c). A matriz que contém as probabilidades de associações entre os vértices da rede bipartida usuário-tag é obtida da seguinte forma:

$$UT^{new}_{n\times m} = A_{n\times k} \times B^{\mathsf{T}}_{m\times k}. \tag{3.1}$$

Por outro lado, a matriz de probabilidades referente à rede bipartida item-tag é definida como:

$$IT^{new}_{q\times m} = A_{q\times k} \times B^{\mathsf{T}}_{m\times k}. \tag{3.2}$$

Figura 16 – Fluxograma da abordagem proposta TRLBG para recomendação de *tags*.



Fonte: Elaborada pelo autor.

Assim, quanto maior a pontuação, maior a certeza da existência de uma relação (conexão) entre os vértices. Portanto, para o grafo bipartido usuário-tag a matriz UT^{new} indica o quanto cada *tag* pode ser interessante para cada usuário e para o grafo item-tag a matriz IT^{new} indica o quanto cada *tag* pode ser apropriada para cada item.

Com as matrizes UT^{new} e IT^{new} computadas, inicia-se o processo de recomendação de *tags* para os pares (u,i). A identificação de *tags* potenciais para os usuários deve ser feita com base nos padrões existentes nos interesses pessoais pelos itens e pelas *tags*, pois cada usuário tem um perfil de preferência e atribuição específico. Desta forma, foram identificadas quais *tags* estão mais relacionadas tanto com um determinado item i quanto com o usuário alvo u, para serem indicadas para o par (u,i). Para isso, as recomendações personalizadas de *tags* para um determinado par usuário-item são geradas com base nas informações das matrizes UT^{new} e IT^{new}, como mostra a Figura 16 (d). A pontuação final de recomendação das *tags* para um usuário u e um item i é dada pela multiplicação das pontuações das entradas UT^{new} e IT^{new} das matrizes, formalmente, definida por:

$$score_{u,i}(t) = UT^{new}_{u,t} \times IT^{new}_{i,t}. \tag{3.3}$$

Por fim, é gerada uma lista de recomendação com base nas pontuações obtidas para cada tupla (u,i,t). Como resultado, as k *tags* com as maiores pontuações serão recomendadas.

3.3.3 Experimentos e Resultados

Nos experimentos, a metodologia de avaliação proposta por Jäschke *et al.* (2008) foi adotada para realizar a comparação de desempenho entre o método proposto e os da literatura. Os autores utilizaram uma adaptação da validação *leave-one-out*, denominada de *LeavePostOut*, em que para cada usuário, um item i é aleatoriamente selecionado e tem as suas atribuições de *tags* removidas do conjunto de treinamento e adicionadas ao conjunto de teste. Por exemplo, dado que o usuário u_1 atribuiu as *tags* t_1, t_4 e t_7 ao item i_3 e que este foi selecionado para validação, todo o conjunto de *tags* $\{t_1,t_4,t_7\}$ atribuído ao item é removido dos dados de atribuição de treinamento. Assim, o objetivo da tarefa de recomendação é, a partir dos dados de atribuição de *tags* de treinamento, identificar as *tags* ocultadas do perfil de atribuição de tags de um usuário u para um item i. O processo de avaliação não foi considerado para os usuários que possuem apenas um item com atribuição de *tags*, pois nesse caso não seria possível extrair dados de treinamento. Este procedimento foi repetido 5 vezes com diferentes amostragens de conjuntos de dados de teste e treinamento. Assim, todos os valores reportados nas análises experimentais são resultados da média dessas 5 execuções.

Para os experimentos, foram utilizados os conjuntos de dados *MovieLens* e *LastFm*, que estão disponíveis na página do 2º *International Workshop on Information Heterogeneity and*

Fusion in Recommender Systems[3]. *MovieLens*[4] é um sistema de classificação de filmes, em que cada usuário pode categorizar os filmes avaliados por meio da atribuição de *tags* e visualizar o conjunto de filmes que possuem a mesma etiqueta. O conjunto de dados Last.fm é obtido do *site* de música online da Last.fm[5] e permite que os usuários realizem a atribuição de *tags* para músicas e artistas. O mesmo pré-processamento descrito em Zuo *et al.* (2016) foi aplicado para a remoção de *tags* infrequentes. Foram selecionadas as *tags* que ocorrem mais do que três vezes no conjunto de dados *MovieLens* e cinco vezes no conjunto de dados *LastFm*. A Tabela 2 resume as estatísticas gerais dos conjuntos de dados após o pré-processamento.

Tabela 2 – Estatísticas gerais dos conjuntos de dados utilizados.

Conjunto de Dados	Usuários	Itens	*Tags*
LastFm	1808	12212	2305
MovieLens	1777	5554	2561

Fonte: Elaborada pelo autor.

O método proposto foi comparado diretamente com os seguintes métodos: *KaztBm25* (RAWASHDEH *et al.*, 2013); FC baseada no perfil de atribuição de *tags* (FCTags) (MARINHO; SCHMIDT-THIEME, 2008); e Mix of Most Popular Tags (MMPT) (JÄSCHKE *et al.*, 2008), que recomenda uma combinação das *tags* mais populares de um usuário alvo com as tags mais populares de um determinado item. Os desempenhos dos métodos foram analisados conforme nove tamanhos de lista de recomendação: $L = \{1,2,4,6,8,10,15,20,30\}$

Para a avaliação dos métodos de recomendação foram utilizadas as medidas precisão e *recall*. A precisão é a quantidade de *tags* recomendadas na lista de recomendação que são realmente relevantes dividida pelo tamanho da lista. Para um conjunto de usuários U, a precisão da recomendação é dada por:

$$P(L) = \frac{1}{n} \sum_{u \in U} \frac{Z_u^t}{L}, \tag{3.4}$$

em que L é o comprimento da lista de recomendação e Z_u^t é a quantidade de *tags* da lista de recomendação gerada para o usuário u que estão de fato no conjunto de teste.

Recall é a taxa de *tags* recomendadas relevantes pela quantidade de *tags* associadas aos pares usuário-item presentes no conjunto de teste. Para um determinado usuário u, a taxa *recall* da recomendação é definida como:

$$R(L) = \frac{1}{n} \sum_{u \in U} \frac{N_t}{Z_u^t}, \tag{3.5}$$

onde N_t é o número total de *tags* de um par (u,i) contidas no conjunto de teste. O comprimento L da lista de recomendações, exerce influência direta nas medidas precisão e *recall*. Assim, quanto maior o valor de L maior é R, porém, quanto menor o valor de L, maior é a precisão P.

[3] http://ir.ii.uam.es/hetrec2011/
[4] https://movielens.org
[5] http://last.fm

Primeiramente, uma análise experimental foi realizada para determinar os valores dos parâmetros do método PBG, método central da abordagem proposta TRLBG, que permitam a melhor extração de informações latentes das redes bipartidas. A combinação de valores dos parâmetros tem relação direta com o resultado da recomendação e, portanto, foi investigado como o desempenho se comporta conforme variações nos valores dos parâmetros alfa (α), beta (β), tamanho dos vetores (k) e quantidade de propagações locais e globais ($prop$). Ambas as bases (*LastFm* e *MoviesLens*) e medidas de avaliação foram consideradas para a análise. Além disso, incluímos nessa análise os dois tipos de modelagem, binário e de frequência, pois os pesos das arestas dos grafos bipartidos influenciam diretamente no processo de propagação.

Para cada parâmetro foi definido o conjunto de valores utilizados na análise, são eles: $\alpha = \{0.1, 0.01, 0.001\}$, $\beta = \{0.0, 0.01, 0.001\}$, $prop = \{10, 50, 100, 150\}$ e $k = \{250, 500, 750, 1000\}$. Como ilustrado na Figura 16, a abordagem é constituída por dois processos sequenciais responsáveis pelo aprendizado de representações, um para o grafo usuário-tag e outro para o grafo item-tag. Desta forma, é possível utilizar diferentes valores de k para cada processo, porém em experimentos iniciais foi identificado não haver melhora significativa com essa variação. O mesmo foi adotado para o número de propagação local e global. Esses aspectos mostram que a abordagem não é sensível a essas variações mais específicas desses parâmetros.

A partir dos conjuntos de valores definidos para cada parâmetro, obteve-se um total de 144 combinações de parâmetros com as quais o método proposto foi executado. A Tabelas 3, 4, 5, 6 mostram os resultados das doze melhores combinações de parâmetros identificadas para cada base e medida de avaliação. Foram observados os resultados de precisão e *recall* para cada um dos nove tamanhos de lista de recomendação adotados, assim como a média destes resultados. Um dos objetivos desta análise foi obter um conjunto de valores *default* dos parâmetros que seja válido para diferentes bases e modelagens. O conjunto de valores $\alpha = 0.001$, $\beta = 0.001$, $prop = 100$ e $k = 500$ foram definidos como referência. É possível observar nas Tabelas 5 e 6 que no caso da base *MovieLens* os valores *default* não geram os melhores resultados, como ocorre com a base *LastFm*. Porém, a perda de desempenho é pequena, mostrando que a abordagem não é extremamente sensível à variação do conjunto de valores dos parâmetros. O conjunto *default* de parâmetros foi utilizado na análise comparativa com métodos da literatura a fim de fornecer uma comparação justa.

A análise de comparação entre a abordagem proposta TRLBG e os métodos da literatura foi realizada com base nas medidas de precisão e *recall*, descritas acima, e conforme a variação do tamanho L da lista de recomendação de *tags*. Desta forma, é possível observar o comportamento dos métodos à medida que a quantidade de *tags* recomendadas aumenta. Além disso, foi investigado o efeito dos pesos das arestas, binário e frequência, sobre o desempenho dos métodos.

As Figuras 17 e 18 mostram os resultados de precisão e *recall* para o conjunto de dados *LastFm*, respectivamente. É possível observar que quando o peso do tipo frequência

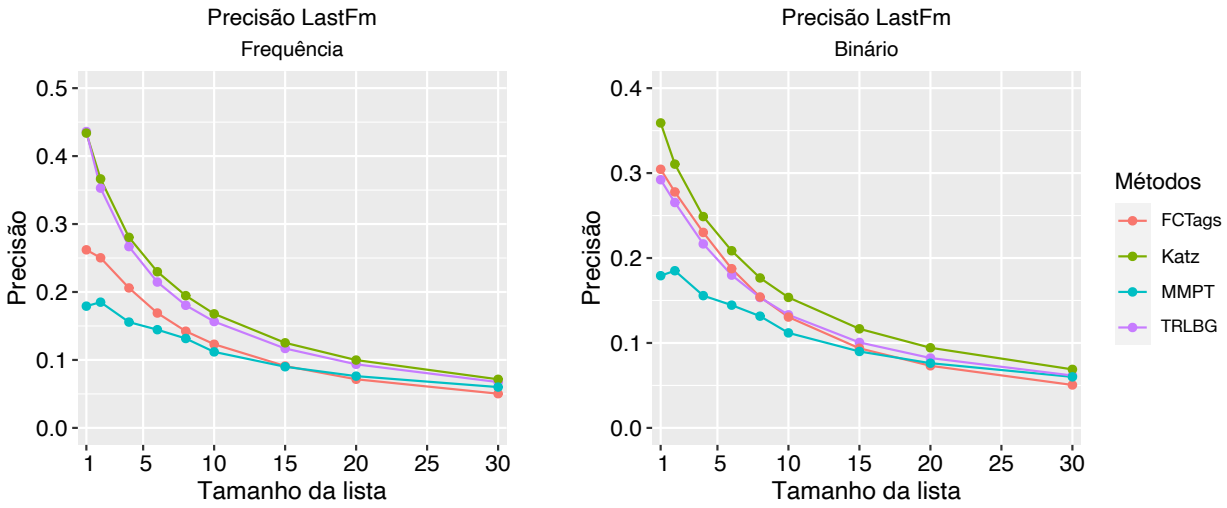
Tabela 3 – Resultado da análise experimental do desempenho do método PBG para a base *LastFm* e medida precisão, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os tamanhos de lista de recomendação. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros				Frequência				Binário			
Alfa	Beta	Prop	K	Média	L1	L2	L4	Média	L1	L2	L4
0.001	**0**	**100**	**500**	**1.885**	0.436	0.353	0.267	**1.485**	0.292	0.265	0.217
0.001	0	100	250	1.857	0.422	0.351	0.266	1.468	0.294	0.259	0.211
0.001	0	100	750	1.792	0.413	0.336	0.249	1.384	0.269	0.245	0.197
0.001	0	50	250	1.852	0.419	0.350	0.265	1.471	0.293	0.261	0.212
0.001	0	50	500	1.881	0.434	0.352	0.267	1.481	0.291	0.265	0.215
0.001	0	50	750	1.797	0.413	0.337	0.250	1.384	0.269	0.245	0.196
0.001	0.001	100	250	1.810	0.413	0.341	0.260	1.437	0.289	0.259	0.209
0.001	0.001	100	500	1.830	0.424	0.346	0.260	1.390	0.277	0.255	0.202
0.001	0.001	100	750	1.814	0.422	0.345	0.259	1.317	0.264	0.241	0.187
0.001	0.001	50	250	1.812	0.414	0.342	0.259	1.437	0.290	0.260	0.209
0.001	0.001	50	500	1.832	0.425	0.346	0.261	1.387	0.277	0.253	0.201
0.001	0.001	50	750	1.816	0.424	0.344	0.259	1.316	0.263	0.240	0.188

Fonte: Elaborada pelo autor.

é considerado, a abordagem proposta TRLBG atinge resultados consideráveis em relação ao método *Katz*, principalmente no caso da medida de precisão. A curva dos resultados da medida *recall* é mais intensa para o caso da modelagem binária, indicando que a qualidade das *tags* acompanha o aumento de tamanho da lista, enquanto no caso da modelagem com frequência a curva sofre uma atenuação quando o tamanho da lista atinge o tamanho dez.

Figura 17 – Resultados da medida precisão para a base *LastFm*.



Fonte: Elaborada pelo autor.

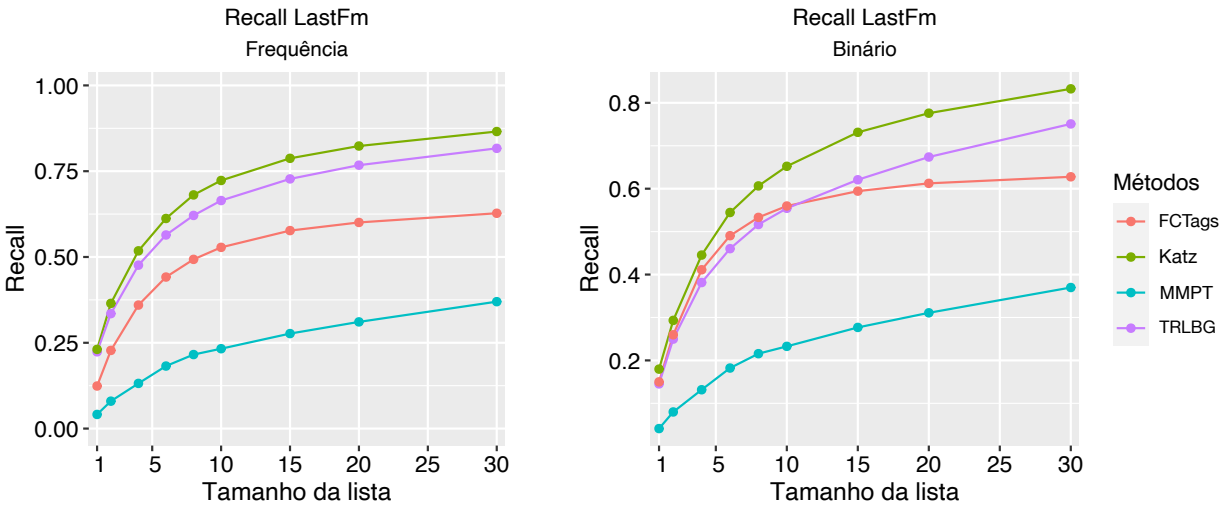
As Figuras 19 e 20 mostram os resultados de precisão e *recall* para o conjunto de dados *MovieLens*, respectivamente. Assim como no caso da base *LastFm*, nota-se o mesmo

Tabela 4 – Resultado da análise experimental do desempenho do método PBG para a base *LastFm* e medida *recall*, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os tamanhos de lista de recomendação. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros				Frequência				Binário			
Alfa	Beta	Prop	K	Média	L1	L2	L4	Média	L1	L2	L4
0.001	**0**	**100**	**500**	**5.196**	0.224	0.335	0.476	**4.353**	0.145	0.250	0.381
0.001	0	100	250	5.073	0.213	0.331	0.472	4.289	0.147	0.246	0.374
0.001	0	100	750	5.026	0.209	0.315	0.441	4.149	0.131	0.229	0.344
0.001	0	50	250	5.058	0.212	0.329	0.471	4.295	0.147	0.249	0.376
0.001	0	50	500	5.186	0.221	0.333	0.476	4.335	0.144	0.249	0.378
0.001	0	50	750	5.034	0.208	0.316	0.444	4.152	0.131	0.228	0.344
0.001	0.001	100	250	4.901	0.205	0.316	0.454	4.097	0.139	0.237	0.362
0.001	0.001	100	500	4.943	0.211	0.320	0.453	3.986	0.136	0.236	0.353
0.001	0.001	100	750	4.879	0.211	0.321	0.453	3.809	0.128	0.222	0.328
0.001	0.001	50	250	4.902	0.207	0.318	0.453	4.089	0.140	0.239	0.361
0.001	0.001	50	500	4.941	0.211	0.321	0.454	3.976	0.135	0.234	0.350
0.001	0.001	50	750	4.879	0.213	0.320	0.453	3.805	0.128	0.220	0.329

Fonte: Elaborada pelo autor.

Figura 18 – Resultados da medida *recall* para a base *LastFm*.



Fonte: Elaborada pelo autor.

comportamento da abordagem proposta TRLBG em relação às modelagens adotadas, em que os melhores resultados são obtidos quando a frequência é utilizada como peso.

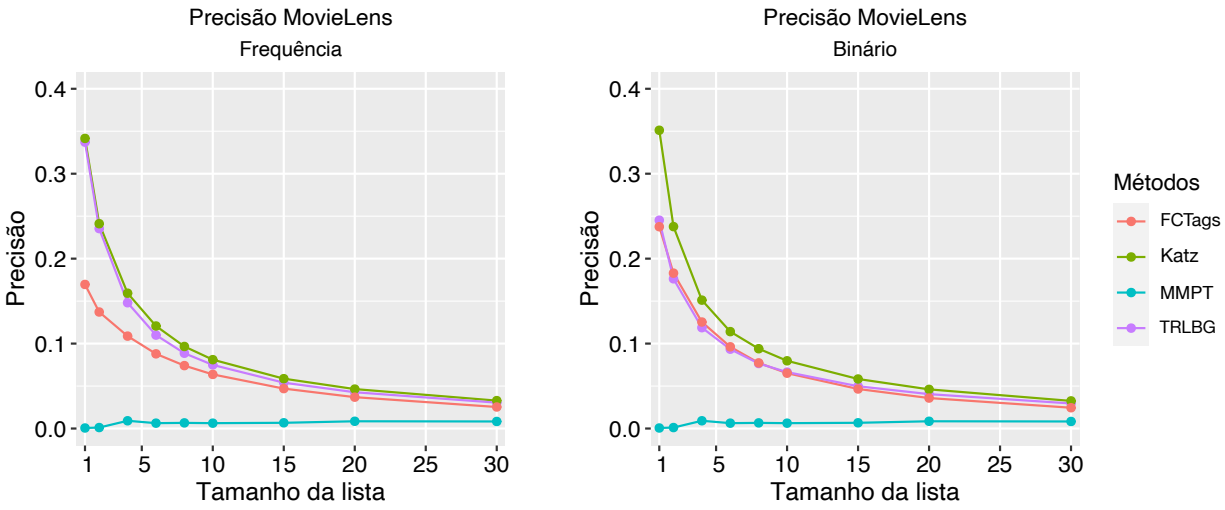
A análise mostra a importância que a modelagem adotada exerce sobre o desempenho dos métodos. Em geral, os resultados da abordagem proposta e do método *Katz* tendem a ser melhores quando a frequência é utilizada como peso, para ambas as bases e medidas adotadas. Este comportamento é observado principalmente na abordagem proposta e pode ser justificado devido à natureza do método PBG, desenvolvido para a modelagem textual em grafos

Tabela 5 – Resultado da análise experimental do desempenho do método PBG para a base *MovieLens* e medida precisão, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os tamanhos de lista de recomendação. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros				Frequência				Binário			
Alfa	Beta	Prop	K	Média	L1	L2	L4	Média	L1	L2	L4
0.001	**0**	**100**	**500**	1.121	0.337	0.235	0.148	0.896	0.245	0.176	0.119
0.001	0	100	250	0.978	0.283	0.202	0.131	0.763	0.207	0.149	0.101
0.001	0	100	750	1.169	0.353	0.245	0.155	**0.964**	0.271	0.190	0.128
0.001	0	50	250	0.978	0.285	0.201	0.130	0.757	0.204	0.148	0.100
0.001	0	50	500	1.115	0.334	0.233	0.148	0.900	0.250	0.175	0.119
0.001	0	50	750	**1.170**	0.357	0.246	0.154	0.958	0.268	0.189	0.127
0.001	0.001	100	250	0.976	0.283	0.202	0.132	0.748	0.196	0.145	0.102
0.001	0.001	100	500	1.081	0.322	0.226	0.144	0.838	0.226	0.167	0.113
0.001	0.001	100	750	1.111	0.334	0.234	0.147	0.857	0.237	0.171	0.115
0.001	0.001	50	250	0.985	0.289	0.203	0.133	0.747	0.196	0.144	0.102
0.001	0.001	50	500	1.083	0.325	0.226	0.144	0.834	0.226	0.165	0.112
0.001	0.001	50	750	1.109	0.331	0.234	0.148	0.852	0.234	0.169	0.114

Fonte: Elaborada pelo autor.

Figura 19 – Resultados da medida precisão para a base *MovieLens*.



