UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Classificação transdutiva em redes heterogêneas de informação, baseada na divergência KL.

**Luzia de Menezes Romanetto**

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

ICMC USP
SÃO CARLOS

**Luzia de Menezes Romanetto**

# Classificação transdutiva em redes heterogêneas de informação, baseada na divergência KL.

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Alneu de Andrade Lopes

**USP – São Carlos**
**Novembro de 2019**

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

R758c

Romanetto, Luzia de Menezes
Classificação transdutiva em redes heterogêneas de informação, baseada na divergência KL / Luzia de Menezes Romanetto; orientador Alneu de Andrade Lopes. -- São Carlos, 2019.
106 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, 2019.

1. Redes Heterogêneas de Informação. 2. Classificação Transdutiva. 3. Divergência KL . I. Lopes, Alneu de Andrade, 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

**Luzia de Menezes Romanetto**

# Transductive classification in heterogeneous information networks based on KL-divergence.

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

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Alneu de Andrade Lopes

**USP – São Carlos**
**November 2019**

*Este trabalho é dedicado a todos que de alguma forma passaram pela minha vida ao longo deste percurso. Que para o bem ou para o mal me fizeram aprender algo novo, pois cada nova lição me fez quem sou hoje.*

# AGRADECIMENTOS

À minha família, com quem pude estar junto em todos os momentos.

Ao Tor que fez minha vida feliz por lindos 6 anos, um ser com quem aprendi o que é o amor puro. À Mantega que me acompanha até hoje e me rouba toda atenção quando estou perdida na escuridão.

Aos meus amigos que perto ou distantes torceram por mim.

Ao Kim, que foi importante em muitos momentos nesta jornada.

Ao Google que me deu a incrível oportunidade de estágio. Uma experiência incrível que me fez aprender muito, tanto na área técnica quanto pessoal. Ao meu time que foi incrível em me guiar e apoiar em todos os momentos.

Ao meu orientador prof. Dr. Alneu de Andrade Lopes por me orientar, apoiar e compreender em todos os momentos. Muito além do obvio desta frase, não teria conseguido terminar essa tese sem seu apoio.

*“Apreciar e compreender a vida em cada instante é uma arte a ser praticada.”*
*(Monja Coen)*

# RESUMO

ROMANETTO, L. M. **Classificação transdutiva em redes heterogêneas de informação, baseada na divergência KL.**. 2019. 106 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, 2019.

A área de esquisa em Redes Heterogêneas de Informação (HIN) é um recente e proeminente tópico, especialmente quando consideramos que grande parte dos dados de mundo real possuem características heterogêneas. Tais dados, com topologias complexas como relações entre diferentes tipos de objetos, o que não é naturalmente representável pelas tradicionais redes homogêneas. Além disso, comparada com as pesquisas existentes em redes homogêneas ou mesmo em redes bipartidas, a área de pesquisa em HIN ainda permanece com diversos pontos inexplorados. Dentre estes, o desenvolvimento de métodos para a classificação transdutiva em HIN apresenta diversas possibilidades de desenvolvimento. Nesta tese foi proposto o método TCHN de classificação transdutiva de HIN. Tal método tem como diferencial a utilização da divergência KL como medida de similaridade para a regularização da propagação de informação pelos vetores de informação. Esta modelagem tem como motivação o fato de tal métrica ser mais apropriada para a regularização de distribuições de probabilidade, considerando que a distribuição de informação na rede tende a se comporta de tal maneira. Experimentos comprovam que o método TCHN produz resultados comparáveis ou até mesmo superiores aos métodos representativos da área, confirmando assim sua efetividade para a classificação em diversos cenários. Além disso, a complexidade do método TCHN para redes esparsas mostra-se bastante atrativa para a aplicação em dados de mundo real, que como já comentado possuem naturalmente características heterogêneas. Além do desenvolvimento do método TCHN, como parte das demandas da área que impactaram neste trabalho, foi desenvolvida uma ferramenta de geração de redes heterogêneas sintéticas, camada HNOC, em parceria com outros pesquisadores do grupo de pesquisa. Esta já se mostrou bastante útil para a validação do método TCHN, pois com seu uso, foi possível a comparação das técnicas em redes com diferentes características com um custo bastante reduzido se comparado com o possível custo de levantamento de redes semelhantes com base em dados reais.