Fonte: Elaborada pelo autor.

bipartido que mapeia as ocorrências dos termos nos documentos. Por outro lado, o método de recomendação de *tags* baseado em FC, mostrou melhores resultados para o peso do tipo binário, que corresponde à modelagem de pesos originalmente proposta para este método (MARINHO; SCHMIDT-THIEME, 2008).

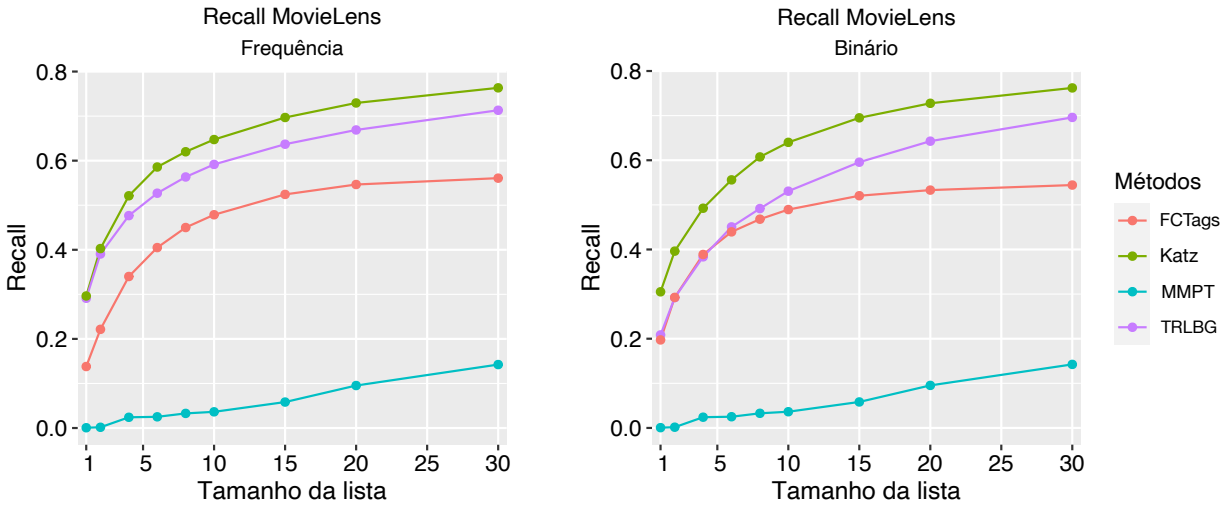
Por fim, é possível realizar uma comparação indireta a partir da análise experimental do artigo do método *KaztBm25* (RAWASHDEH *et al.*, 2013) que mostra que ele tem melhor desempenho do que os métodos *User-Centric Tag Model* (UCTM) (WETZKER *et al.*, 2010) e o

Tabela 6 – Resultado da análise experimental do desempenho do método PBG para a base *MovieLens* e medida *recall*, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os tamanhos de lista de recomendação. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros				Frequência				Binário			
Alfa	Beta	Prop	K	Média	L1	L2	L4	Média	L1	L2	L4
0.001	**0**	**100**	**500**	4.859	0.291	0.391	0.477	4.291	0.209	0.292	0.383
0.001	0	100	250	4.339	0.242	0.333	0.419	3.674	0.176	0.245	0.324
0.001	0	100	750	5.028	0.305	0.406	0.500	4.533	0.231	0.316	0.416
0.001	0	50	250	4.333	0.243	0.330	0.418	3.662	0.172	0.243	0.320
0.001	0	50	500	4.837	0.289	0.387	0.476	4.293	0.213	0.290	0.385
0.001	0	50	750	**5.012**	0.307	0.406	0.497	**4.528**	0.228	0.315	0.414
0.001	0.001	100	250	4.296	0.241	0.332	0.424	3.612	0.165	0.238	0.326
0.001	0.001	100	500	4.660	0.275	0.370	0.463	3.937	0.190	0.273	0.362
0.001	0.001	100	750	4.738	0.283	0.385	0.475	3.974	0.200	0.281	0.374
0.001	0.001	50	250	4.315	0.245	0.333	0.425	3.621	0.165	0.237	0.328
0.001	0.001	50	500	4.662	0.277	0.371	0.464	3.925	0.189	0.271	0.361
0.001	0.001	50	750	4.748	0.282	0.385	0.478	3.960	0.198	0.279	0.369

Fonte: Elaborada pelo autor.

Figura 20 – Resultados da medida *recall* para a base *MovieLens*.



Fonte: Elaborada pelo autor.

algoritmo *FolkRank* Hotho *et al.* (2006b), Hotho *et al.* (2006a), métodos de referência dentre os algoritmos baseados em folksonomia para recomendação de *tags*. Sendo assim, considerando que a abordagem proposta obteve resultados equivalentes ao método *KaztBm25*, pode-se assumir o mesmo em relação ao UCTM e *FolkRank*.

Com base nas análises realizadas, foi possível observar o comportamento da abordagem proposta TRLBG em relação à variação dos parâmetros do método PBG, ao tipo de peso utilizado na modelagem dos grafos bipartidos e aumento do tamanho da lista de recomendação. Acima

de tudo, o aspecto mais importante foi mostrar que as representações aprendidas para cada uma das duas visões, que correspondem aos relacionamentos das *tags* com os usuários e itens, colaboram entre si, permitindo a integração de informação para a recomendação de *tags* para pares usuário-item.

3.4 Recomendação de Itens com Grafo Tripartido

Como mencionado anteriormente, a FC é uma das técnicas mais utilizadas nos sistemas de recomendação tradicionais em que os perfis de preferência dos usuários são utilizados para a recomendação de itens (TSO-SUTTER; MARINHO; SCHMIDT-THIEME, 2008; AGGARWAL *et al.*, 2016). Os perfis de marcação de *tags* dos usuários possuem grande potencial de contribuição para a melhora do desempenho da recomendação de itens pela filtragem colaborativa, pois adicionam informações dos dados de preferência dos usuários por itens. Como consequência, muitos esforços têm sido feitos para o desenvolvimento de sistemas de recomendação de itens baseados em informações de *tags*, do inglês *tag-aware recommender systems* (ZHANG; ZHOU; ZHANG, 2010).

Porém, no caso dos sistemas colaborativos de marcação, em que a estrutura formada pelo relacionamento entre as entidades do sistema é mais complexa, a FC tradicional só pode ser aplicada a partir da decomposição dos dados em relações binárias (RICCI; ROKACH; SHAPIRA, 2011; ZUO *et al.*, 2016) a partir da relação ternária formadas pelos elementos usuário, item e *tag*. Sendo assim, estudos foram desenvolvidos a fim de incorporar as informações de atribuições de *tags* feitas por usuários no processo de recomendação tradicional de itens, pois as *tags* representam o discernimento dos usuários sobre os itens e podem ser vistas como descritores dos itens.

Neste contexto, métodos de aprendizado de representação possuem grande potencial para extrair informações latentes e gerar atributos mais abstratos a partir dos perfis de preferência e atribuição dos usuários (ZUO *et al.*, 2016). A partir desta motivação, esta seção apresenta uma abordagem de recomendação de itens baseada em FC e aprendizado de representação em grafo. A contribuição principal é a proposta de um *framework* de propagação para a extração de representação em grafo tripartido, modelado a partir das relações da estrutura folksonomia, que utiliza como base teórica o método de propagação PBG (FALEIROS; VALEJO; LOPES, 2020). Uma vez que, o padrão de atribuição de *tags* auxilia o processo de descoberta de itens que se ajustam à preferência dos usuários, a base principal da FC para a geração das recomendações será a representação aprendida do domínio de *tags*.

3.4.1 Trabalhos relacionados

Uma variedade de algoritmos de recomendação foram propostos para sistemas colaborativos de marcação. As propostas giram em torno da inserção das informações de atribuição

de *tags* e se diferenciam da forma de como este processo é feito. Encontram-se na literatura métodos: baseados em grafos que exploram a estrutura relacional dos elementos dos sistemas; não supervisionados que utilizam conceitos de agrupamento para categorizar as *tags* e encontrar nos grupos padrões que auxiliem nas recomendações; fatoração de matrizes para a extração de vetores de variáveis latentes inferidos a partir dos padrões dos perfis de marcação dos usuários e redes neurais profundas para o aprendizado de representações. A seguir, são reportados diferentes métodos pertencentes a estas categorias.

Nakamoto *et al.* (2007) criaram dois modelos de FC baseados em contexto através das informações de atribuição de *tags*, incorporadas nas equações de cálculo da similaridade entre os usuários e das pontuações de recomendação. Tso-Sutter, Marinho e Schmidt-Thieme (2008) propuseram um método de fusão para reduzir o espaço tridimensional das relações entre usuário-item-tag em uma matriz que capturam as correlações entre usuários, itens e *tags*. Pan *et al.* (2021) propuseram um modelo de expansão de *tags* para agregar mais informações aos perfis de *tags* dos usuários, por meio de aprendizado de redes bayesianas para a descoberta de relações entre as *tags*.

Alguns estudos são baseados na alocação de recursos em grafos bipartidos proposto por Zhou *et al.* (2007). Wu e Zhang (2010) utilizaram um processo de difusão em rede bipartida ponderada usuário-item gerada conforme as atribuições de *tags* feitas pelos usurários combinadas com um atrator de interesse de cada par usuário-item. Shang *et al.* (2010) aplicaram o método de alocação de recursos em um grafo tripartido modelado a partir das informações de atribuições de *tags* para extrair as similaridades entre os usuários utilizadas pela FC. De forma similar, Zhang, Zhou e Zhang (2010), propuseram um método de recomendação baseado em alocação de recursos em grafo tripartido, porém sem a necessidade de FC, pois o processo de difusão extrai as pontuações de recomendação dos itens para cada usuário alvo.

A inferência dos perfis de interesse dos usuários pode ser obtida com base na importância de um grupo de *tags* para um dado usuário (GEMMELL *et al.*, 2008). Assim, estudos foram desenvolvidos utilizando técnicas de aprendizado não supervisionado. Shepitsen *et al.* (2008) propuseram um algoritmo que utiliza a recomendação gerada pela FC tradicional considerando informações obtidas com base nos grupos de *tags* gerados pela técnica de Agrupamento Hierárquico: o interesse dos usuários por cada grupo de *tags* e os grupos mais relacionados com os itens. Zhen, Li e Yeung (2009) incorpora as informações de marcação de *tags* no processo de Fatoração de Matrizes do método *Probabilistic Matrix Factorization* (PMF) (MNIH; SALAKHUTDINOV, 2007) desenvolvido para a FC tradicional. Wu *et al.* (2012) utilizam os dados de marcação de *tags* para determinar os vizinhos mais próximos (vizinhança) dos usuários e itens e incorporam essa informação no processo de fatoração de matrizes para reforçar que as características latentes dos usuários, ou itens, sejam semelhantes as das suas respectivas vizinhanças.

Zuo *et al.* (2016) propuseram um novo algoritmo de recomendação baseado em redes neurais profundas por meio de um modelo de *autoencoder* esparso, em que os perfis de atribuição

de *tags* são utilizados para a extração de um nova representação mais abstrata utilizada na FC baseada em usuário para gerar recomendações. Por outro lado, Xu *et al.* (2016) utilizam redes neurais profundas para mapear os perfis de usuários e itens gerados pelas atribuições de *tags* para duas novas representações, utilizadas por uma função de recomendação que mede a semelhança e relevância entre um par usuário-item. De forma similar, Xu *et al.* (2017) utilizam esta abordagem, porém com algumas mudanças no modelo de neurais profundas para garantir mais escalabilidade e eficiência do treinamento. Liang *et al.* (2018) desenvolveram um processo de recomendação baseado na construção de novos perfis de usuários com base nos comportamentos de marcação de *tags* dos usuários, que utiliza redes neurais profundas e redes neurais recorrentes para extrair representações latentes dos itens e dos usuários, respectivamente. Ahmadian, Ahmadian e Jalili (2022) propuseram um novo método de recomendação que utiliza redes neurais profundas para modelar representações de informações adicionais aos perfis de preferências dos usuários. Para este fim, um *autoencoder* esparso é usado para extrair representações a partir de relações usuário-usuário e tag-usuário, utilizadas em um processo de filtragem colaborativa.

3.4.2 Proposta

Diante da necessidade de métodos que forneçam recomendação de itens que promovam a satisfação dos usuários e da lacuna existente no método PBG (FALEIROS; VALEJO; LOPES, 2020) para redes k-partidas, esta proposta possui dois objetivos principais: (i) o desenvolvimento de uma nova abordagem de FC baseada em dados de atribuição de *tags* para a recomendação em sistemas colaborativos de marcação e (ii) a criação e análise experimental de um *framework* de aprendizado de representações no qual ocorre a propagação dos vetores de informações latentes dos vértices por todas as camadas de uma rede tripartida, com base na teoria do método não supervisionado de propagação em grafos bipartidos PBG.

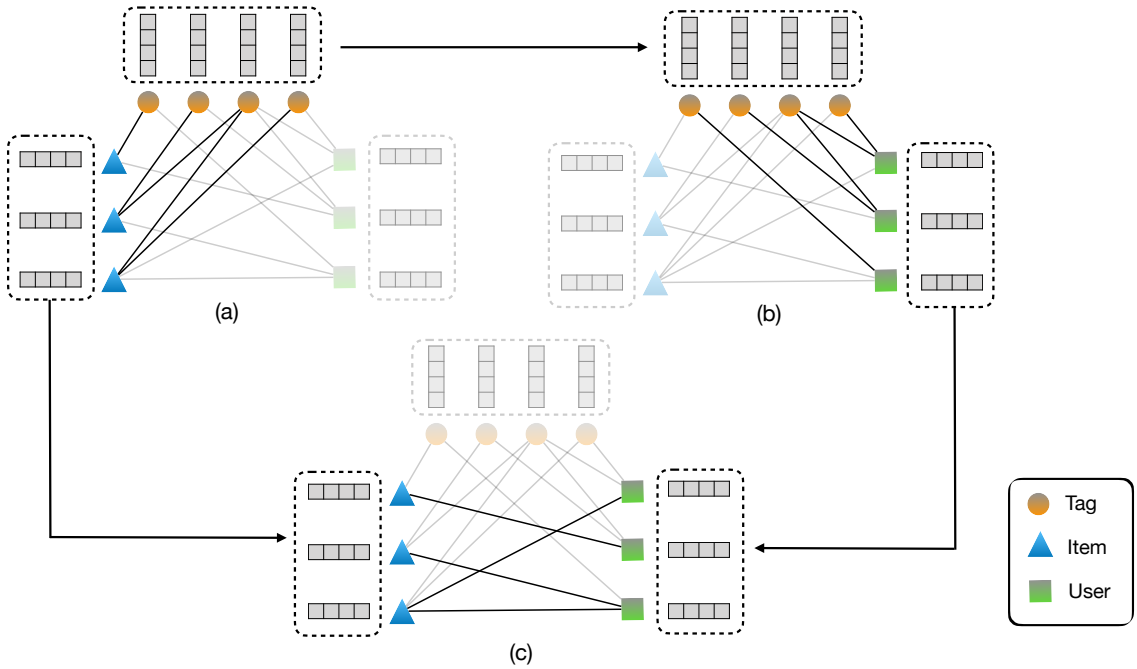
Normalmente, os algoritmos tradicionais de FC operam na relação binária entre usuários e itens (RICCI; ROKACH; SHAPIRA, 2011; ZHANG; ZHOU; ZHANG, 2011). Portanto, como foi adotado a FC como algoritmo de recomendação, o grafo tripartido com as relações usuário-item-tag foi reduzido em três grafos bipartidos: usuário-item $G1(U,I,E)$, usuário-tag $G2(U,T,E)$ e item-tag $G3(I,T,E)$, descritos como três matrizes de adjacências A, A' e A'', respectivamente. Além disso, são definidas as funções de ponderação dos grafos. A partir dos dados de marcação de *tag* é possível adotar tanto peso binário como utilizar a frequência, como descrito na Seção 3.3.2.

É importante destacar que, mesmo com a decomposição do grafo tripartido, as relações entre cada par de grafo bipartido gerado se mantém, por exemplo: o grafo $G1(U,I,E)$ e $G2(U,T,E)$ possuem os vértices do tipo usuário em comum e assim por diante. Este aspecto permite que vetores de informações latentes aprendidos em diferentes camadas bipartidas de uma estrutura k-partida colaborem entre si através de um processo iterativo de transferência de informação na rede. Isto permite que ocorra a difusão de informações ao longo do grafo

tripartido completo. Esta é a concepção do *framework* proposto, denominado ***P**ropagation in **T**ripartite **G**raph for **R**epresentation **L**earning* (PTGRL), que utiliza como base teórica o método de propagação PBG.

O *framework* funciona como um processo de inicialização contínuo. Primeiramente, ocorre um processo de atribuição de vetores de tamanho k, inicializados aleatoriamente, para cada vértice das redes. Esses vetores são as estruturas que armazenam as informações latentes aprendidas. Em seguida, o processo de propagação inicia em uma das redes bipartidas. A Figura 21 ilustra um ciclo do *framework*, no qual o PBG é executado na rede bipartida item-tag (Figura 21(a)) para que, em seguida, os vetores de *tags* com as informações latentes extraídas sejam utilizadas como inicialização do processo de propagação na rede bipartida usuário-tag (Figura 21(b)) e, por fim, a rede bipartida usuário-item é inicializada com os vetores latentes aprendidos anteriormente para estes dois tipos de vértices (Figura 21(c)). Este processo pode ser repetido e na segunda execução a rede usuário-tag inicia com as informações dos vetores aprendidas no ciclo anterior. Portanto, o segundo ciclo não inicia com informações aleatórias.

Figura 21 – Esquema do método proposto PTGRL para propagação em rede k-partida.



Fonte: Elaborada pelo autor.

No Algoritmo 8 são detalhas as etapas do *framework* proposto que tem como entrada os três grafos bipartidos UT, UI e IT que modelam as relações da folksonomia. O método PBG é executado em cada rede bipartida de forma que os vetores de entrada são atualizados com o resultado da propagação anterior, formando um processo sequencial de propagação de informação. Este processo é repetido até atingir um número máximo N_p pré-definido. Por fim, os vetores aprendidos na última execução para cada um dos vértices do tipo usuário são

considerados as suas respectivas novas representações. A matriz retornada $U_{n,k}$ corresponde a nova representação de cada vértice do tipo usuário:

$$\vec{U}_i = (x_1, x_2, ..., x_k) \tag{3.6}$$

em que x_k representa o k-ésimo fator latente, sendo k menor que o número total de *tags*. No caso desta abordagem, a matriz U é resultado da última propagação do grafo bipartido usuários-tag, como indica a linha 9 do algoritmo 8.

Algoritmo 8 – Algoritmo do *framework* PTGRL de propagação em rede k-partida.

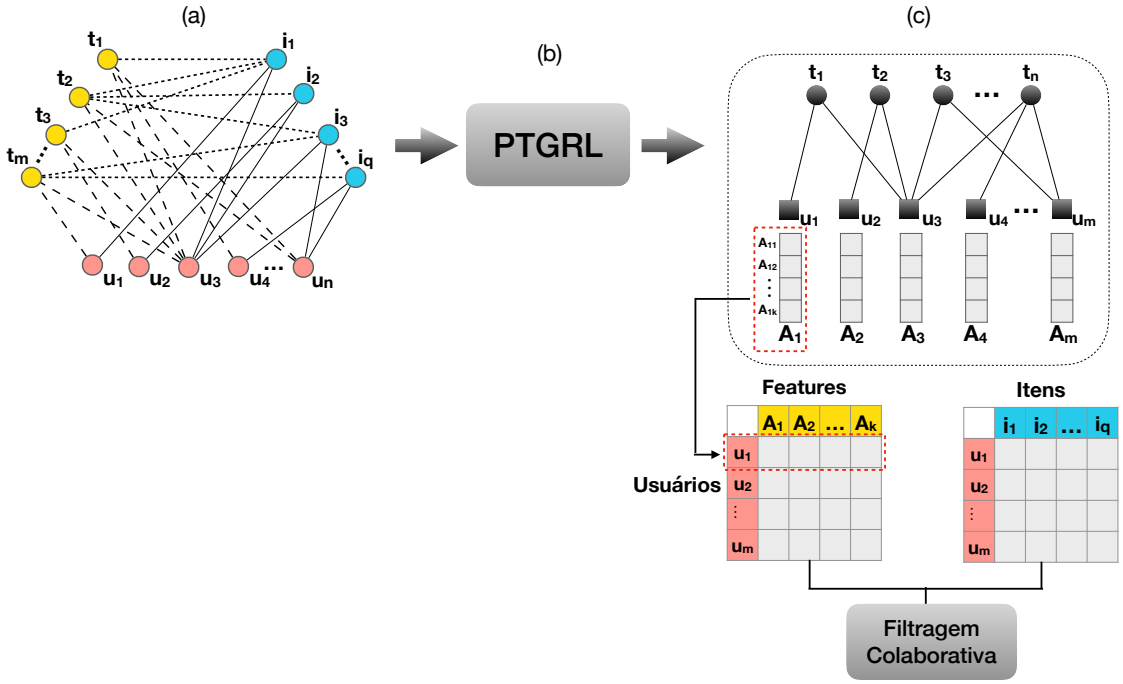
```
1:  procedimento TRIPBG(UT, UI, IT, Np)
2:      User ← Random()
3:      Item ← Random()
4:      Tag ← Random()
5:      para prop ← 1 to Np faça
6:          Item', Tag' ← PBG(Item, Tag, IT)
7:          User', Item'' ← PBG(User, Item', UI)
8:          User'', Tag'' ← PBG(User', Tag', UT)
9:          User ← User''
10:         Item ← Item''
11:         Tag ← Tag''
12:     fim para
13:     retorna User                ▷ Conjunto de vetores gerados para cada vértice u
14: fim procedimento
```

A matriz que contém as representações obtidas para cada usuário é utilizada no algoritmo de filtragem colaborativa para que recomendações personalizadas de itens sejam geradas. Na Figura 22 é ilustrada a abordagem de recomendação proposta, composta por três etapas principais. Primeiramente, os dados são modelados em rede tripartida formada pelas relações entre usuários, itens e tags. Em seguida, o *framework* de propagação em grafo tripartido é usado a extração de novas representações vetoriais (*features*) para os usuários. Finalmente, a representação e as informações dos itens são utilizados pela filtragem colaborativa para gerar recomendações.

Nesta abordagem é utilizada a FC baseada em memória e em usuário, detalhada anteriormente. Porém, para o caso dos dados de folksonomia existem duas possíveis formas para o cálculo da vizinhança N_u^k de um usuário u: considerando perfil de preferência por itens ou o perfil de atribuição de *tags* como dados de referência (RICCI; ROKACH; SHAPIRA, 2011). Deste modo, como o objetivo é integrar as informações do perfil de atribuição de *tags*, a matriz U de representações aprendidas para os usuários do grafo bipartido usuário-tag pelo *framework* é utilizada como a base para o cálculo das vizinhanças dos usuários.

A etapa de FC para a geração das recomendações personalizadas inicia com o cálculo de similaridade entre os usuários. Com base na matriz U e um dado tamanho de vizinhança k, a vizinhança N_u^k do usuário alvo u é obtida a partir do cálculo da similaridade de cosseno entre os

Figura 22 – Representação da abordagem de FC para recomendação em sistemas colaborativos de marcação, composta por três etapas: (a) modelagem da folksonomia como grafo tripartido (b) aplicação do *framework* de aprendizado de representação em grafo tripartido (c) recomendação utilizando as novas representações aprendidas para os usuários.



Fonte: Elaborada pelo autor.

usuários, dada por:

$$sim_{u,v} = \frac{U_u \cdot U_v}{\|U_u\| \, \|U_v\|} \tag{3.7}$$

Então, os perfis de preferência dos usuários (avaliações) que compõem a vizinhança do usuário alvo são combinados para prever a possível preferência do usuário-alvo u pelo item i, da seguinte forma:

$$score_{u,i} = \sum_{v \in N_u^k} sim_{u,v} \times UI_{v,i} \tag{3.8}$$

Por fim, a lista de recomendação é gerada com base nas pontuações obtidas para cada par (u,i). Como resultado, os k itens com as maiores pontuações serão recomendados.

Assim, é possível incorporar as informações de atribuição de *tags* com as de preferência dos usuários pelos itens. Em resumo, a matriz usuário-tag é usada para encontrar usuários semelhantes e a matriz user-item é usada para fazer recomendações de itens com base nos usuários semelhantes identificados. Desta forma, a relação ternária entre usuários, itens e *tags* é explorada.

3.4.3 Experimentos e Resultados

Para fins de avaliação de desempenho, a técnica proposta PTGRL foi comparada com o método PBG utilizado apenas na camada usuário-tag para o aprendizado de representações; e com a abordagem de FC que utiliza *autoencoder* profundo para o aprendizado de representações (ZUO *et al.*, 2016). Para a análise experimental dos métodos, foi utilizada a técnica de validação cruzada 5-*folds*, nas quais 80% (4-*folds*) dos dados de atribuição de *tags* (u,t,i) foi destinado para o treinamento e os 20% (1-*fold*) restantes para teste. Esse processo foi repetido usando todas as *folds* para teste. A informação conhecida no conjunto de treinamento é usada para gerar recomendações, enquanto o conjunto de teste é usado para avaliar o desempenho de recomendação dos algoritmos.

A confiança do usuário no sistema de recomendação aumenta à medida que esse usuário concorda com o conjunto de recomendações sugeridas pelo sistema (BOBADILLA *et al.*, 2013). Assim, para medir a capacidade preditiva dos métodos, foram empregadas as métricas de avaliação precisão e *recall*, detalhadas na Seção 3.3.3, e *rank score*. Geralmente, os usuários tendem a considerar os itens que aparecem no topo de uma lista de recomendações. Portanto, a experiência do usuário se torna menos agradável quando esses itens não são consistentes com a sua preferência (BOBADILLA *et al.*, 2013). Considerando isso, Zhou *et al.* (2007) propuseram uma nova métrica denominada *rank score*, definida como:

$$RankScore\,(L) = \frac{1}{n} \sum_{u \in U} \sum_{i \in E_i^u} \frac{pos_i\,(L)}{W_u}, \tag{3.9}$$

onde $pos_i(L)$ indica a posição do item i na lista de recomendações e W_u representa a quantidade total de itens desconhecidos pelo usuário u. Por exemplo, se houver 500 itens desconhecidos para um dado usuário u e se o item i que está no conjunto de teste, pois foi realmente avaliado por u, estiver na décima posição na lista de recomendações, o *rankScore* do item i será $\frac{10}{500} = 0.02$. Espera-se que um bom algoritmo de recomendação produza valores de *rankScore* pequenos, pois os itens que estão no conjunto de teste devem estar o mais próximo do topo da lista Zhou *et al.* (2007). O valor final do *rankScore* é dado pela média dos resultados obtidos para todos os usuários.

Em um primeiro momento, um estudo foi realizado para definir os valores dos parâmetros do método PBG, usado como base da técnica proposta PTGRL, que permitam a melhor extração de informações latentes das redes. A combinação de valores dos parâmetros tem relação direta com o resultado da recomendação e, portanto, foi estudado como o desempenho se comporta à medida que são variados os valores dos parâmetros: α, β, tamanho dos vetores (k) e quantidade de propagações locais e globais (*prop*). O PTGRL é constituído por um processo iterativo, definido neste experimento com duas repetições. As duas bases de dados e medidas de avaliação foram consideradas para a análise. Além disso, nesse estudo foram considerados os dois tipos de modelagem, binário e de frequência, uma vez que os pesos das arestas dos grafos bipartidos

influenciam diretamente no processo de propagação.

Para cada parâmetro foi definido o conjunto de valores utilizados na análise, são eles: $\alpha = \{0.1, 0.01, 0.001\}$, $\beta = \{0.0, 0.001\}$, $prop = \{10, 20, 50\}$ e $k = \{250, 500, 750\}$. A partir da definição desses conjuntos de valores, obteve-se um total de 54 combinações de parâmetros com as quais o método proposto foi executado. As Tabelas 7, 8, 9, 10 mostram os resultados das doze melhores combinações de parâmetros identificadas para cada base e medida de avaliação. Os resultados de precisão e *recall* foram observados para cada um dos 50 tamanhos de lista de recomendação adotados, assim como a média destes resultados. Esta análise permite que seja obtido um conjunto de valores *default* dos parâmetros que seja válido para diferentes bases e modelagens. O conjunto de valores $\alpha = 0.001$, $\beta = 0.0$, $prop = 10$ e $k = 750$ foram definidos como referência a partir dos resultados alcançados.

Tabela 7 – Resultado da análise experimental do desempenho do método PTGRL para a base *LastFm* e medida precisão, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os 50 tamanhos de lista de recomendação adotados. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros					Frequência				Binário			
Alfa	Beta	Prop	K	PropTrip	Média	L1	L2	L4	Média	L1	L2	L4
0.001	0	10	250	2	1.859	0.096	0.082	0.073	1.934	0.102	0.086	0.076
0.001	0	10	500	2	1.871	0.099	0.084	0.073	1.913	0.097	0.084	0.075
0.001	**0**	**10**	**750**	**2**	1.897	0.098	0.083	0.075	1.900	0.096	0.084	0.074
0.001	0	50	250	2	1.849	0.096	0.082	0.072	**1.948**	0.102	0.086	0.077
0.001	0	50	500	2	1.874	0.098	0.084	0.073	1.921	0.099	0.085	0.075
0.001	0	50	750	2	**1.906**	0.099	0.084	0.075	1.914	0.097	0.084	0.075
0.01	0	10	250	2	1.848	0.096	0.080	0.072	1.657	0.083	0.072	0.064
0.01	0	10	500	2	1.695	0.083	0.067	0.060	1.506	0.069	0.061	0.055
0.01	0	10	750	2	1.606	0.075	0.060	0.053	1.383	0.060	0.053	0.049
0.01	0	50	250	2	1.860	0.098	0.082	0.072	1.683	0.085	0.071	0.065
0.01	0	50	500	2	1.713	0.085	0.069	0.061	1.528	0.073	0.063	0.057
0.01	0	50	750	2	1.636	0.076	0.060	0.055	1.399	0.061	0.054	0.049

Fonte: Elaborada pelo autor.

A análise de comparação entre a abordagem proposta PTGRL e os demais métodos foi realizada com base nas medidas de precisão, *recall* e *rankScore*, descritas acima, e conforme a variação do tamanho L da lista de recomendação de itens aos usuários. Desta forma, analisar os resultados para diferentes tamanhos de lista permite que seja observado o comportamento dos métodos à medida que a quantidade de itens recomendados aumenta. Além disso, foi investigado o efeito dos pesos das arestas, binário e frequência, sobre o desempenho dos métodos.

A Figura 23 mostra os resultados de precisão, *recall* e *rankScore* para o conjunto de dados *LastFm*. Observa-se que os métodos PTGRL e PBG possuem melhores desempenhos do que o método baseado em *autoencoder* para a extração de representação. A Figura 24 mostra os resultados de precisão, *recall* e *rankScore* para o conjunto de dados *MovieLens*. Com relação às

Tabela 8 – Desempenho do método proposto PTGRL para a base *LastFm* e medida *recall*, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os 50 tamanhos de lista de recomendação adotados. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros					Frequência				Binário			
Alfa	Beta	Prop	K	PropTrip	Média	L1	L2	L4	Média	L1	L2	L4
0.001	0	10	250	2	**10.106**	0.034	0.056	0.072	10.381	0.038	0.060	0.075
0.001	0	10	500	2	9.942	0.034	0.057	0.071	10.369	0.036	0.058	0.075
0.001	**0**	**10**	**750**	**2**	9.972	0.033	0.055	0.073	10.237	0.036	0.058	0.076
0.001	0	50	250	2	9.972	0.034	0.057	0.071	**10.461**	0.038	0.059	0.076
0.001	0	50	500	2	10.035	0.035	0.058	0.073	10.308	0.037	0.060	0.076
0.001	0	50	750	2	10.053	0.034	0.056	0.074	10.325	0.036	0.058	0.076
0.01	0	10	250	2	9.828	0.032	0.053	0.070	9.480	0.034	0.053	0.068
0.01	0	10	500	2	8.995	0.028	0.042	0.053	8.816	0.028	0.046	0.059
0.01	0	10	750	2	8.485	0.024	0.035	0.047	8.232	0.025	0.039	0.051
0.01	0	50	250	2	9.824	0.033	0.054	0.069	9.612	0.034	0.052	0.070
0.01	0	50	500	2	9.083	0.027	0.044	0.055	8.921	0.030	0.047	0.060
0.01	0	50	750	2	8.581	0.024	0.036	0.049	8.306	0.025	0.039	0.050

Fonte: Elaborada pelo autor.

Tabela 9 – Resultado da análise experimental do desempenho do método PTGRL para a base *MovieLens* e medida *precisão*, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os 50 tamanhos de lista de recomendação adotados. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros					Frequência				Binário			
Alfa	Beta	Prop	K	PropTrip	Média	L1	L2	L4	Média	L1	L2	L4
0.001	0	10	250	2	0.852	0.045	0.036	0.032	0.873	0.042	0.034	0.031
0.001	0	10	500	2	0.888	0.049	0.037	0.034	0.912	0.044	0.037	0.032
0.001	**0**	**10**	**750**	**2**	0.871	0.045	0.037	0.034	0.922	0.047	0.039	0.034
0.001	0	50	250	2	0.886	0.046	0.039	0.034	0.884	0.041	0.036	0.032
0.001	0	50	500	2	**0.890**	0.047	0.038	0.034	0.915	0.046	0.038	0.033
0.001	0	50	750	2	0.885	0.047	0.038	0.034	**0.934**	0.049	0.039	0.035
0.01	0	10	250	2	0.753	0.036	0.030	0.027	0.805	0.032	0.029	0.026
0.01	0	10	500	2	0.716	0.031	0.027	0.026	0.790	0.032	0.029	0.026
0.01	0	10	750	2	0.674	0.027	0.025	0.023	0.779	0.031	0.028	0.025
0.01	0	50	250	2	0.761	0.035	0.031	0.028	0.816	0.035	0.031	0.027
0.01	0	50	500	2	0.718	0.031	0.027	0.025	0.796	0.034	0.028	0.026
0.01	0	50	750	2	0.676	0.028	0.025	0.023	0.781	0.030	0.027	0.025

Fonte: Elaborada pelo autor.

características das redes, em todos os cenários analisados os métodos mostraram pouca variância ao tipo de peso das arestas dotado para a modelagem das redes.

A partir dos experimentos realizados, é possível observar que o método proposto PTGRL

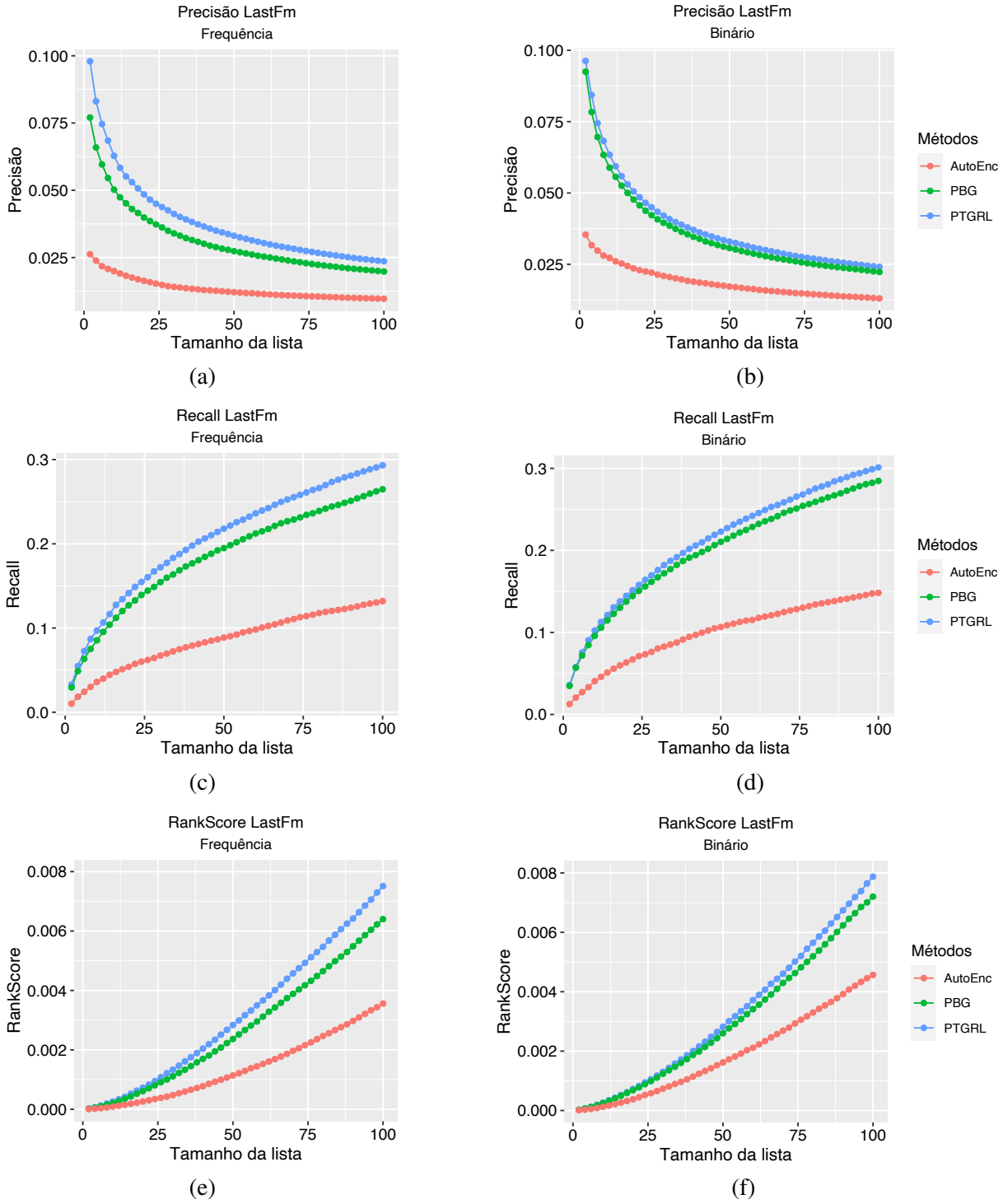
Tabela 10 – Resultados do método PTGRL para a base *MovieLens* e medida *recall*, conforme variação dos parâmetros alfa, beta, número de propagação e tamanho dos vetores. As colunas L1, L2 e L4 e média correspondem, respectivamente, aos resultados de precisão das listas de tamanho um, dois e quatro e a média dos resultados de todos os 50 tamanhos de lista de recomendação adotados. Os números em negrito destacam o conjunto de valores de parâmetros *default* e as maiores médias.

Parâmetros					Frequência				Binário			
Alfa	Beta	Prop	K	PropTrip	Média	L1	L2	L4	Média	L1	L2	L4
0.001	0	10	250	2	6.267	0.022	0.034	0.043	6.369	0.020	0.030	0.037
0.001	0	10	500	2	**6.557**	0.023	0.033	0.044	6.581	0.020	0.032	0.040
0.001	**0**	**10**	**750**	**2**	6.491	0.022	0.035	0.043	6.522	0.024	0.034	0.043
0.001	0	50	250	2	6.422	0.021	0.035	0.043	6.277	0.019	0.031	0.037
0.001	0	50	500	2	6.518	0.023	0.035	0.044	6.610	0.022	0.033	0.041
0.001	0	50	750	2	6.475	0.024	0.036	0.045	**6.763**	0.024	0.036	0.045
0.01	0	10	250	2	5.502	0.017	0.025	0.033	5.783	0.012	0.021	0.027
0.01	0	10	500	2	5.180	0.014	0.023	0.030	5.658	0.013	0.022	0.027
0.01	0	10	750	2	4.883	0.011	0.019	0.027	5.509	0.012	0.020	0.025
0.01	0	50	250	2	5.572	0.016	0.027	0.032	5.883	0.014	0.022	0.028
0.01	0	50	500	2	5.160	0.014	0.023	0.029	5.660	0.014	0.021	0.029
0.01	0	50	750	2	4.947	0.012	0.020	0.027	5.544	0.011	0.020	0.026

Fonte: Elaborada pelo autor.

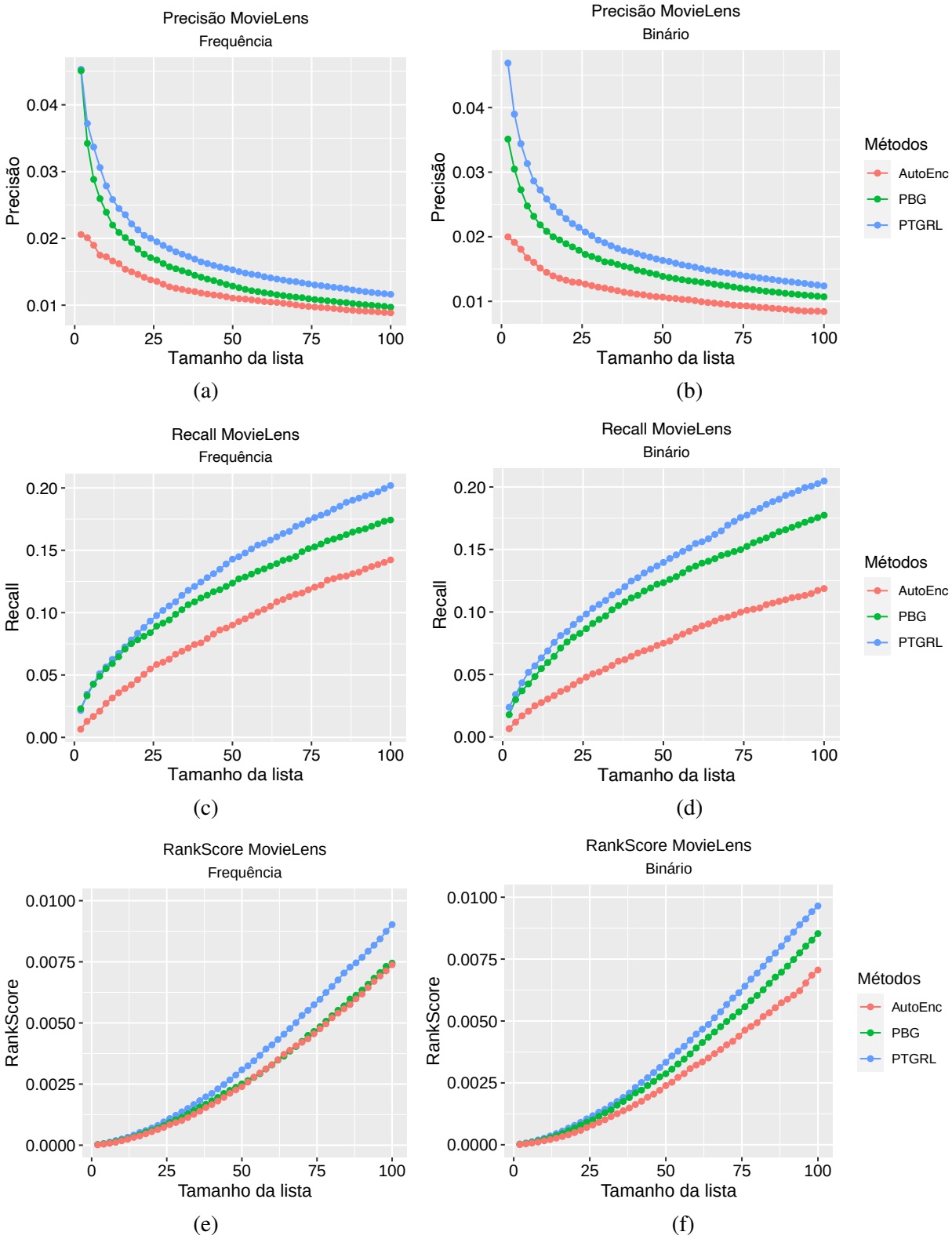
atinge os melhores resultados em todos os cenários. Além disso, é importante destacar a diferença entre os resultados das abordagens de recomendação que utilizam o método proposto PTGRL e o PBG. Isto mostra que o processo de transferência de informações entre as camadas bipartidas da estrutura *k*-partida, da técnica PTGRL, enriquece as representações aprendidas para cada vértice da rede, em comparação com as representações aprendidas utilizando apenas uma rede bipartida.