**Palavras-chave:** Redes Heterogêneas de Informação, Classificação Transdutiva, Divergência KL.

# ABSTRACT

ROMANETTO, L. M. **Transductive classification in heterogeneous information networks based on KL-divergence.**. 2019. 106 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, 2019.

Heterogeneous Information Networks (HIN) is a promising and recent research topic, specially considering that much real world data are heterogeneous. Those data, with complex topology such as relation among different types of objects, are not naturally represented by usual homogeneous networks. Moreover, compared to existing research on homogeneous networks, or even on bipartite networks, HIN research field still presents many unexplored points. Among these, the development of methods for transductive classification in HIN presents several development possibilities. In this thesis we propose a new transductive classification method on HIN called TCHN. This method has as a differential the use of KL divergence as a similarity measure to regularize the information propagation through information vectors. This modeling is motivated by the fact that such metric is more appropriate for the regularization of probability distributions, considering that the information distribution in the network tends to behave in such a way. Experiments show the TCHN method produces results comparable or even superior to representative methods of the area, thus confirming its effectiveness for classification in different scenarios. Moreover, the complexity of the TCHN method for sparse networks is attractive for application to real world data, which as already discussed naturally have heterogeneous characteristics. In addition to the development of the TCHN method, as part of the demands of the area that impacted this work, it was developed a tool for synthetic heterogeneous network generation, this development was made in partnership with other researchers of our group. HNOC has already proved to be very useful in the validation of the TCHN method, with its use it was possible to compare the techniques in networks with different characteristics at a very low cost compared to the possible cost of surveying similar networks based on real data.

**Keywords:** Heterogeneous Information Networks, KL-divergence, Transductive Classification..

# LISTA DE ILUSTRAÇÕES

# LISTA DE ALGORITMOS

# LISTA DE TABELAS

# SUMÁRIO

CAPÍTULO

1

# INTRODUÇÃO

Grande parte dos sistemas de mundo real possuem como parte de sua estrutura componentes ou entidades multi-tipos, ou mesmo interações internas ou externas com tal característica. Podemos citar como exemplos de tais sistemas : atividades sociais humanas, sistemas comunicação e sistemas biológicos. Sendo muitos dos dados gerados por tais sistemas representados via redes complexas. Tais redes, ou grafos, com topologias não triviais, em muitos casos, são assim naturalmente heterogêneas, compostas por elementos multi-tipos. Em tais casos, a representação tradicional via redes homogênea (compostas apenas por um tipo de vértices e arestas) não é capaz de considerar toda a informação presente nos dados originais. De forma que, a representação de tais dados via redes heterogêneas de informação (HIN) [1] se mostra uma melhor opção, capaz de contemplar a complexidade do dado e levar a melhores resultados em diversas tarefas computacionais.

A área de pesquisa em HIN introduzida por Sun *et al.* (2009) teve muito de seu progresso contido ao longo da ultima década, o que faz de tal área um campo de pesquisa bastante recente. Desde então, tal área vem recebendo um crescente interesse e vem ganhando a atenção de muitos pesquisadores em diversas áreas de aprendizado de maquina, tais como : predição de *links* (SHAHREZA *et al.*, 2017; SHI *et al.*, 2012), agrupamento (PIO *et al.*, 2018) e classificação (SUN; HAN, 2012; FALEIROS; ROSSI; LOPES, 2017). Neste trabalho, o principal foco de interesse é a classificação de dados dentro do contexto de redes heterogêneas.