Figura 23 – Resultados das medidas precisão, *recall* e *rankScore* para a base *LastFm*.



Fonte: Elaborada pelo autor.

Figura 24 – Resultados das medidas precisão, *recall* e *rankScore* para a base *MovieLens*.



Fonte: Elaborada pelo autor.

3.4.4 Experimentos e Resultados com Redes Sintéticas

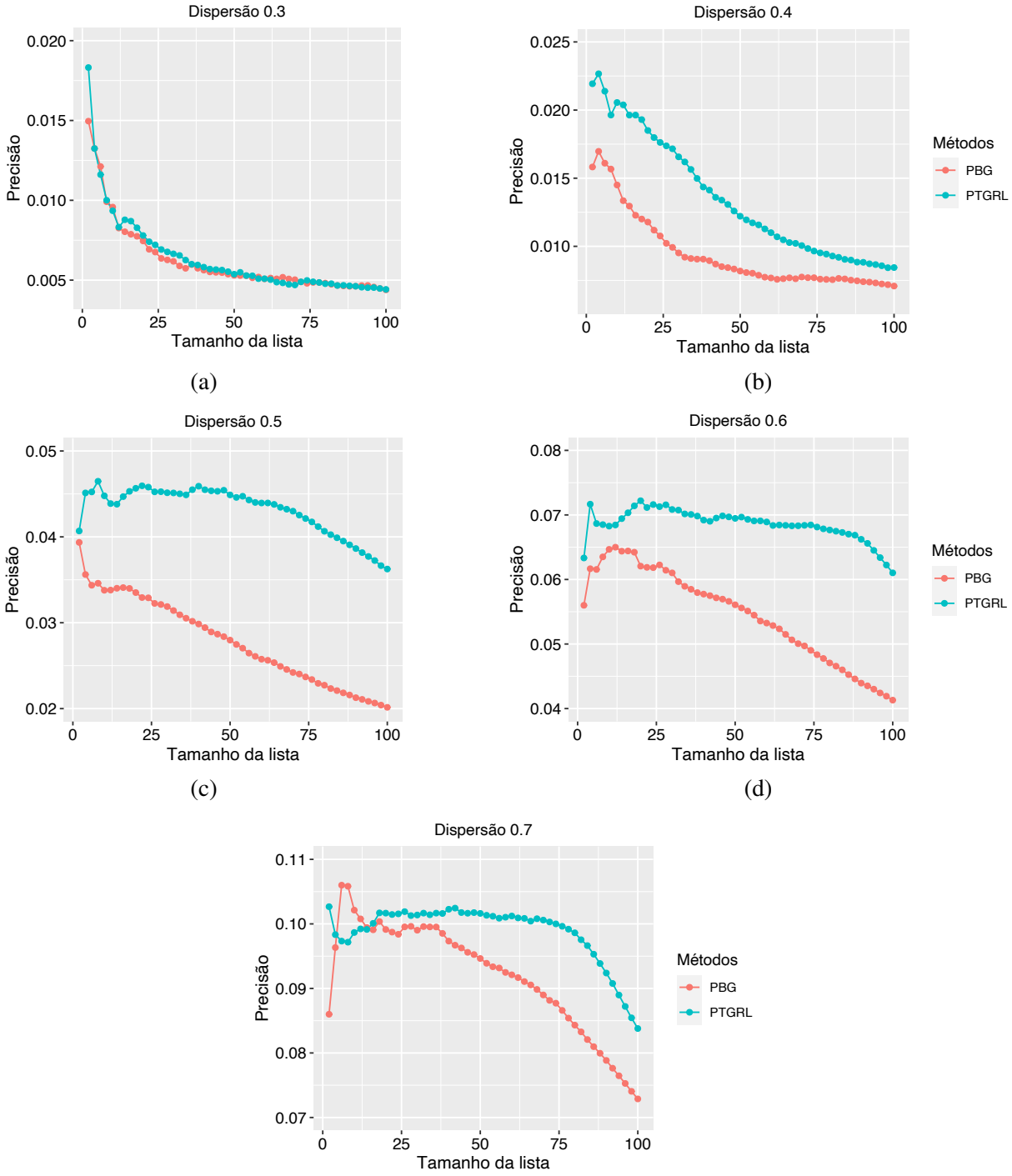
Além de realizar experimentos com dados reais, o desempenho dos métodos foi analisado em dados sintéticos. Para isso, foi utilizada a ferramenta de geração de redes sintéticas BNOC (VALEJO *et al.*, 2019), desenvolvida pela autora desta tese em colaboração com outros membros do grupo de pesquisa. Os detalhes da ferramenta estão descritos na Seção 2.1.3. Este tipo de análise possibilita uma investigação em redes geradas em um ambiente controlado para testes. Neste contexto, o nível de esparsidade (quantidade de arestas) das redes é uma característica importante que exerce influência no desempenho dos métodos. Em geral, redes de reposicionamento de fármacos doença-gene e redes textuais termo-documento são esparsas, enquanto redes biológicas de expressão gênica gene-paciente são caracterizadas por estruturas densas. Como as redes reais *MovieLens* e *LastFm* utilizadas nos experimentos são esparsas, a capacidade de controlar o nível de esparsidade das camadas da rede permite que o desempenho dos métodos PTGRL e PBG sejam analisados conforme a variação desta característica.

Para tanto, 6 redes tripartidas sintéticas compostas por três camadas de vértices de tamanho $\{250, 500, 500\}$ foram geradas. A camada de 250 vértices foi considerada a camada alvo de recomendação, representando os usuários, e as demais as camadas de *tags* e itens. Os pesos entre as arestas que conectam vértices entre as diferentes camadas das redes foram definidos como binários. As redes geradas possuem diferentes níveis de esparsidade, gerados pelo parâmetro de dispersão da ferramenta BNOC por meio do conjunto de valores $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. O parâmetro de dispersão permite controlar a esparsidade da estrutura topológica das redes, de forma que valores de dispersão inferiores $(d \approx 0)$ produzem redes de grau médio mais baixo, enquanto valores de dispersão mais altos $(d \approx 1)$ produzem redes com alto grau médio (VALEJO *et al.*, 2019).

Assim como nos demais experimentos deste capítulo, as medidas de precisão e *recall* foram empregadas para avaliação. A Figura 25 mostra o desempenho dos métodos para cada rede sintética. Pode-se observar que, para os dois métodos a taxa de precisão melhora conforme o aumento de densidade nas redes. No menor grau de dispersão, o método PTGRL tem resultado similar com o PBG, porém essa diferença aumenta no caso das redes mais densas. O mesmo é observado para a medida *recall*, como mostra a Figura 26. Esses resultados sugerem que o aumento de conexões no grafo acrescenta informações pertinentes para recomendação.

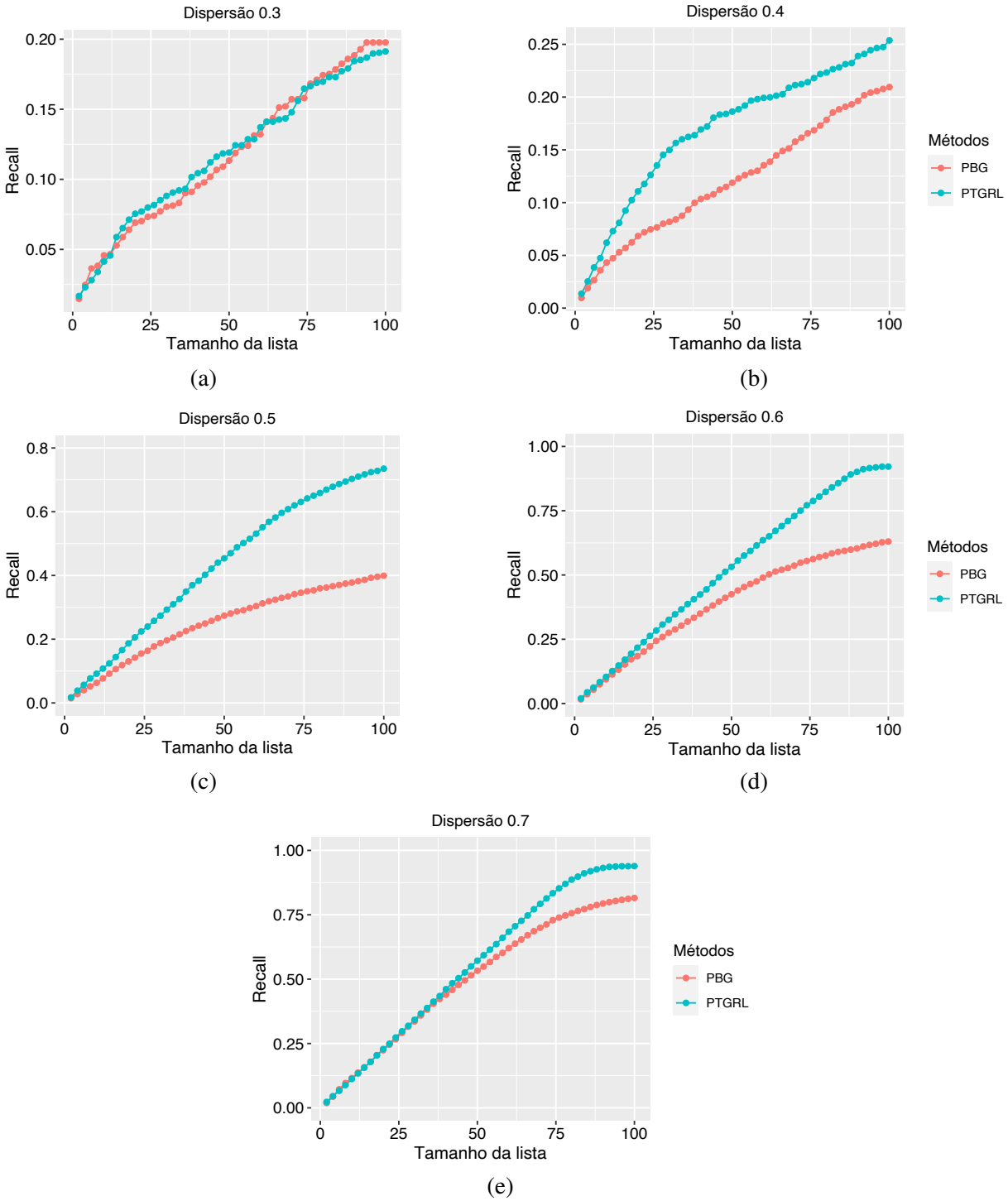
Nas mídias sociais ocorre o efeito "câmaras de eco", do inglês *echo chamber*, definido como o compartilhamento de informações, como opiniões e crenças, intensificado entre usuários que possuem a tendência a concordarem entre si (CINELLI *et al.*, 2021). Isto tende a restringir que um determinado grupo receba fontes diversificadas de conhecimento, reforçando a perspectiva e visão de quem os compõe. Este mesmo comportamento pode ser encontrado no contexto dos sistemas de recomendação (JIANG *et al.*, 2019; KALIMERIS *et al.*, 2021; TOMMASEL; RODRIGUEZ; GODOY, 2021; CINUS *et al.*, 2022). Desta forma, com a possibilidade da geração de redes sintéticas com estruturas de comunidades, ao longo da pesquisa foi identificada

Figura 25 – Resultados de precisão para redes com diferentes níveis de dispersão.



Fonte: Elaborada pelo autor.

Figura 26 – Resultados de *recall* para redes com diferentes níveis de dispersão.



Fonte: Elaborada pelo autor.

a possibilidade da elaboração de medidas que permitam a análise dos métodos conforme a importância das comunidades dos usuários para o processo de recomendação. Abaixo são descritas as duas medidas propostas para esta análise.

Cada usuário j pertence a uma comunidade dada pelo índice $C(j)$. Para obter o conjunto de usuários que estão na mesma comunidade que j fazemos $C_j = \{j' \in j | C(j) = C(j') \wedge j \neq j'\}$. As medidas foram propostas para verificar como os recomendadores estão indicando itens populares ou "preferidos" pela comunidade. A primeira medida $\theta_1(j)$ calcula a taxa de recomendação de itens i conectados aos usuários j' pertencentes a comunidade do usuário alvo j. Se os demais usuários da comunidade do usuário j estão relacionados com 10 itens e o recomendador recomendou 5 dos 10 itens, a porcentagem será de 50%, conforme:

$$\theta_1(j) = \frac{1}{|I_{C_j}|} \sum_{i \in I_{C_j}} x \begin{cases} x = 1, & \text{if } e_{j',i} \in E_{C_j} \\ x = 0, & \text{caso contrário} \end{cases} \tag{3.10}$$

em que I_{C_j} corresponde ao conjunto de itens que possuem conexão com os usuários j' do conjunto C_j e E_{C_j} denota o conjunto de arestas que conectam itens com os usuários j' pertencentes a comunidade C_j.

A segunda medida $\theta_2(j)$ avalia o quanto os itens relacionados com a comunidade C_j, a qual o usuário alvo j pertence, estão sendo recomendados. Imagine que, os demais usuários da comunidade tem relação com dois itens i_1 e i_2 e que o i_1 tem relação com 7 dos 10 desses usuários, enquanto o item i_2 tem relação com apenas 4 desses usuários. Se apenas o item i_1 estiver na lista de recomendação do usuário j, o valor dessa medida será $\frac{7}{11} = 0.63$. Se os dois itens estiverem na lista de recomendação, o valor dessa medida será $\frac{11}{11} = 1$. Assim, $\theta_2(j)$ pode ser definida como:

$$\theta_2(j) = \frac{1}{|E_{C_j}|} \sum_{j' \in C_j} \phi(i, j') \tag{3.11}$$

$$\phi(i, j') = \sum_{j' \in C_j} x \begin{cases} x = 1, & \text{if } e_{j',i} \in E_{C_j} \\ x = 0, & \text{caso contrário} \end{cases} \tag{3.12}$$

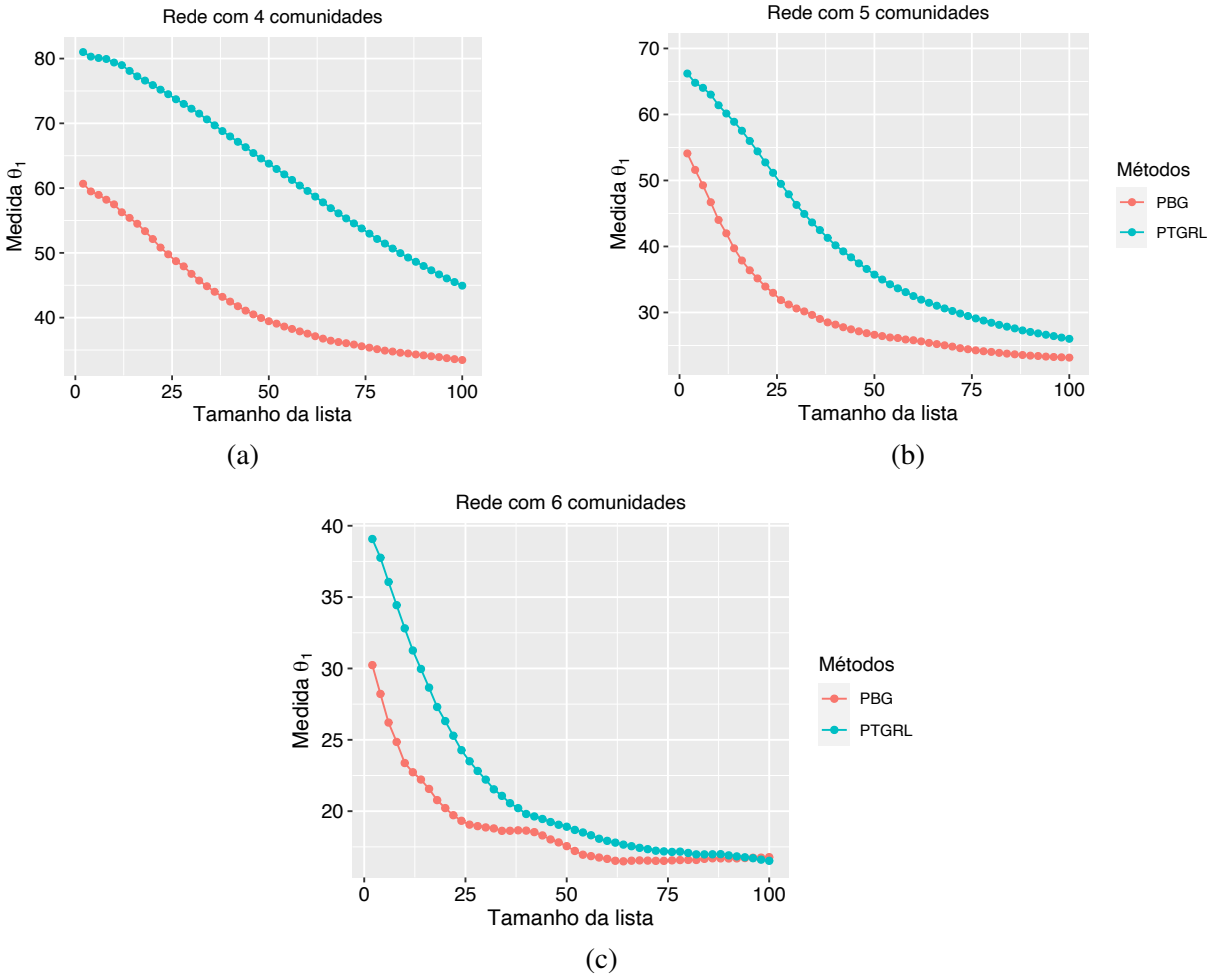
em que $\left|E_{C_j}\right|$ a denota a cardinalidade do conjunto de arestas que conectam itens com os usuários pertencentes a comunidade C_j do usuário j. Note que, estas medidas são calculadas para cada usuário. Sendo assim, o valor final é dado pela média dos valores obtidos para a recomendação gerada para cada usuário.

Experimentos foram realizados novamente com os métodos PTGRL e PBG para avaliar os resultados em relação às medidas propostas θ_1 e θ_2. Para analisar o comportamento das medidas em relação ao número de comunidades, foram geradas três redes tripartidas sintéticas compostas por três camadas de vértices de tamanho $\{250, 500, 500\}$. Cada rede possui uma das seguintes quantidades de comunidades em todas as suas camadas de vértices: $\{4, 5, 6\}$. Como no

experimento anterior, a camada de 250 vértices foi considerada como a camada de usuários e as demais como as camadas de *tags* e itens. O nível de dispersão 0.4 foi utilizado em todas as redes.

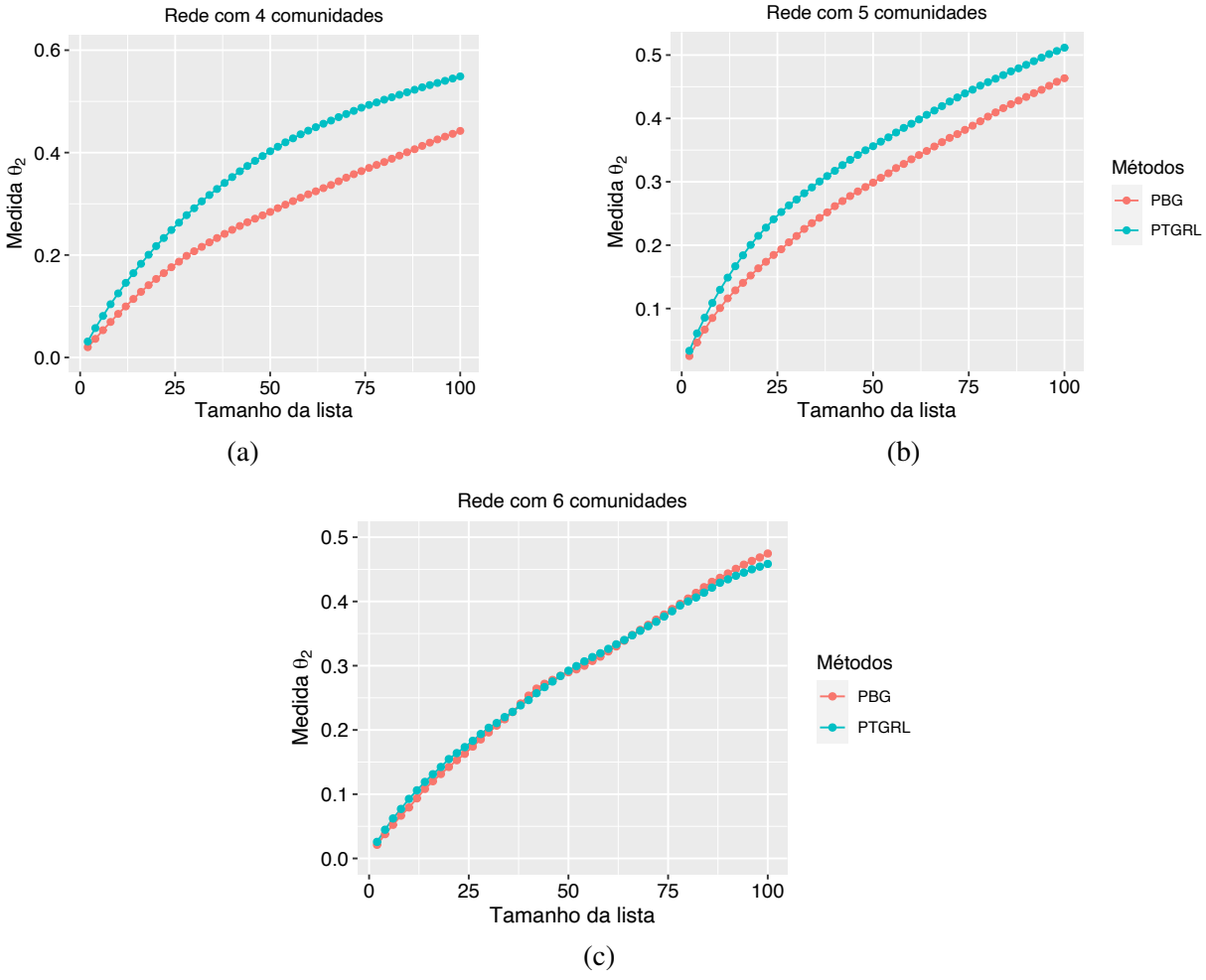
Na Figura 27 são mostrados os resultados para a medida θ_1. É possível perceber que o método proposto PTGRL tende a recomendar mais itens relacionados com os usuários pertencentes a mesma comunidade do usuário alvo que recebe as recomendações. Dependendo da aplicação, esta característica pode ser vantajosa, pois o método consegue captar melhor a preferência e os relacionamentos que os elementos de uma mesma comunidade possuem. A Figura 28 mostra os resultados para a medida θ_2. Nota-se que não há tanta variação conforme o aumento da quantidade de comunidades, como ocorre na medida θ_1. Isto pode indicar que os itens recomendados a um usuário alvo tendem a ter a mesma taxa de conexão com os demais usuários da comunidade, independente da quantidade de comunidades.

Figura 27 – Resultados da medida θ_1 para redes com diferentes quantidades de comunidades.



Fonte: Elaborada pelo autor.

Figura 28 – Resultados da medida θ_2 para redes com diferentes quantidades de comunidades.



Fonte: Elaborada pelo autor.

CAPÍTULO

4

PROPAGAÇÃO EM REDES K-PARTIDAS PARA PREDIÇÃO DE ASSOCIAÇÃO

Neste capítulo é apresentada uma abordagem desenvolvida para a predição de associação (conexão) entre lncRNAs e doenças. Na Seção 4.1 são apresentados os conceitos necessários para compreensão do contexto dos dados biológicos. A Seção 4.2 apresenta os trabalhos relacionados com este projeto. A fundamentação da proposta desenvolvida é apresentada na Seção 4.3. Por fim, os experimentos e resultado obtidos são descritos na Seção 4.4.

4.1 Introdução

Na década de 90, o Projeto Genoma Humano promoveu rápidos avanços nos experimentos de sequenciamento genômico para a obtenção das informações básicas da sequência genômica humana. Através de tecnologias de sequenciamento é possível ler uma amostra de DNA e determinar a sequência de nucleotídeos que a compõem. Essa sequência é uma peça fundamental para o conhecimento e estudo do genoma, que carrega a informação hereditária de um organismo. Assim, os constantes avanços nas tecnologias de sequenciamento permitiram o desenvolvimento de análises genômicas e transcriptômicas mais sofisticadas, como descoberta de funções biológicas associadas aos genes, mecanismos de regulação genética e processos celulares (FANG; FULLWOOD, 2016).

O início dos avanços da pesquisa genômica teve como objetivo principal o estudo do DNA e a identificação de seus genes. Esse estudo forneceu importantes percepções e mudou a compreensão da biologia (BRENT, 2000; EDDY, 2001). Porém, por volta dos anos 2000, muitas pesquisas apontaram a identificação de um número surpreendente de RNAs que não codificam proteínas e ampliaram os estudos sobre os genes que produzem RNAs não-codificantes. Assim, os RNAs podem ser divididos em RNAs codificantes e RNAs não-codificantes (ncRNAs).

Os RNAs não-codificantes são também chamados de RNAs funcionais, uma vez que

desempenham diferentes funções celulares, como estruturais, reguladoras e metabólicas (EDDY, 2001; MATTICK; MAKUNIN, 2006). Hoje, sabe-se que os ncRNAs compõem uma grande parte do genoma humano (DINGER *et al.*, 2008; FRITH; PHEASANT; MATTICK, 2005) e podem ser classificados em dois grupos com base em seu tamanho (FANG; FULLWOOD, 2016). Os ncRNAs com menos de 200 nucleotídeos (nt) de comprimento são classificados como RNAs curtos, tais como o microRNA (miRNA), *piwi-interacting* RNA (piRNA), RNA transportador (tRNA) e ribossomal (rRNA). Por outro lado, ncRNAs maiores que 200 nt são classificados como RNA longos (lncRNAs).

Muitos estudos evidenciam que os lncRNAs possuem relações com o desenvolvimento de várias doenças humanas complexas e importantes processos biológicos (WAPINSKI; CHANG, 2011; YANG *et al.*, 2014a). Porém, ainda é necessário avançar em direção a compreensão do papel dos lncRNAs na causa de doenças, devido à dificuldade em desvendar a concepção do mecanismo que está por trás desta relação (YUAN *et al.*, 2021). Estudos demonstraram que a alteração e desregulação de lncRNAs podem desencadear várias doenças cardiovasculares, neurodegenerativas e vários tipos de câncer (WAPINSKI; CHANG, 2011; FANG; FULLWOOD, 2016). Assim, a identificação de associações entre doenças e lncRNAs é um ponto essencial para ampliar a compreensão dos mecanismos moleculares responsáveis pelos desenvolvimentos de doenças, identificação de biomarcadores e alvos terapêuticos a fim de aumentar a eficácia e eficiência do diagnóstico e tratamento de doenças (MORI *et al.*, 2018; YAN *et al.*, 2020).

Além do papel como potenciais biomarcadores para indicação de funções patológicas em um organismo, estudos revelaram que os lncRNAs estão relacionados as respostas a agentes farmacológicos, tornando-se peças importantes para o reposicionamento de fármacos, como a descoberta de drogas anticancerígenas para a terapia do câncer (LING; FABBRI; CALIN, 2013; LI *et al.*, 2016). Além disso, a descoberta de associações entre doenças e lncRNAs também contribui para a investigação do papel de genes codificadores de proteínas em doenças e, consequentemente, sua importância para o diagnóstico, prognóstico e tratamento.

Devido ao avanço nas tecnologias e projetos de sequenciamento, a geração de dados de RNA aumentou significativamente nos últimos anos (SHENDURE *et al.*, 2017). O conhecimento atual sobre as associações entre lncRNAs e doenças ainda não é satisfatório dada a complexidade da biologia humana (YAN *et al.*, 2020) e do processo de análise envolvido. Experimentos biológicos para a identificação e validação de associação entre elementos são demorados e caros. Além disso, organizar e integrar os dados de associações registradas na literatura e em bancos de dados adquiridos por diferentes estudos experimentais e tecnologias é um processo nada trivial. Portanto, devido à crescente quantidade disponível destas informações biológicas, a identificação de relações entre lncRNAs e doenças demanda uma abordagem interdisciplinar, na qual estão envolvidos métodos experimentais desenvolvidos em laboratórios e métodos computacionais.

Neste contexto, há uma necessidade iminente de metodologias computacionais que possam efetivamente inferir relevantes associações entre doenças e lncRNAs em larga escala. As

associações já conhecidas e validadas são usadas como base para criar uma série de métodos computacionais para a predição de novas associações. O desenvolvimento de métodos computacionais voltados para a identificação de potenciais associações entre lncRNAs e doenças auxiliam a otimizar e priorizar o que é mais relevante de ser analisado *in vitro* por especialistas da área, a fim de beneficiar a identificação da influência de lncRNAs alvos no diagnóstico, tratamento e prevenção de doenças.

Neste estudo, uma nova abordagem foi proposta com base em um algoritmo de propagação em rede tripartida para inferir potenciais interações entre doenças e lncRNAs. A proposta foi inspirada no método não supervisionado de propagação em redes bipartidas PBG (FALEIROS; VALEJO; LOPES, 2020), estendendo a estratégia de propagação em redes bipartidas do PBG para um *framework* de propagação que possibilita a troca de informações entre as camadas bipartidas de uma rede tripartida. A abordagem proposta foi avaliada através de uma investigação detalhada de seu desempenho em comparação com outros métodos da literatura.

4.2 Trabalhos relacionados

Devido à importância da descoberta de novas associações entre doenças e lncRNAs, muitas abordagens computacionais têm sido desenvolvidas para a identificação automática e em larga escala destas relações. Os trabalhos se diferenciam em vários aspectos, desde a teoria matemática empregada até os tipos de dados utilizados. O estado da arte possui trabalhos baseados em aprendizado de máquina semissupervisionado e não supervisionado, alocação de recursos, propagação em grafos, fatoração de matrizes, dentre outros. Em relação aos dados, encontra-se na literatura trabalhos que abordam diferentes fontes de dados para a investigação de princípios biológicos relacionados com a descoberta de novas associações entre lncRNAs e doenças, tais como similaridade entre lncRNAs, similaridade semântica entre doenças, associação entre miRNA e lncRNAs, interação entre proteínas etc.

Chen e Yan (2013) desenvolveram um *framework* semissupervisionado baseado em *Laplacian Regularized Least Squares*, denominado LRLSLDA, para a identificação de pontenciais novas associações a partir de informações conhecidas de lncRNA-doença, similaridade semânticas entre doenças e similaridade funcional entre os lncRNAs. Ganegoda *et al.* (2015) propuseram o método KRWRH que utiliza *Random Walk with Restart* em uma rede heterogênea para a predição de um tipo específico de lncRNA com doenças, considerando matriz de similaridade dos lincRNAs gerada a partir de dados de expressão, matriz de similaridade das doenças e as associações entre lincRNA e doença validadas experimentalmente. Ding *et al.* (2018) propuseram o TPGLDA que utiliza o método de alocação de recursos em uma rede tripartida que integra associações gene-doença com associações lncRNA-doença, juntamente com informações de similaridades entre lncRNA e doenças, com o objetivo de representar heterogeneidade do problema de identificação de associação de doença de lncRNA.

Yang *et al.* (2014a) construíram uma rede bipartida projetada através do método de alocação de recurso (ZHOU *et al.*, 2007) em dados de associações de doenças tanto com os lncRNAs como os genes codificantes e utilizaram um algoritmo de propagação, baseado em Zhou *et al.* (2004), para realizar a predição de novas associações. Zhang *et al.* (2017) desenvolveram um *framework* baseado em rede heterogênea, denominado LncRDNetFlow, que utiliza um algoritmo de propagação de fluxo para percorrer uma rede construída com base em diferentes fontes de dados, como similaridade entre as doenças, interações entre proteínas, similaridade entre lncRNAs e interações entre estas três entidades biológicas (lncRNA-doença, doença-proteína e lncRNA-proteína), para inferir associações de lncRNA-doença.

Sun *et al.* (2014) utilizaram o método de caminhada aleatória com reinício para inferir potenciais associações de lncRNA para uma dada doença de interesse a partir da integração das redes lncRNA-doença, similaridade entre doenças e similaridade funcional entre lncRNAs. Zhang *et al.* (2020) propuseram um método baseado em projeção vetorial para gerar pontuações que indicam a possibilidade de existência de novas associações a partir de uma rede formada por associações conhecidas de lncRNA-doença, similaridade semânticas entre doenças e similaridade funcional entre os lncRNAs. Nesta mesma linha, Chen *et al.* (2020) propuseram um novo método de predição de lncRNA-doença, denominado LRLSSP, que utiliza *Laplacian regularized least squares* para realizar uma estimativa inicial de novas potenciais associações entre lncRNAs e doenças, que serve como base para a aplicação de projeção vetorial responsável por refinar e gerar a predição final de associações.

Muitos estudos utilizam *matrix completion* e fatoração de matrizes para identificar potenciais associações lncRNA-doença. Lu *et al.* (2018) propuseram o uso de Análise de Componentes Principais para extrair vetores de características para as doenças e lncRNAs, com base em dados de associações entre doença-gene e lncRNA-doença, utilizados pelo método *Inductive Matrix Completion* (IMC) para preencher a matriz lncRNA-doença indicando potenciais associações. Fu *et al.* (2018) utilizaram diferentes fontes de dados heterogêneas para modelar matrizes que representam diferentes relações entre diferentes dados biológicos, processadas por um modelo de fatoração de matriz que explora a relação latente entre objetos do mesmo tipo e de tipos diferentes e realiza a predição de potenciais associações entre lncRNAs e doenças. Yu *et al.* (2018) propuseram o uso de fatoração de matrizes em uma rede heterogênea composta por inter e intra associações entre diferentes tipos de entidades, como miRNAs, genes, lncRNAs, doenças e fármacos. Xuan *et al.* (2019) investigaram a construção de uma rede ponderada lncRNA-doença gerada a partir de três redes bipartidas geradas pelas relações entre lncRNA-doença-miRNA para a inferência de associações entre lncRNAs e doenças com fatoração de matrizes.

4.3 Proposta

Diante dos desafios da detecção de possíveis relações entre lncRNAs e doenças e da lacuna existente do método PBG (FALEIROS; VALEJO; LOPES, 2020) para redes k-partidas, esta proposta tem dois objetivos principais: (i) o desenvolvimento de uma nova abordagem computacional para a identificação de associações entre lncRNAs e doenças e (ii) a criação e análise experimental de um *framework* de propagação em rede tripartida para a predição de novas potenciais conexões entre vértices de diferentes tipos. Para tal, é utilizada a fundamentação teórica o método de propagação em redes bipartidas PBG.

Uma das bases teóricas do PBG é a fatoração de matrizes não-negativa (NMF) (FALEIROS; VALEJO; LOPES, 2020; LEE; SEUNG, 1999), que tem sido amplamente utilizada para analisar dados em diversas aplicações. Esse método fatora uma matriz R_{NxM} em duas matrizes W e H com menor dimensão, seguindo a propriedade de que as matrizes possuem apenas elementos com valores não negativos. O produto das matrizes decompostas se aproxima o melhor possível da matriz original, como $R_{NxM} \approx W_{NxP}H_{PxM}$. O PBG converte os princípios da NMF em um algoritmo baseado na estrutura de um grafo bipartido. Assim, esse método computacional permite realizar tarefas não supervisionadas, como extração de representação (*features*) e agrupamento de dados. O algoritmo PBG desempenha um processo de propagação iterativo em uma rede bipartida, em que a informação latente relacionada a um vértice influencia as informações de seus vizinhos.

Neste contexto, duas questões que cercam o método PBG possibilitam o desenvolvimento de novas pesquisas. Primeiramente, este método possui bons resultados para o problema de extração de tópicos quando comparado com métodos da literatura, provendo indícios de que pode ser utilizado em outras aplicações. Além disso, o PBG possui uma lacuna para o contexto de redes tripartidas, fato que instigou o desenvolvimento de uma abordagem experimental em que a informação latente obtida para os vértices de uma camada da rede tripartida é propagada para a segunda partição que compõem a rede, funcionando como um processo de inicialização contínuo, para a obtenção de vetores de representação para cada vértice da rede. Com isso, foi desenvolvido o *framework* **P**ropagation in **T**ripartite **G**raph for **A**ssociation **P**rediction (PTGAP), que utiliza como entrada um grafo tripartido para a geração de vetores de representação para cada vértice e, com isso, gera como saída uma pontuação que indica a possibilidade da ocorrência de conexão entre cada par de vértice.

Esta proposta utiliza uma rede tripartida que integra informações biológicas derivadas das relações entre doenças, genes e lncRNAs. Na abordagem proposta, a construção da rede é baseada nas matrizes de interação doença-gene e doença-lncRNA. Assim, a etapa inicial é a construção da rede tripartida $GR(D,L,T,E)$, em que $D = \{d_1, d_2, ..., d_m\}$, $L = \{l_1, l_2, ..., l_n\}$ e $T = \{t_1, t_2, ..., t_q\}$, correspondem aos conjuntos de m doenças, n lncRNAs e q genes, respectivamente, interligados por um conjunto de arestas denotado por E. $A^{LD}_{n,m}$ representa a matriz de adjacência das interações entre doenças e lncRNAs, de forma que se existir uma associação verificada entre um lncRNA l_i

e uma doença d_j o elemento $A^{LD}(i,j)$ recebe o valor 1, caso contrário 0. Por sua vez, $A^{TD}_{q,m}$ denota a matriz de adjacência correspondente as conexões entre genes e doenças, em que $A^{TD}(i,j) = 1$ se t_i e d_j possuem relação verificada e $A^{TD}(i,j) = 0$, caso contrário. Além disso, duas redes de similaridade entre lncRNAs $A^{L}_{n,n}$ e doenças $A^{D}_{m,m}$ foram utilizadas para os casos de inferência para vértices desconexos.

Uma vez que, uma rede tripartida pode ser decomposta em duas redes bipartidas, foi possível constituir o PBG como a parte central do *framework* PTGAP. Desta forma, dado uma rede tripartida lncRNA-doença-gene, duas redes bipartidas doença-lncRNA e doença-gene são geradas. Primeiramente, para cada vértice das redes é associado um vetor de tamanho k inicializado aleatoriamente, responsável por armazenar as informações latentes. Então, o processo de propagação inicia em uma das redes bipartidas, neste caso a rede doença-gene. Para esta rede, a propagação foi realizada considerando os vértices em T (genes) como alvos e os vértices em D (doenças) como pontes, de forma que as informações da camada T são propagadas para a camada D e, ao final, retornam para T conforme a estrutura da rede. A partir desta execução do PBG os vetores são inicializados com informações da rede doença-gene e são utilizados na execução da propagação na partição doença-lncRNA. Neste segundo passo, a propagação é executada considerando os vértices L (lncRNAs) como alvos e os vértices em D (doenças) como pontes. Ao término da execução na segunda partição, os vetores para os vértices em D são utilizados para uma nova execução do PBG na primeira partição doença-gene. Este processo é repetido até atingir um número máximo de iterações definidas à priori. A Figura 29 ilustra as etapas principais do funcionamento do PTGAP.

Por fim, as matrizes $A_{n,k}$ e $B_{m,k}$ compostas pelos vetores computados para os vértices em L e D, respectivamente, são combinadas para que sejam obtidas as pontuações que indicam as possibilidades de existência de associações entre os pares de lncRNAs e doenças. Para isso, a matriz $AS^{LD}_{m,n}$ que contém as probabilidades (pontuação) de associação entre os lncRNAs e as doenças é obtida da seguinte forma:

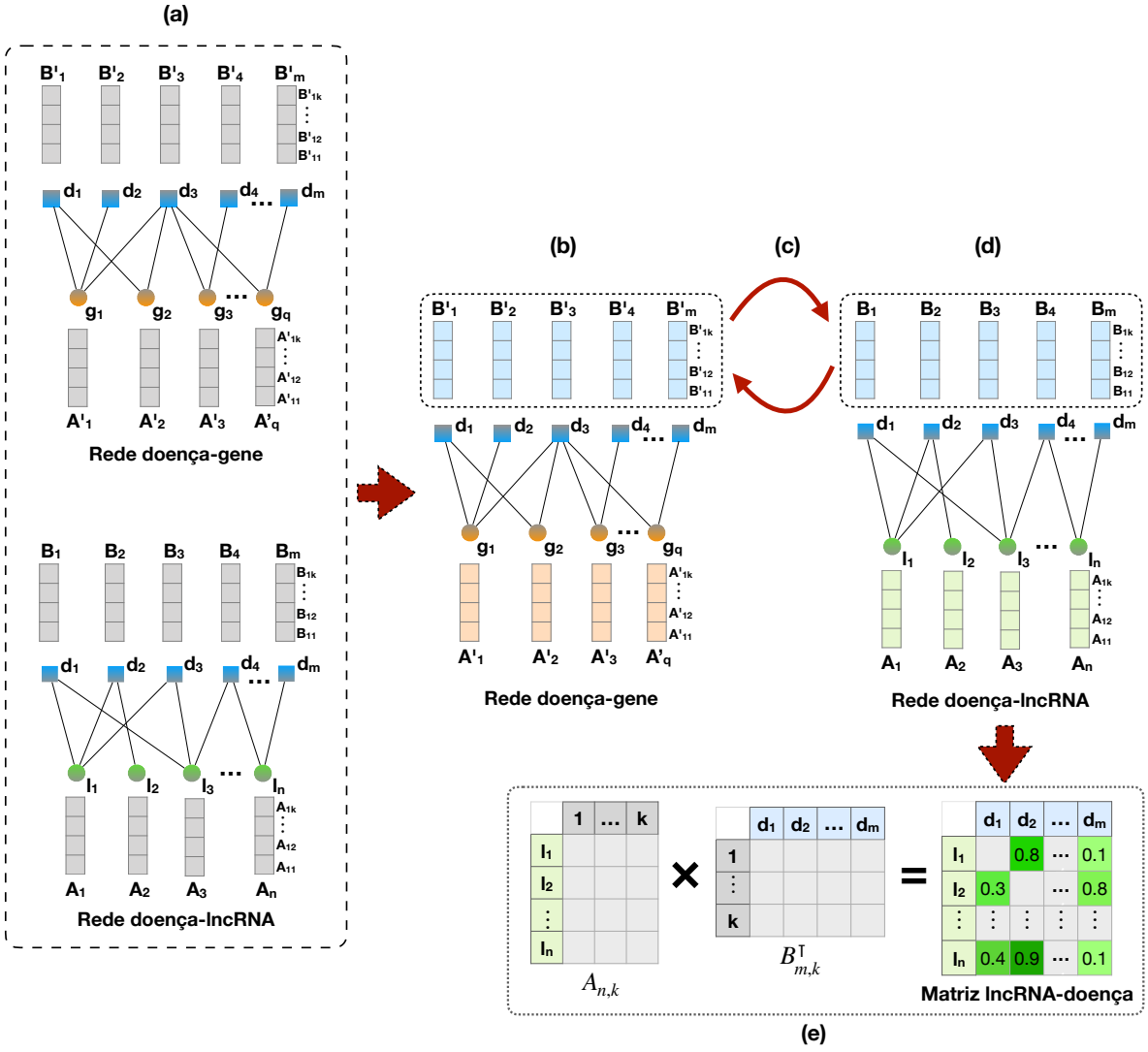
$$AS^{DL} = A_{n,k} \times B^{\mathsf{T}}_{m,k}, \tag{4.1}$$

em que $A_{n,k}$ e $B_{m,k}$ correspondem, respectivamente, as matrizes que contém os vetores obtidos para cada vértice do tipo doença e lncRNA obtidos pelo método PTGAP. Assim, quanto maior a pontuação, maior a certeza de que o lncRNA pode ter uma relação com uma doença específica.

4.4 Experimentos e Resultados

A existência de bases de dados de associação entre lncRNA-doença de alta qualidade é extremamente importante para o estudo e compreensão de como os lncRNAs contribuem para o desenvolvimento de doenças complexas e a inferência de novas relações lncRNA-doença. Para

Figura 29 – Diagrama do funcionamento do PTGAP composto por quatro etapas: (a) inicialização aleatória dos vetores de cada vértice, (b) execução do PBG na primeira partição doença-gene, (c) transferência de informação dos vetores dos vértices em D, obtidos pelo PBG, na rede doença-gene para os vetores dos vértices em D na rede doença-lncRNA, (d) execução do PBG na segunda partição doença-lncRNA utilizando as informações transferidas na etapa anterior e (e) predição das pontuações das possíveis associações entre os pares de lncRNAs-doenças.



Fonte: Elaborada pelo autor.

tanto, Chen *et al.* (2012) desenvolveram o conjunto de dados *LncRNADisease*[1] por meio da seleção manual de associações lncRNA-doença validadas experimentalmente e publicadas na literatura. Esse conjunto de dados foi utilizado para a realização de experimentos em dados reais desta proposta. Para fornecer uma visão geral dos dados, na Tabela 11 são mostradas as estatísticas da rede de associação lncRNA-doença-gene utilizada.

A rede tripartida lncRNA-doença-gene foi utilizada como a base principal para a predição de associações. Além disso, duas redes homogêneas foram utilizadas para prover informações

[1] Publicamente disponível em http://cmbi.bjmu.edu.cn/lncrnadisease

para os casos em que os vértices da rede possuem apenas uma conexão e, consequentemente, se tornam desconexos da rede quando a associação é removida para o processo de validação cruzada. Abaixo são descritos em detalhes os dados utilizados neste trabalho conforme coletado [2] e detalhado por Ding *et al.* (2018):

Associações lncRNA–doença e gene-doenças: A partir da base *LncRNADisease*, Ding *et al.* (2018) realizaram uma filtragem para remover os lncRNAs sem dados de expressão na base *ArrayExpress* [3] e as doenças sem registos na base *Disease Ontology*[4] que fornece uma ontologia que unifica as variadas terminologias utilizada para definir as doenças. Com base nas doenças relacionadas aos lncRNAs selecionadas, 5.212 associações entre genes e doenças do banco de dados DisGeNET[5] foram coletadas para formar a partição doença-gene do grafo tripartido. Assim, a rede tripartida utilizada possui 115 lncRNAs, 178 doenças, 1.415 genes e 540 associações lncRNA-doença, utilizadas como referência na validação cruzada realizada para a avaliação experimental, como detalhado na Tabela 11.

Rede de similaridade entre LncRNAs: Foram coletados os dados de expressão dos LncRNAs selecionados na base *ArrayExpres* para a construção de uma rede de similaridade entre eles. Para isso, o Coeficiente de Correlação de *Spearman* foi utilizado para calcular a similaridade entre os perfis de expressão entre cada par de lncRNAs.

Rede de similaridade entre doenças: Ding *et al.* (2018) utilizaram a definição do cálculo de similaridade semântica entre doenças descrito por Wang *et al.* (2010). As informações semânticas das doenças foram obtidas do *Medical Subject Headings* (MeSH), pertencente a *National Library of Medicine* (NLM)[6], sendo um vocabulário que cataloga termos de áreas relacionadas à saúde. Assim, uma doença pode ser vista como um termo e as suas relações semânticas com as demais doenças podem ser representadas pela estrutura de um Grafo Acíclico Dirigido (GAD). A similaridade semântica entre um par de doenças é dada pelo grau de correspondência entre seus GADs.

A validação cruzada *leave-one-out* (LOO) foi utilizada para mostrar o desempenho do método proposto em predizer potenciais associações entre doenças e lncRNAs. O processo de aprendizado é executado uma vez para cada associação existente. De forma que, em cada etapa da validação LOO, uma associação conhecida de lncRNA-doença da rede tripartida é removida e torna-se amostra de teste, enquanto as demais associações conhecidas são utilizadas na etapa de aprendizado, como ilustra a Figura 30. A pontuação de relevância de uma associação teste entre

[2] Dados disponíveis em: https://github.com/USTC-HIlab/TPGLDA
[3] Disponível em: http://www.ebi.ac.uk/arrayexpress/
[4] Disponível em: http://disease-ontology.org/
[5] Disponível em: http://www.disgenet.org/web/DisGeNET/
[6] Disponível em: http://www.nlm.nih.gov/

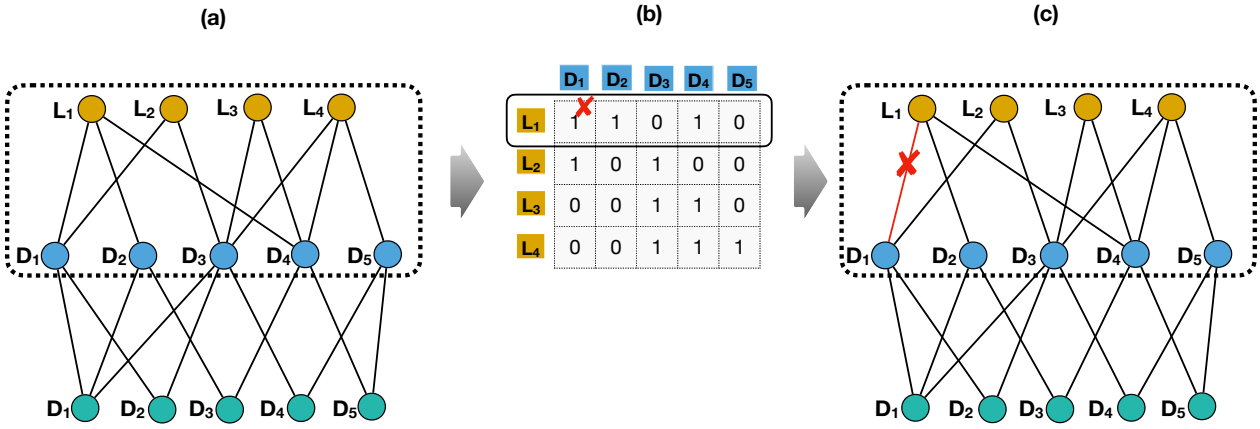
Tabela 11 – Estatísticas básicas da base LncRNADisease. $\langle k_d \rangle$, $\langle k_l \rangle$ e $\langle k_g \rangle$ correspondem, respectivamente, aos graus médios dos vértices dos tipos doença, lncRNA e gene. $kmax_d$, $kmax_l$ e $kmax_g$ representam, respectivamente, os graus máximos dos vértices dos tipos doença, lncRNA e gene.

Associações	$\langle k_d \rangle$	$\langle k_l \rangle$	$\langle k_g \rangle$	$kmax_d$	$kmax_l$	$kmax_g$	Total
doença-lncRNA	3.03	4.69	-	24	54	-	540
doença-gene	29.28	-	3.68	641	-	46	5212

Fonte: Elaborada pelo autor.

um lncRNA e uma doença é calculada em cada etapa da validação e quanto maior a pontuação mais provável é a existência da associação.

Figura 30 – Exemplo de validação cruzada *leave-one-out* utilizada para avaliar a predição de associação entre lncRNA e doença. O X em vermelho representa a remoção de uma associação para a validação.



Fonte: Elaborada pelo autor.

A curva de características operacionais do receptor (*Receiver Operating Characteristic Curve* (ROC)) foi utilizada para apresentar o desempenho preditivo do método proposto e dos métodos da literatura utilizados como referência para a comparação dos resultados obtidos. A curva traça a representação da relação entre a taxa de falsos positivos (1-especificidade) e a taxa de verdadeiros positivos (sensibilidade) ao longo de diferentes limiares ou valores de corte. Com base na curva ROC também foi calculada a área correspondente sob a curva (*Area Under the ROC Curve* (AUC)), que pode possuir valores que variam entre 0 (quando as predições estão 100% erradas) e 1 (quando as predições estão 100% corretas), em que $0,5$ indica desempenho aleatório. A especificidade (ESPEC) corresponde a taxa de associações negativas (não comprovadas) que foram detectados corretamente e a sensibilidade (SENS) representa o percentual de associações existentes que foram corretamente identificados, definidas da seguinte forma:

$$ESPEC = \frac{TN}{TN+FP} = \frac{TN}{N}, \tag{4.2}$$

$$SENS = \frac{TP}{TP+FN} = \frac{TP}{P}. \tag{4.3}$$

Além disso, a partir destas taxas de predição foram calculadas outras medidas de avaliação como acurácia (ACC), precisão (PREC) e coeficiente de correlação de *Matthews* (MCC), definidas da seguinte forma:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{P+N}, \tag{4.4}$$

$$PREC = \frac{TP}{TP+FP}, \tag{4.5}$$

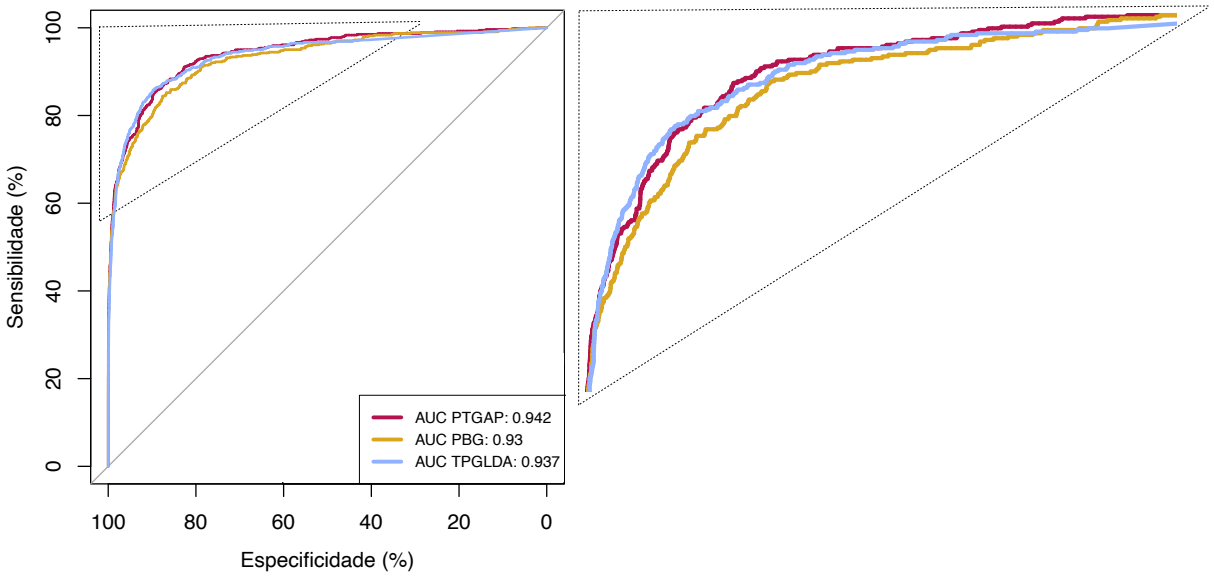
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}. \tag{4.6}$$

A avaliação experimental da capacidade preditiva de associações de lncRNA-doença do PTGAP visa mostrar que o seu desempenho supera o do método PBG que utiliza apenas a rede bipartida e se compara aos desempenhos dos métodos da literatura TPGLDA (DING *et al.*, 2018), KRWRH (GANEGODA *et al.*, 2015) e LRLSLDA (CHEN; YAN, 2013). Este trabalho utilizou os resultados dos métodos KRWRH e LRLSLDA disponíveis no artigo do TPGLDA (DING *et al.*, 2018) e o código disponível deste método para a reprodução dos seus resultados. Isto foi possível devido à utilização do mesmo processo de validação. Nesta análise, os parâmetros do método PBG foram definidos como: quantidade de propagação local e global igual a 5, tamanho dos vetores $k = 20$, $\alpha = 0.005$ e $\beta = 0.1$. Para o caso da técnica TPGLDA, utilizou-se o processo iterativo com número máximo de repetições igual 2.

As curvas ROC correspondentes aos métodos PTGAP, PBG e TPGLDA são mostradas na Figura 31. Pode-se observar que, todos alcançam um desempenho confiável com valores de AUC de 0.942, 0.93 e 0.937 para os métodos PTGAP, PBG e TPGLDA, respectivamente. Apesar de possuírem desempenhos similares, o método proposto PTGAP alcança melhor AUC do que o PBG. Essa melhoria deve-se a utilização da transferência de informações entre diferentes camadas bipartidas da estrutura *k*-partida lncRNA-doença-gene proposta pelo método desenvolvido PTGAP, enquanto o PBG usa apenas informação biparta da rede lncRNA-doença. Essa diferença pode ser significativa principalmente em aplicações biológicas e médicas.

Na Tabela 12 são mostrados os resultados das demais medidas de avaliação usadas para medir o desempenho preditivo dos métodos, Sens, Espec, Prec, Acc e MCC. Nessa análise, dois níveis de rigor para medir o desempenho preditivo foram adotados, como utilizado em outros trabalhos da literatura, fixando a especificidade em 95% e 99% para os cálculos das medidas (DING *et al.*, 2018; SUN *et al.*, 2016; LI *et al.*, 2015). Quando o nível de especificidade atinge um valor igual a 99%, o método PTGAP apresenta desempenho melhor que o PBG e TPGLDA, que possuem maiores valores de AUC que os demais métodos da literatura KRWRH e LRLSLDA.

Figura 31 – Curva ROC e os valores de AUC obtidos para os métodos PTGAP, PBG e TPGLDA na validação cruzada LOO no conjunto de dados de referência.



Fonte: Elaborada pelo autor.

Por outro lado, no caso em que o nível de rigor da especificidade diminui para o ponto de 95%, o desempenho do método proposto se torna ligeiramente menor do que o TPGLDA, mas ainda maior que o PBG.

Tabela 12 – Comparação da abordagem bipartida e tripartida com outros três métodos da literatura considerando dois níveis de especificidade como corte: $Espec = 99\%$ e $Espec = 95\%$.

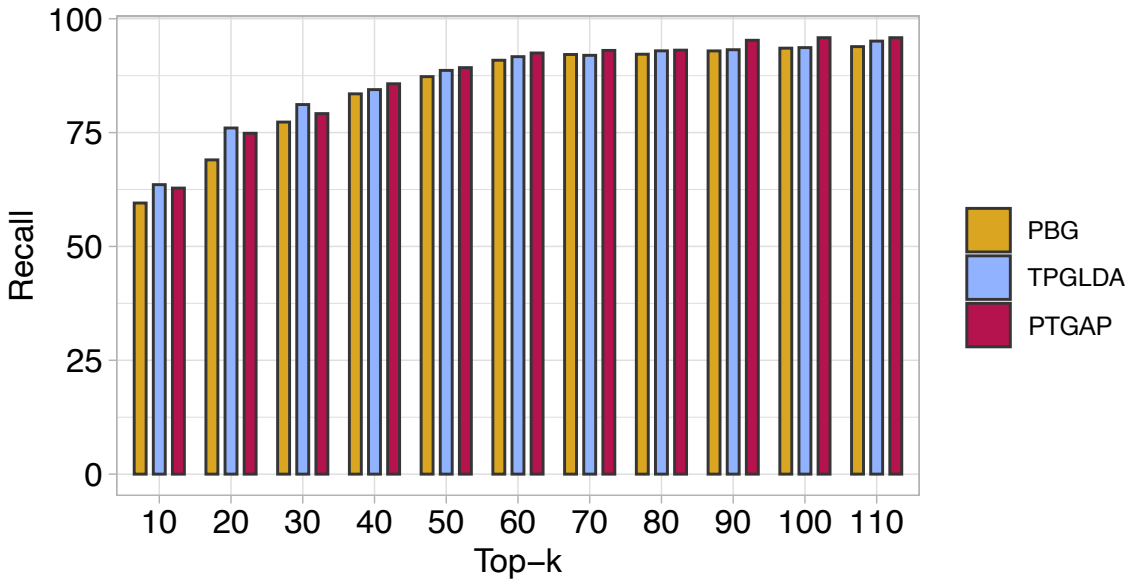
Medidas	**PTGAP**	**PBG**	**TPGLDA**	**KRWRH**	**LRLSLDA**
Espec = 95%					
Sens	74.8%	72%	76.9%	42.6%	35.2%
Acc	94.5%	94.4%	94.5%	93.6%	93.4%
Prec	28.8%	28.1%	29.4%	18.7%	16.0%
MCC	44.3%	42.7%	45.5%	25.4%	20.7%
Espec = 99%					
Sens	56.7%	55.4%	53.5%	11.7%	10.7%
Acc	97.9%	97.8%	97.8%	96.7%	96.7%
Prec	60.6%	60%	59.2%	24.1%	22.6%
MCC	57.5%	56.6%	55.1%	15.2%	14.0%

Fonte: Elaborada pelo autor.

O desempenho dos métodos também foi analisado em relação às k pontuações de associações mais altas. Para tanto, é calculado o *recall* dos métodos, que avalia a quantidade de lncRNAs que realmente possuem associações com as doenças, presente entre as k maiores pontuações geradas pelos métodos. Os resultados obtidos de *recall* para onze diferentes valores de top-k são apresentados na Figura 32. Os métodos PTGAP e TPGLDA possuem os melhores

desempenhos de *recall*. O TPGLDA alcança os maiores valores de *recall* para os ranqueamento com as 10, 20 e 30 maiores pontuações. Porém, quando o tamanho de ranqueamento atinge o valor 40 o método proposto PTGAP demonstra melhor performance. Esta análise demonstra que o PTGAP pode inferir uma grande quantidade de lncRNAs realmente relacionados as doenças em diferentes tamanhos de ranqueamento.

Figura 32 – Média dos valores de *recall* de todas as doenças de acordo com onze diferentes valores de top-*k*.



Fonte: Elaborada pelo autor.

Por fim, a análise da capacidade preditiva dos métodos foi realizada para diferentes doenças humanas, por meio do cálculo da AUC para as predições feitas para cada doença separadamente. Na Tabela 13 são mostrados os valores de AUC para 13 doenças obtidos pelos métodos PTGAP, PBG e TPGLDA. De modo geral, o método proposto PTGAP obtém valores de AUC superiores aos demais métodos. Além disso, um ponto importante para observar é que o PTGAP possui bons resultados independente da quantidade de conexões que as doenças têm com lncRNAs, por exemplo, tanto para o câncer de próstata (13 conexões) quanto para o câncer de cólon (6 conexões) o método atinge valores de AUC de no mínimo 0.9. Assim, pode-se assumir que os resultados do PTGAP fornece bons indícios de sua eficácia para identificar associações entre lncRNAs e uma determinada doença.

Tabela 13 – Resultados de AUC para os métodos PTGAP, PBG e TPGLDA para 13 doenças.

Doença	**N° de associações com LncRNAs**	**AUC**		
		PTGAP	**PBG**	**TPGLDA**
Câncer de Estômago	24	0.837	0.811	0.893
Esophageal Squamous Cell Carcinoma	13	0.841	0.836	0.822
Câncer de Próstata	13	0.90	0.863	0.886
Câncer de bexiga	11	0.869	0.79	0.883
Câncer de Pulmão	9	0.864	0.854	0.828
Melanoma	9	0.985	0.958	0.939
Doença de Huntington	6	0.888	0.68	0.7
Câncer de colon	6	0.959	0.942	0.948
Linfoma	6	0.644	0.657	0.592
Diabetes Tipo 2	4	0.734	0.741	0.555
Câncer de Pâncreas	3	0.973	0.973	0.920
Diabetes Tipo 1	3	0.824	0.857	0.765
Adenocarcinoma de Esôfago	2	0.712	0.664	0.69

Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÃO

Nesta tese foi investigado o desenvolvimento de abordagens de propagação em redes *k*-partidas através da transferência de informação entre diferentes camadas das redes, baseando-se na premissa que diferentes visões de dados tendem a contribuir entre si. Tal comportamento é o fundamento do método semissupervisionado Co-training, que permite uma nova visão de como diferentes informações de uma mesma amostra podem contribuir para a inferência de classe. As abordagens desenvolvidas se mostraram promissoras para exploração de redes *k*-partidas para diferentes problemas reais.

Com o rápido crescimento da Internet uma enorme quantidade de dados e recursos são criados a partir dos usuários. Nesse sentido, os sistemas de recomendação surgem como um meio eficaz de filtragem de informações para solucionar o problema de sobrecarga de informações, fornecendo recomendações personalizadas por meio de técnicas de descoberta de conhecimento. Os sistemas colaborativos de marcação são sistemas de informação que permitem aos usuários atribuir livremente *tags* às suas coleções de itens. Esses sistemas fornecem uma rica estrutura de dados, conhecida como folksonomia, que pode ser modelada como um grafo tripartido. A partir desses dados, duas abordagens foram desenvolvidas nesta tese, uma para a recomendação de *tags* e outra para a recomendação de itens. Para a recomendação de *tags* foi desenvolvida a abordagem TRLBG, que explora a integração entre as representações extraídas para os vértices das redes bipartidas usuário-tag e item-tag. Desse modo, foi possível confirmar que as informações latentes obtidas para cada visão dos dados podem contribuir entre si. Para a recomendação de itens, foi desenvolvida a técnica PTGRL que utiliza toda a informação da estrutura tripartida dos dados para a extração de novas representações para os vértices do tipo usuário. Esta baseia-se em um processo iterativo em que as informações aprendidas em cada rede bipartida são propagadas pela estrutura tripartida do grafo. As representações aprendidas são incorporadas no processo de filtragem colaborativa para a recomendação de itens. Nesse caso, experimentos mostram que o método PTGRL melhorou os resultados em relação ao PBG, sugerindo que o processo de transferência de informações entre as camadas bipartidas da

estrutura k-partida proposta enriquece as representações aprendidas para cada vértice da rede, em comparação com as representações aprendidas utilizando apenas uma rede bipartida. De modo geral, os resultados experimentais também indicaram que ambas as abordagens propostas fornecem resultados competitivos em comparação com métodos da literatura.

Uma outra proposta desta tese foi a abordagem PTGAP de propagação em rede tripartida para o aprendizado de representação visando a predição de associação lncRNA-doença. A descoberta de potenciais associações entre lncRNA e doenças contribui para com a pesquisa de doenças humanas complexas que, como consequência, resulta no aumento da possibilidade de prevenção de doenças, na efetiva melhora da qualidade de diagnósticos, prognósticos e terapias especializadas. Portanto, o desenvolvimento de abordagens computacionais que auxiliem esse processo de descoberta é essencial. Na abordagem proposta, o processo de propagação é dado pela transferência de vetores com informações latentes extraídos para cada vértice do tipo doença entre as camadas lncRNa-doença e gene-doença, de modo iterativo. Esse processo permite que o método de propagação utilizado inicie com informações previamente descobertas em outra camada da rede (visão). A abordagem se mostrou competitiva em comparação com métodos da literatura e com o método PBG aplicado apenas na rede bipartida lncRNA-doença. O aumento no desempenho da abordagem proposta em comparação ao PBG mais uma vez indica que a transferência de informações entre diferentes camadas potencializa a representação aprendida.

Como trabalho futuro, deseja-se incluir as relações entre vértices do mesmo tipo, pois em redes k-partidas são permitidas apenas conexões entre conjuntos disjuntos de vértices. Por exemplo, considerando o contexto de predição de associação entre lncRNA e doenças, seria possível incluir relações entre os vértices do tipo doença, agregando mais informações biológicas para o grafo tripartido. Quanto as melhorias das abordagens em redes k-partidas, pode-se destacar que não foi identificado um critério de parada ou para determinar a quantidade máxima de iterações necessárias, sendo necessário definir mais esse parâmetro além dos que já existem no método de propagação utilizado. Além disso, pretende-se investigar a transferência de informação de rótulos entre camadas de rede k-partida em problemas de contexto semissupervisionado. Por fim, as técnicas desenvolvidas fornecem possibilidades de extensões e refinamentos conforme as necessidades das aplicações. Isso se deve pela modelagem em grafo e formulação de processos iterativos simples e intuitivos.

Para a divulgação dos resultados das contribuições principais desta tese, artigos estão em processo de submissão para periódicos. A seguir são apresentados as principais contribuições resultantes deste trabalho de doutorado:

- **Ferramenta BNOC**
  Valejo, A., Góes, F., Romanetto, L. Oliveira, M. C.; Lopes, A. A. A benchmarkingtool for the generation of bipartite network models with overlapping communities. **Knowledge Information Systems** (2019). https://doi.org/10.1007/s10115-019-01411-9

- **Recomendação de *tags* em sistemas colaborativos de marcação**
  GÓES, F. R.; MILIOS E.; LOPES, A. A.. Tag Recommender based on Propagation in Bipartite Graphs. **Journal of intelligent information systems**. (em fase de submissão)

- **Recomendação de itens em sistemas colaborativos de marcação**
  GÓES, F. R.; VALEJO, A.; MILIOS E.; LOPES, A. A.. Tag-aware recommendation based on Representation Learning in Tripartite Graph. **Expert systems with applications**. (em fase de submissão)

- **Predição de associação entre lncRNA e doenças**
  GÓES, F. R.; LOPES, A. A.. PTGAP: LncRNA-disease associations prediction based on Representation Learning in Tripartite Graph. **BMC Bioinformatics**. (em fase de submissão)

Além disso, atividades e colaborações desenvolvidas ao longo deste trabalho de doutorado são descritas abaixo:

- **Estágio no exterior**
  O projeto de pesquisa foi contemplado com uma bolsa de estágio do *Emerging Leaders in the Americas Program* (ELAP) financiado pelo Governo Canadense para um período de estudo de 6 meses na *Dalhousie University* sob a supervisão do Dr. Evangelos Milios. O supervisor de estágio tem uma ampla experiência em Aprendizado de Máquina e grafos, e colaborou com novos *insights* sobre a proposta desta tese. Ademais, as contribuições relacionadas com a aplicação de Sistemas Colaborativos de marcação tiveram origem no período de pesquisa no Canadá, devido à experiência do supervisor com esses dados (LIPCZAK *et al.*, 2009; LIPCZAK; MILIOS, 2010; LIPCZAK; MILIOS, 2011).

- **Curso na área de Bioinformática**
  Participação no *workshop Resources and tools for functional genomics and chemical biology* (2018, Argentina) organizado pelo CABANA, um consórcio para fortalecimento da Bioinformática na América Latina através de um programa de capacitação sustentável.

REFERÊNCIAS

ABNEY, S. **Semisupervised learning for computational linguistics**. [S.l.]: CRC Press, 2007. Citado na página 41.

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE transactions on knowledge and data engineering**, IEEE, v. 17, n. 6, p. 734–749, 2005. Citado na página 64.

AGGARWAL, C. C. **Data mining: the textbook**. [S.l.]: Springer, 2015. Citado nas páginas 43 e 45.

AGGARWAL, C. C. *et al.* **Recommender systems**. [S.l.]: Springer, 2016. v. 1. Citado nas páginas 63, 64, 65, 66, 67 e 83.

AGGARWAL, C. C.; ZHAI, C. An introduction to text mining. In: **Mining text data**. [S.l.]: Springer, 2012. p. 1–10. Citado na página 54.

AHMADIAN, S.; AHMADIAN, M.; JALILI, M. A deep learning based trust-and tag-aware recommender system. **Neurocomputing**, Elsevier, v. 488, p. 557–571, 2022. Citado na página 85.

AHN, Y.-Y.; BAGROW, J. P.; LEHMANN, S. Link communities reveal multiscale complexity in networks. **nature**, Nature Publishing Group, v. 466, n. 7307, p. 761–764, 2010. Citado na página 39.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citado na página 31.

AMANCIO, D. R.; COMIN, C. H.; CASANOVA, D.; TRAVIESO, G.; BRUNO, O. M.; RODRIGUES, F. A.; COSTA, L. da F. A systematic comparison of supervised classifiers. **PloS one**, Public Library of Science San Francisco, USA, v. 9, n. 4, p. e94137, 2014. Citado na página 42.

AMARAL, L. A. N.; UZZI, B. Complex systems—a new paradigm for the integrative study of management, physical, and technological systems. **Management science**, INFORMS, v. 53, n. 7, p. 1033–1035, 2007. Citado na página 25.

BALCAN, M.-F.; BLUM, A.; YANG, K. Co-training and expansion: Towards bridging theory and practice. **Advances in neural information processing systems**, MIT Press, v. 17, p. 89–96, 2005. Citado na página 46.

BARABÁSI, A.-L. **Network science**. [S.l.]: Cambridge University Press, 2016. Citado na página 32.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado nas páginas 32 e 38.

BELÉM, F. M.; ALMEIDA, J. M.; GONÇALVES, M. A. A survey on tag recommendation methods. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 68, n. 4, p. 830–844, 2017. Citado nas páginas 70, 71 e 72.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural computation**, MIT Press, v. 15, n. 6, p. 1373–1396, 2003. Citado na página 58.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013. Citado nas páginas 26, 27 e 54.

BERTINI, J. R.; ZHAO, L.; MOTTA, R.; LOPES, A. de A. A nonparametric classification method based on k-associated graphs. **Information Sciences**, Elsevier, v. 181, n. 24, p. 5435–5456, 2011. Citado na página 43.

BERTON, L.; FALEIROS, T. de P.; VALEJO, A.; VALVERDE-REBAZA, J.; LOPES, A. de A. Rgcli: Robust graph that considers labeled instances for semi-supervised learning. **Neurocomputing**, Elsevier, v. 226, p. 238–248, 2017. Citado na página 43.

BERTON, L.; LOPES, A. d. A. *et al.* Network construction and applications for semi-supervised learning. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO-SBC. **Conference on Graphics, Patterns and Images, XXIX; Workshop of Theses and Dissertations**. [S.l.], 2016. Citado nas páginas 26 e 43.

BICEGO, M.; LOVATO, P.; OLIBONI, B.; PERINA, A. Expression microarray classification using topic models. In: **Proceedings of the 2010 ACM Symposium on Applied Computing**. [S.l.: s.n.], 2010. p. 1516–1520. Citado na página 28.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012. Citado na página 55.

BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: ACM. **Proceedings of the eleventh annual conference on Computational learning theory**. [S.l.], 1998. p. 92–100. Citado nas páginas 28 e 46.

BOBADILLA, J.; ORTEGA, F.; HERNANDO, A.; GUTIÉRREZ, A. Recommender systems survey. **Knowledge-based systems**, Elsevier, v. 46, p. 109–132, 2013. Citado nas páginas 67 e 89.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4-5, p. 175–308, 2006. Citado nas páginas 25 e 32.

BRENT, R. Genomic biology. **Cell**, Elsevier, v. 100, n. 1, p. 169–183, 2000. Citado na página 101.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998. Citado na página 71.

BYDE, A.; WAN, H.; CAYZER, S. Personalized tag recommendations via tagging and content-based similarity metrics. In: **ICWSM**. [S.l.: s.n.], 2007. Citado na página 72.

CAI, H.; ZHENG, V. W.; CHANG, K. C.-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 30, n. 9, p. 1616–1637, 2018. Citado nas páginas 56, 57 e 58.

CHAKRABORTY, D.; MAULIK, U. Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning. **IEEE journal of translational engineering in health and medicine**, IEEE, v. 2, p. 1–11, 2014. Citado na página 42.

CHANG, S.; HAN, W.; TANG, J.; QI, G.-J.; AGGARWAL, C. C.; HUANG, T. S. Heterogeneous network embedding via deep architectures. In: **Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2015. p. 119–128. Citado na página 59.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. Semi-supervised learning. 2006. **Cambridge, Massachusettes: The MIT Press View Article**, 2006. Citado nas páginas 26 e 44.

CHEN, F.; WANG, Y.-C.; WANG, B.; KUO, C.-C. J. Graph representation learning: A survey. **APSIPA Transactions on Signal and Information Processing**, Cambridge University Press, v. 9, 2020. Citado na página 58.

CHEN, G.; WANG, Z.; WANG, D.; QIU, C.; LIU, M.; CHEN, X.; ZHANG, Q.; YAN, G.; CUI, Q. Lncrnadisease: a database for long-non-coding rna-associated diseases. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D983–D986, 2012. Citado na página 107.

CHEN, H.; ZHANG, Z. Prediction of drug-disease associations for drug repositioning through drug-mirna-disease heterogeneous network. **IEEE Access**, IEEE, v. 6, p. 45281–45287, 2018. Citado na página 27.

CHEN, M.; PENG, Y.; LI, A.; DENG, Y.; LI, Z. A novel lncrna-disease association prediction model using laplacian regularized least squares and space projection-federated method. **IEEE Access**, IEEE, v. 8, p. 111614–111625, 2020. Citado na página 104.

CHEN, R.; HUA, Q.; CHANG, Y.-S.; WANG, B.; ZHANG, L.; KONG, X. A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. **IEEE Access**, IEEE, v. 6, p. 64301–64320, 2018. Citado nas páginas 63, 64, 65, 66 e 67.

CHEN, X.; HE, T.; HU, X.; ZHOU, Y.; AN, Y.; WU, X. Estimating functional groups in human gut microbiome with probabilistic topic models. **IEEE transactions on nanobioscience**, IEEE, v. 11, n. 3, p. 203–215, 2012. Citado na página 28.

CHEN, X.; HU, X.; LIM, T. Y.; SHEN, X.; PARK, E.; ROSEN, G. L. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, v. 9, n. 4, p. 980–991, 2011. Citado na página 28.

CHEN, X.; YAN, G.-Y. Novel human lncrna–disease association inference based on lncrna expression profiles. **Bioinformatics**, Oxford University Press, v. 29, n. 20, p. 2617–2624, 2013. Citado nas páginas 103 e 110.

CINELLI, M.; MORALES, G. D. F.; GALEAZZI, A.; QUATTROCIOCCHI, W.; STARNINI, M. The echo chamber effect on social media. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 118, n. 9, p. e2023301118, 2021. Citado na página 95.

CINUS, F.; MINICI, M.; MONTI, C.; BONCHI, F. The effect of people recommenders on echo chambers and polarization. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2022. v. 16, p. 90–101. Citado na página 95.

COSTA, A. F. D.; MANZATO, M. G.; CAMPELLO, R. J. Boosting collaborative filtering with an ensemble of co-trained recommenders. **Expert Systems with Applications**, Elsevier, v. 115, p. 427–441, 2019. Citado na página 67.

COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. Citado nas páginas 32, 34 e 38.

COSTA, L. da F.; BOAS, P. V.; SILVA, F.; RODRIGUES, F. A pattern recognition approach to complex networks. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2010, n. 11, p. P11015, 2010. Citado nas páginas 25 e 34.

CRAIN, S. P.; ZHOU, K.; YANG, S.-H.; ZHA, H. Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In: **Mining text data**. [S.l.]: Springer, 2012. p. 129–161. Citado na página 55.

CUI, M.; PRASAD, S. Sparsity promoting dimensionality reduction for classification of high dimensional hyperspectral images. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on**. [S.l.], 2013. p. 2154–2158. Citado na página 42.

DENG, L.; YU, D. Deep learning: methods and applications. **Foundations and trends in signal processing**, Now Publishers Inc. Hanover, MA, USA, v. 7, n. 3–4, p. 197–387, 2014. Citado nas páginas 55 e 56.

DING, L.; WANG, M.; SUN, D.; LI, A. Tpglda: Novel prediction of associations between lncrnas and diseases via lncrna-disease-gene tripartite graph. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 1–11, 2018. Citado nas páginas 103, 108 e 110.

DINGER, M. E.; PANG, K. C.; MERCER, T. R.; MATTICK, J. S. Differentiating protein-coding and noncoding rna: challenges and ambiguities. **PLoS computational biology**, Public Library of Science San Francisco, USA, v. 4, n. 11, p. e1000176, 2008. Citado na página 102.

DONG, Y.; CHAWLA, N. V.; SWAMI, A. metapath2vec: Scalable representation learning for heterogeneous networks. In: **Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2017. p. 135–144. Citado na página 59.

DONG, Y.; HU, Z.; WANG, K.; SUN, Y.; TANG, J. Heterogeneous network representation learning. In: **IJCAI**. [S.l.: s.n.], 2020. v. 20, p. 4861–4867. Citado na página 56.

EDDY, S. R. Non–coding rna genes and the modern rna world. **Nature Reviews Genetics**, Nature Publishing Group, v. 2, n. 12, p. 919–929, 2001. Citado nas páginas 101 e 102.

ERDOS, P.; RÉNYI, A. *et al.* On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, Citeseer, v. 5, n. 1, p. 17–60, 1960. Citado nas páginas 31 e 38.

EULER, L. Solutio problematis ad geometriam situs pertinentis. **Commentarii academiae scientiarum Petropolitanae**, p. 128–140, 1741. Citado na página 31.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2011. Citado nas páginas 25 e 40.

FALEIROS, T. d. P. **Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado na página 49.

FALEIROS, T. de P.; ROSSI, R. G.; LOPES, A. de A. Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs. **Pattern Recognition Letters**, Elsevier, v. 87, p. 127–138, 2017. Citado nas páginas 27, 28, 49 e 50.

FALEIROS, T. de P.; VALEJO, A.; LOPES, A. de A. Unsupervised learning of textual pattern based on propagation in bipartite graph. **Intelligent Data Analysis**, IOS Press, v. 24, n. 3, p. 543–565, 2020. Citado nas páginas 27, 28, 70, 73, 83, 85, 103 e 105.

FANG, Y.; FULLWOOD, M. J. Roles, functions, and mechanisms of long non-coding rnas in cancer. **Genomics, proteomics & bioinformatics**, Elsevier, v. 14, n. 1, p. 42–54, 2016. Citado nas páginas 101 e 102.

FOUMANI, S. N. M.; NICKABADI, A. A probabilistic topic model using deep visual word representation for simultaneous image classification and annotation. **Journal of Visual Communication and Image Representation**, Elsevier, v. 59, p. 195–203, 2019. Citado na página 28.

FRITH, M. C.; PHEASANT, M.; MATTICK, J. S. The amazing complexity of the human transcriptome. **European journal of human genetics: EJHG**, v. 13, n. 8, p. 894–897, 2005. Citado na página 102.

FU, G.; WANG, J.; DOMENICONI, C.; YU, G. Matrix factorization-based data fusion for the prediction of lncrna–disease associations. **Bioinformatics**, Oxford University Press, v. 34, n. 9, p. 1529–1537, 2018. Citado na página 104.

GANEGODA, G. U.; LI, M.; WANG, W.; FENG, Q. Heterogeneous network model to infer human disease-long intergenic non-coding rna associations. **IEEE transactions on nanobioscience**, IEEE, v. 14, n. 2, p. 175–183, 2015. Citado nas páginas 103 e 110.

GARG, N.; WEBER, I. Personalized, interactive tag recommendation for flickr. In: **Proceedings of the 2008 ACM conference on Recommender systems**. [S.l.: s.n.], 2008. p. 67–74. Citado na página 72.

GEMMELL, J.; SHEPITSEN, A.; MOBASHER, B.; BURKE, R. Personalizing navigation in folksonomies using hierarchical tag clustering. In: SPRINGER. **International Conference on Data Warehousing and Knowledge Discovery**. [S.l.], 2008. p. 196–205. Citado na página 84.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado nas páginas 32 e 38.

GOLDER, S. A.; HUBERMAN, B. A. Usage patterns of collaborative tagging systems. **Journal of information science**, Sage Publications Sage CA: Thousand Oaks, CA, v. 32, n. 2, p. 198–208, 2006. Citado na página 68.

GOLDMAN, S.; ZHOU, Y. Enhancing supervised learning with unlabeled data. In: CITESEER. **ICML**. [S.l.], 2000. p. 327–334. Citado na página 46.

GONG, C.; TAO, D.; YANG, J.; FU, K. Signed laplacian embedding for supervised dimension reduction. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2014. v. 28, n. 1. Citado na página 27.

GOYAL, P.; FERRARA, E. Graph embedding techniques, applications, and performance: A survey. **Knowledge-Based Systems**, Elsevier, v. 151, p. 78–94, 2018. Citado na página 56.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. In: **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 855–864. Citado nas páginas 54 e 59.

GUILLORY, A.; BILMES, J. A. Label selection on graphs. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2009. p. 691–699. Citado na página 43.

HAFFARI, G. R.; SARKAR, A. Analysis of semi-supervised learning with the yarowsky algorithm. **arXiv preprint arXiv:1206.5240**, 2012. Citado na página 45.

HAMILTON, W. L.; YING, R.; LESKOVEC, J. Representation learning on graphs: Methods and applications. **arXiv preprint arXiv:1709.05584**, 2017. Citado na página 56.

HAMOUDA, S.; WANAS, N. Put-tag: personalized user-centric tag recommendation for social bookmarking systems. **Social network analysis and mining**, Springer, v. 1, n. 4, p. 377–385, 2011. Citado nas páginas 70 e 72.

HE, X.; LIAO, L.; ZHANG, H.; NIE, L.; HU, X.; CHUA, T.-S. Neural collaborative filtering. In: **Proceedings of the 26th international conference on world wide web**. [S.l.: s.n.], 2017. p. 173–182. Citado na página 67.

HOTHO, A.; JÄSCHKE, R.; SCHMITZ, C.; STUMME, G. Folkrank: A ranking algorithm for folksonomies. 2006. Citado nas páginas 71 e 82.

____. Information retrieval in folksonomies: Search and ranking. In: SPRINGER. **European semantic web conference**. [S.l.], 2006. p. 411–426. Citado nas páginas 69, 71 e 82.

HUANG, M.; ZHUANG, F.; ZHANG, X.; AO, X.; NIU, Z.; ZHANG, M.-L.; HE, Q. Supervised representation learning for multi-label classification. **Machine Learning**, Springer, v. 108, n. 5, p. 747–763, 2019. Citado na página 27.

JÄSCHKE, R.; MARINHO, L.; HOTHO, A.; SCHMIDT-THIEME, L.; STUMME, G. Tag recommendations in social bookmarking systems. **Ai Communications**, IOS Press, v. 21, n. 4, p. 231–247, 2008. Citado nas páginas 69, 76 e 77.

JEBARA, T.; WANG, J.; CHANG, S.-F. Graph construction and b-matching for semi-supervised learning. In: ACM. **Proceedings of the 26th Annual International Conference on Machine Learning**. [S.l.], 2009. p. 441–448. Citado na página 43.

JI, M.; SUN, Y.; DANILEVSKY, M.; HAN, J.; GAO, J. Graph regularized transductive classification on heterogeneous information networks. In: SPRINGER. **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. [S.l.], 2010. p. 570–586. Citado na página 49.

JIANG, R.; CHIAPPA, S.; LATTIMORE, T.; GYÖRGY, A.; KOHLI, P. Degenerate feedback loops in recommender systems. In: **Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society**. [S.l.: s.n.], 2019. p. 383–390. Citado na página 95.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 40.

KALIMERIS, D.; BHAGAT, S.; KALYANARAMAN, S.; WEINSBERG, U. Preference amplification in recommender systems. In: **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**. [S.l.: s.n.], 2021. p. 805–815. Citado na página 95.

KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, Springer, v. 18, n. 1, p. 39–43, 1953. Citado na página 71.

KITSAK, M.; KRIOUKOV, D. Hidden variables in bipartite networks. **Physical Review E**, APS, v. 84, n. 2, p. 026114, 2011. Citado na página 37.

KONG, X.; YU, P. S.; DING, Y.; WILD, D. J. Meta path-based collective classification in heterogeneous information networks. In: **Proceedings of the 21st ACM international conference on Information and knowledge management**. [S.l.: s.n.], 2012. p. 1567–1571. Citado na página 27.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. **Computer**, IEEE, v. 42, n. 8, p. 30–37, 2009. Citado na página 67.

KRESTEL, R.; FANKHAUSER, P.; NEJDL, W. Latent dirichlet allocation for tag recommendation. In: **Proceedings of the third ACM conference on Recommender systems**. [S.l.: s.n.], 2009. p. 61–68. Citado nas páginas 28 e 70.

LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. Benchmark graphs for testing community detection algorithms. **Physical review E**, APS, v. 78, n. 4, p. 046110, 2008. Citado na página 38.

LANCICHINETTI, A.; RADICCHI, F.; RAMASCO, J. J.; FORTUNATO, S. Finding statistically significant communities in networks. **PloS one**, Public Library of Science San Francisco, USA, v. 6, n. 4, p. e18961, 2011. Citado na página 39.

LARGERON, C.; MOUGEL, P.-N.; RABBANY, R.; ZAÏANE, O. R. Generating attributed networks with communities. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 4, p. e0122777, 2015. Citado na página 38.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, v. 401, n. 6755, p. 788–791, 1999. Citado na página 105.

LEI, K.; FU, Q.; YANG, M.; LIANG, Y. Tag recommendation by text classification with attention-based capsule network. **Neurocomputing**, Elsevier, v. 391, p. 65–73, 2020. Citado na página 73.

LI, A.; GE, M.; ZHANG, Y.; PENG, C.; WANG, M. Predicting long noncoding rna and protein interactions using heterogeneous network model. **BioMed research international**, Hindawi, v. 2015, 2015. Citado na página 110.

LI, J.; ZHENG, S.; CHEN, B.; BUTTE, A. J.; SWAMIDASS, S. J.; LU, Z. A survey of current trends in computational drug repositioning. **Briefings in bioinformatics**, Oxford University Press, v. 17, n. 1, p. 2–12, 2016. Citado na página 102.

LIANG, N.; ZHENG, H.-T.; CHEN, J.-Y.; SANGAIAH, A. K.; ZHAO, C.-Z. Trsdl: Tag-aware recommender system based on deep learning–intelligent computing systems. **Applied Sciences**, MDPI, v. 8, n. 5, p. 799, 2018. Citado na página 85.

LING, H.; FABBRI, M.; CALIN, G. A. Micrornas and other non-coding rnas as targets for anticancer drug development. **Nature reviews Drug discovery**, Nature Publishing Group, v. 12, n. 11, p. 847–865, 2013. Citado na página 102.

LIPCZAK, M.; HU, Y.; KOLLET, Y.; MILIOS, E. E. Tag sources for recommendation in collaborative tagging systems. In: CITESEER. **DC@ PKDD/ECML**. [S.l.], 2009. Citado nas páginas 68, 72 e 117.

LIPCZAK, M.; MILIOS, E. Learning in efficient tag recommendation. In: ACM. **Proceedings of the fourth ACM conference on Recommender systems**. [S.l.], 2010. p. 167–174. Citado na página 117.

____. Efficient tag recommendation for real-life data. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 3, n. 1, p. 1–21, 2011. Citado nas páginas 72 e 117.

LIU, Q.; XIE, R.; CHEN, L.; LIU, S.; TU, K.; CUI, P.; ZHANG, B.; LIN, L. Graph neural network for tag ranking in tag-enhanced video recommendation. In: **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**. [S.l.: s.n.], 2020. p. 2613–2620. Citado na página 73.

LOPS, P.; GEMMIS, M. D.; SEMERARO, G. Content-based recommender systems: State of the art and trends. **Recommender systems handbook**, Springer, p. 73–105, 2011. Citado na página 65.

LU, C.; YANG, M.; LUO, F.; WU, F.-X.; LI, M.; PAN, Y.; LI, Y.; WANG, J. Prediction of lncrna–disease associations based on inductive matrix completion. **Bioinformatics**, Oxford University Press, v. 34, n. 19, p. 3357–3364, 2018. Citado na página 104.

LÜ, L.; MEDO, M.; YEUNG, C. H.; ZHANG, Y.-C.; ZHANG, Z.-K.; ZHOU, T. Recommender systems. **Physics reports**, Elsevier, v. 519, n. 1, p. 1–49, 2012. Citado na página 65.

LUO, Y.; ZHAO, X.; ZHOU, J.; YANG, J.; ZHANG, Y.; KUANG, W.; PENG, J.; CHEN, L.; ZENG, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. **Nature communications**, Nature Publishing Group, v. 8, n. 1, p. 1–13, 2017. Citado na página 27.

MARINHO, L. B.; NANOPOULOS, A.; SCHMIDT-THIEME, L.; JÄSCHKE, R.; HOTHO, A.; STUMME, G.; SYMEONIDIS, P. Social tagging recommender systems. In: **Recommender systems handbook**. [S.l.]: Springer, 2011. p. 615–644. Citado na página 68.

MARINHO, L. B.; SCHMIDT-THIEME, L. Collaborative tag recommendations. In: **Data Analysis, Machine Learning and Applications**. [S.l.]: Springer, 2008. p. 533–540. Citado nas páginas 72, 77 e 81.

MATTICK, J. S.; MAKUNIN, I. V. Non-coding rna. **Human molecular genetics**, Oxford University Press, v. 15, n. suppl_1, p. R17–R29, 2006. Citado na página 102.

MENEZES, G. V.; ALMEIDA, J. M.; BELÉM, F.; GONÇALVES, M. A.; LACERDA, A.; MOURA, E. S. d.; PAPPA, G. L.; VELOSO, A.; ZIVIANI, N. Demand-driven tag recommendation. In: SPRINGER. **Joint European conference on machine learning and knowledge discovery in databases**. [S.l.], 2010. p. 402–417. Citado na página 72.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citado na página 59.

MISHNE, G. Autotag: a collaborative approach to automated tag assignment for weblog posts. In: **Proceedings of the 15th international conference on World Wide Web**. [S.l.: s.n.], 2006. p. 953–954. Citado nas páginas 71 e 72.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2. Citado na página 40.

MNIH, A.; SALAKHUTDINOV, R. R. Probabilistic matrix factorization. **Advances in neural information processing systems**, v. 20, 2007. Citado na página 84.

MORI, T.; NGOUV, H.; HAYASHIDA, M.; AKUTSU, T.; NACHER, J. C. ncrna-disease association prediction based on sequence information and tripartite network. **BMC systems biology**, BioMed Central, v. 12, n. 1, p. 41–51, 2018. Citado na página 102.

MOUSSIADES, L.; VAKALI, A. Benchmark graphs for the evaluation of clustering algorithms. In: IEEE. **2009 Third International Conference on Research Challenges in Information Science**. [S.l.], 2009. p. 197–206. Citado na página 38.

NAKAMOTO, R.; NAKAJIMA, S.; MIYAZAKI, J.; UEMURA, S. Tag-based contextual collaborative filtering. **IAENG International Journal of Computer Science**, v. 34, n. 2, 2007. Citado na página 84.

NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford university press, 2010. Citado nas páginas 25, 32 e 38.

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado nas páginas 25, 31 e 38.

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical review E**, APS, v. 69, n. 2, p. 026113, 2004. Citado na página 38.

PAN, Y.; HUO, Y.; TANG, J.; ZENG, Y.; CHEN, B. Exploiting relational tag expansion for dynamic user profile in a tag-aware ranking recommender system. **Information Sciences**, Elsevier, v. 545, p. 448–464, 2021. Citado na página 84.

PARK, M.-H.; HONG, J.-H.; CHO, S.-B. Location-based recommendation system using bayesian user's preference model in mobile devices. In: SPRINGER. **International conference on ubiquitous intelligence and computing**. [S.l.], 2007. p. 1130–1139. Citado na página 67.

PAVLOPOULOS, G. A.; KONTOU, P. I.; PAVLOPOULOU, A.; BOUYIOUKOS, C.; MARKOU, E.; BAGOS, P. G. Bipartite graphs in systems biology and medicine: a survey of methods and applications. **GigaScience**, Oxford University Press, v. 7, n. 4, p. giy014, 2018. Citado nas páginas 27 e 33.

PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado na página 54.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. In: **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2014. p. 701–710. Citado na página 59.

RAMEZANI, M. Improving graph-based approaches for personalized tag recommendation. **Journal of Emerging Technologies in Web Intelligence**, v. 3, n. 2, p. 168–176, 2011. Citado na página 71.

RAWASHDEH, M.; KIM, H.-N.; ALJA'AM, J. M.; SADDIK, A. E. Folksonomy link prediction based on a tripartite graph for tag recommendation. **Journal of Intelligent Information Systems**, Springer, v. 40, n. 2, p. 307–325, 2013. Citado nas páginas 71, 77 e 81.

REES, B. S.; GALLAGHER, K. B. Overlapping community detection using a community optimized graph swarm. **Social Network Analysis and Mining**, Springer, v. 2, n. 4, p. 405–417, 2012. Citado na página 38.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. In: **Recommender systems handbook**. [S.l.]: Springer, 2011. p. 1–35. Citado nas páginas 63, 64, 83, 85 e 87.

ROSSI, R. G.; LOPES, A. de A.; REZENDE, S. O. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. **Information Processing & Management**, Elsevier, v. 52, n. 2, p. 217–257, 2016. Citado nas páginas 27 e 49.

____. Using bipartite heterogeneous networks to speed up inductive semi-supervised learning and improve automatic text categorization. **Knowledge-Based Systems**, Elsevier, 2017. Citado na página 47.

ROSSI, R. G.; REZENDE, S. O.; LOPES, A. de A. Term network approach for transductive classification. In: SPRINGER. **International Conference on Intelligent Text Processing and Computational Linguistics**. [S.l.], 2015. p. 497–515. Citado na página 49.

SABOUR, S.; FROSST, N.; HINTON, G. E. Dynamic routing between capsules. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 73.

SARWAR, B.; KARYPIS, G.; KONSTAN, J.; RIEDL, J. Item-based collaborative filtering recommendation algorithms. In: **Proceedings of the 10th international conference on World Wide Web**. [S.l.: s.n.], 2001. p. 285–295. Citado na página 67.

SHAHREZA, M. L.; GHADIRI, N.; MOUSAVI, S. R.; VARSHOSAZ, J.; GREEN, J. R. Heter-lp: A heterogeneous label propagation algorithm and its application in drug repositioning. **Journal of biomedical informatics**, Elsevier, v. 68, p. 167–183, 2017. Citado nas páginas 27 e 36.

SHANG, M.-S.; ZHANG, Z.-K.; ZHOU, T.; ZHANG, Y.-C. Collaborative filtering with diffusion-based similarity on tripartite graphs. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 6, p. 1259–1264, 2010. Citado na página 84.

SHENDURE, J.; BALASUBRAMANIAN, S.; CHURCH, G. M.; GILBERT, W.; ROGERS, J.; SCHLOSS, J. A.; WATERSTON, R. H. Dna sequencing at 40: past, present and future. **Nature**, Nature Publishing Group, v. 550, n. 7676, p. 345–353, 2017. Citado na página 102.

SHEPITSEN, A.; GEMMELL, J.; MOBASHER, B.; BURKE, R. Personalized recommendation in social tagging systems using hierarchical clustering. In: **Proceedings of the 2008 ACM conference on Recommender systems**. [S.l.: s.n.], 2008. p. 259–266. Citado na página 84.

SHI, C.; LI, Y.; ZHANG, J.; SUN, Y.; PHILIP, S. Y. A survey of heterogeneous information network analysis. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 29, n. 1, p. 17–37, 2017. Citado nas páginas 25, 26, 27, 28 e 35.

SIGURBJÖRNSSON, B.; ZWOL, R. V. Flickr tag recommendation based on collective knowledge. In: **Proceedings of the 17th international conference on World Wide Web**. [S.l.: s.n.], 2008. p. 327–336. Citado na página 72.

SILVA, T. C.; ZHAO, L. **Machine learning in complex networks**. [S.l.]: Springer, 2016. v. 1. Citado nas páginas 26, 40, 41, 44, 45, 47 e 48.

SUN, D.; LI, A.; FENG, H.; WANG, M. Ntsmda: prediction of mirna–disease associations by integrating network topological similarity. **Molecular biosystems**, Royal Society of Chemistry, v. 12, n. 7, p. 2224–2232, 2016. Citado na página 110.

SUN, J.; SHI, H.; WANG, Z.; ZHANG, C.; LIU, L.; WANG, L.; HE, W.; HAO, D.; LIU, S.; ZHOU, M. Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. **Molecular BioSystems**, Royal Society of Chemistry, v. 10, n. 8, p. 2074–2081, 2014. Citado na página 104.

SUN, J.; ZHU, M.; JIANG, Y.; LIU, Y.; WU, L. Hierarchical attention model for personalized tag recommendation. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 72, n. 2, p. 173–189, 2021. Citado na página 73.

SUN, K.; WANG, L.; XU, B.; ZHAO, W.; TENG, S. W.; XIA, F. Network representation learning: From traditional feature learning to deep learning. **IEEE Access**, IEEE, v. 8, p. 205600–205617, 2020. Citado nas páginas 27, 54, 56, 58 e 59.

SUN, Y.; HAN, J. Mining heterogeneous information networks: a structural analysis approach. **Acm Sigkdd Explorations Newsletter**, ACM New York, NY, USA, v. 14, n. 2, p. 20–28, 2013. Citado nas páginas 27 e 35.

TANG, S.; YAO, Y.; ZHANG, S.; XU, F.; GU, T.; TONG, H.; YAN, X.; LU, J. An integral tag recommendation model for textual content. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, n. 01, p. 5109–5116. Citado nas páginas 71 e 73.

TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. Citado na página 58.

TOMMASEL, A.; RODRIGUEZ, J. M.; GODOY, D. I want to break free! recommending friends from outside the echo chamber. In: **Fifteenth ACM Conference on Recommender Systems**. [S.l.: s.n.], 2021. p. 23–33. Citado na página 95.

TRIGUERO, I.; GARCÍA, S.; HERRERA, F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. **Knowledge and Information systems**, Springer, v. 42, n. 2, p. 245–284, 2015. Citado na página 45.

TSO-SUTTER, K. H.; MARINHO, L. B.; SCHMIDT-THIEME, L. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In: **Proceedings of the 2008 ACM symposium on Applied computing**. [S.l.: s.n.], 2008. p. 1995–1999. Citado nas páginas 83 e 84.

UYSAL, A. K.; GUNAL, S. A novel probabilistic feature selection method for text classification. **Knowledge-Based Systems**, Elsevier, v. 36, p. 226–235, 2012. Citado na página 42.

VALEJO, A.; GÓES, F.; ROMANETTO, L.; OLIVEIRA, M. C. Ferreira de; LOPES, A. de A. A benchmarking tool for the generation of bipartite network models with overlapping communities. **Knowledge and Information Systems**, Springer, v. 62, n. 4, p. 1641–1669, 2019. Citado nas páginas 30, 38, 39, 40 e 95.

VAPNIK, V. N. **Statistical Learning Theory**. [S.l.]: Wiley-Interscience, 1998. Citado na página 41.

WANG, D.; CUI, P.; ZHU, W. Structural deep network embedding. In: **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1225–1234. Citado na página 27.

WANG, D.; WANG, J.; LU, M.; SONG, F.; CUI, Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. **Bioinformatics**, Oxford University Press, v. 26, n. 13, p. 1644–1650, 2010. Citado na página 108.

WANG, W.; ZHOU, Z.-H. Analyzing co-training style algorithms. In: SPRINGER. **European conference on machine learning**. [S.l.], 2007. p. 454–465. Citado na página 46.

WAPINSKI, O.; CHANG, H. Y. Long noncoding rnas and human disease. **Trends in cell biology**, Elsevier, v. 21, n. 6, p. 354–361, 2011. Citado na página 102.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado nas páginas 32 e 38.

WETZKER, R.; ZIMMERMANN, C.; BAUCKHAGE, C.; ALBAYRAK, S. I tag, you tag: translating tags for advanced user models. In: **Proceedings of the third ACM international conference on Web search and data mining**. [S.l.: s.n.], 2010. p. 71–80. Citado na página 81.

WU, L.; CHEN, E.; LIU, Q.; XU, L.; BAO, T.; ZHANG, L. Leveraging tagging for neighborhood-aware probabilistic matrix factorization. In: **Proceedings of the 21st ACM international conference on Information and knowledge management**. [S.l.: s.n.], 2012. p. 1854–1858. Citado na página 84.

WU, L.; YANG, L.; YU, N.; HUA, X.-S. Learning to tag. In: **Proceedings of the 18th international conference on World wide web**. [S.l.: s.n.], 2009. p. 361–370. Citado na página 72.

WU, P.; ZHANG, Z.-K. Enhancing personalized recommendations on weighted social tagging networks. **Physics Procedia**, Elsevier, v. 3, n. 5, p. 1877–1885, 2010. Citado na página 84.

XU, Z.; CHEN, C.; LUKASIEWICZ, T.; MIAO, Y.; MENG, X. Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling. In: **Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**. [S.l.: s.n.], 2016. p. 1921–1924. Citado na página 85.

XU, Z.; LUKASIEWICZ, T.; CHEN, C.; MIAO, Y.; MENG, X. Tag-aware personalized recommendation using a hybrid deep model. In: AAAI PRESS/INTERNATIONAL JOINT CONFERENCES ON ARTIFICIAL INTELLIGENCE. [S.l.], 2017. Citado na página 85.

XUAN, Z.; LI, J.; YU, J.; FENG, X.; ZHAO, B.; WANG, L. A probabilistic matrix factorization method for identifying lncrna-disease associations. **Genes**, MDPI, v. 10, n. 2, p. 126, 2019. Citado na página 104.

YALAMANCHILI, H. B.; KHO, S. J.; RAYMER, M. L. Latent dirichlet allocation for classification using gene expression data. In: IEEE. **2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE)**. [S.l.], 2017. p. 39–44. Citado na página 28.

YAN, C.; ZHANG, Z.; BAO, S.; HOU, P.; ZHOU, M.; XU, C.; SUN, J. Computational methods and applications for identifying disease-associated lncrnas as potential biomarkers and therapeutic targets. **Molecular Therapy-Nucleic Acids**, Elsevier, v. 21, p. 156–171, 2020. Citado na página 102.

YANG, C.; XIAO, Y.; ZHANG, Y.; SUN, Y.; HAN, J. Heterogeneous network representation learning: A unified framework with survey and benchmark. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, 2020. Citado na página 56.

YANG, X.; GAO, L.; GUO, X.; SHI, X.; WU, H.; SONG, F.; WANG, B. A network based method for analysis of lncrna-disease associations and prediction of lncrnas implicated in diseases. **PloS one**, Public Library of Science San Francisco, USA, v. 9, n. 1, p. e87797, 2014. Citado nas páginas 102 e 104.

YANG, X.; GUO, Y.; LIU, Y.; STECK, H. A survey of collaborative filtering based social recommender systems. **Computer communications**, Elsevier, v. 41, p. 1–10, 2014. Citado na página 67.

YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. In: **33rd annual meeting of the association for computational linguistics**. [S.l.: s.n.], 1995. p. 189–196. Citado na página 45.

YIN, Z.; LI, R.; MEI, Q.; HAN, J. Exploring social tagging graph for web object classification. In: ACM. **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2009. p. 957–966. Citado na página 49.

YU, G.; WANG, Y.; WANG, J.; FU, G.; GUO, M.; DOMENICONI, C. Weighted matrix factorization based data fusion for predicting lncrna-disease associations. In: IEEE. **2018 IEEE international conference on bioinformatics and biomedicine (BIBM)**. [S.l.], 2018. p. 572–577. Citado na página 104.

YUAN, L.; ZHAO, J.; SUN, T.; SHEN, Z. A machine learning framework that integrates multi-omics data predicts cancer-related lncrnas. **BMC bioinformatics**, BioMed Central, v. 22, n. 1, p. 1–18, 2021. Citado na página 102.

ZANIN, M.; PAPO, D.; SOUSA, P. A.; MENASALVAS, E.; NICCHI, A.; KUBIK, E.; BOCCALETTI, S. Combining complex networks and data mining: why and how. **Physics Reports**, Elsevier, v. 635, p. 1–44, 2016. Citado nas páginas 25 e 26.

ZHANG, D.; YIN, J.; ZHU, X.; ZHANG, C. Network representation learning: A survey. **IEEE transactions on Big Data**, IEEE, v. 6, n. 1, p. 3–28, 2018. Citado nas páginas 27 e 58.

ZHANG, J.; ZHANG, Z.; CHEN, Z.; DENG, L. Integrating multiple heterogeneous networks for novel lncrna-disease association inference. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, v. 16, n. 2, p. 396–406, 2017. Citado na página 104.

ZHANG, Y.; CHEN, M.; LI, A.; CHENG, X.; JIN, H.; LIU, Y. Ldai-isps: Lncrna–disease associations inference based on integrated space projection scores. **International journal of molecular sciences**, MDPI, v. 21, n. 4, p. 1508, 2020. Citado na página 104.

ZHANG, Z.-K.; ZHOU, T.; ZHANG, Y.-C. Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 1, p. 179–186, 2010. Citado nas páginas 83 e 84.

_____. Tag-aware recommender systems: a state-of-the-art survey. **Journal of computer science and technology**, Springer, v. 26, n. 5, p. 767–777, 2011. Citado nas páginas 65, 68 e 85.

ZHEN, Y.; LI, W.-J.; YEUNG, D.-Y. Tagicofi: tag informed collaborative filtering. In: **Proceedings of the third ACM conference on Recommender systems**. [S.l.: s.n.], 2009. p. 69–76. Citado na página 84.

ZHOU, D.; BOUSQUET, O.; LAL, T. N.; WESTON, J.; SCHÖLKOPF, B. Learning with local and global consistency. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2004. p. 321–328. Citado nas páginas 48 e 104.

ZHOU, T.; REN, J.; MEDO, M.; ZHANG, Y.-C. Bipartite network projection and personal recommendation. **Physical review E**, APS, v. 76, n. 4, p. 046115, 2007. Citado nas páginas 84, 89 e 104.

ZHOU, Z.-H.; LI, M. Tri-training: Exploiting unlabeled data using three classifiers. **IEEE Transactions on knowledge and Data Engineering**, IEEE, v. 17, n. 11, p. 1529–1541, 2005. Citado na página 46.

ZHU, X. **Semi-Supervised Learning with Graphs**. Tese (Doutorado) — Carnegie Mellon University, 5 2005. Citado nas páginas 43, 47 e 48.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In: **Proceedings of the 20th International conference on Machine learning (ICML-03)**. [S.l.: s.n.], 2003. p. 912–919. Citado na página 49.

ZUO, Y.; ZENG, J.; GONG, M.; JIAO, L. Tag-aware recommender systems based on deep neural networks. **Neurocomputing**, Elsevier, v. 204, p. 51–60, 2016. Citado nas páginas 77, 83, 84 e 89.


ICMC
USP
SÃO CARLOS