Em geral, quando tratamos dados reais possuímos informações de classe de apenas um subconjunto dos dados, e queremos conhecer e extrair toda ou ao menos ter uma visão sobre a informação contida neles. Ao mesmo tempo, a classificação exata de todo o conjunto é impraticável, em especial em grades volumes devido ao custo de mão de obra e tempo. Assim, os métodos de classificação computacional se tornam ferramentas úteis na tarefa de estimar a informação desejada para os dados desconhecidos, e extrair uma melhor visão da informação

[1] Acrônimo do inglês para *Heterogeneous Information Network*

contida nestes.

A classificação transdutiva tem como objetivo resolver a tarefa de classificar um conjunto de dados, com apenas um subconjunto previamente rotulado, fazendo isso sem gerar um modelo (ou hipótese) que generaliza a classificação para novas amostras. Em redes heterogêneas, os métodos de classificação transdutiva se aproveitam da estrutura e significado da rede para propagar as informações, dos elementos conhecidos para os elementos desconhecidos com base em algumas pressuposições sobre a classificação dos elementos e suas relações dentro da rede.

A formulação dos métodos de classificação transdutivos partem de duas importantes premissas (ZHOU *et al.*, 2004):

- **Consistência local ou restrição de suavidade:** pontos próximos e/ou relacionados tendem a ter um mesmo rótulo.
- **Consistência com rótulos pré existente:** bons métodos não deverem divergir muito dos rótulos conhecidos *a priori*.

Partindo de tais premissas, o modelagem da função de custo é equacionada com base em dois termos, um para cada pressuposição e, em geral, o problema é solucionado via métodos iterativos.

Nos últimos anos diversas técnicas vem sendo desenvolvidas com tal embasamento voltado para HIN, mostrando resultados superiores aos de técnicas para rede homogêneas em problemas como: classificação em redes bibliográficas (HWANG; KUANG, 2010), classificação de texto (ROSSI; LOPES; REZENDE, 2016) e reposicionamento de fármacos (LUO *et al.*, 2017).

Apesar disso, comparando com a área de pesquisa em redes homogêneas, ou mesmo em redes bipartidas que é um caso articular de redes heterogêneas, o campo de pesquisa em HIN ainda apresenta abordagens inexploradas. Dentro de tais abordagens ainda carentes de estudo podemos citar: a modelagem de dados e geração de dados sintéticos (VALEJO *et al.*, 2019), estudos do impacto do esquema de HIN sobre os métodos de propagação, a modelagem da função de custo, dentre outras. Neste trabalho, buscou-se o desenvolvimento de uma nova técnica de classificação transdutiva em HIN, tendo como abordagem um nova modelagem para a função de custo com motivações guiada por características dos dados.

Assim, neste trabalho é apresentada um novo método de classificação transdutiva em HIN chamado TCNH [2]. No qual busca-se abordar alguns dos desafios sobre métodos de propagação em HIN usando de uma modelagem diferente para a função de custo. Em especial, buscou-se ao longo do desenvolvimento estudar estratégias de propagação e inicialização dos elementos modelados, conduzindo experimentos para otimizar os resultados do método proposto. Além

[2] Acrônimo do inglês para *Transductive Classification in Heterogeneous Network*

disso, o estudo apresentado nesta tese estende para classificação de HIN gerais o trabalho de Faleiros, Rossi e Lopes (2017), o qual se concentrava apenas em redes bipartidas. Neste trabalho, assim com no de Faleiros e colegas, utiliza-se da divergência de Kullback-Leibler (KL) como medida de similaridade entre vértices conectados.

Para a validação dos resultados do método TCHN, este foi comparado com outros três métodos bastante significativos da área de classificação transdutiva em HIN. Foram consideradas técnicas com diferentes abordagens na modelagem da propagação sobre a mesma rede, dentre elas, técnicas de propagação geral, onde TCHN também se encontra, tendo como comparação a técnica GNetMine; e técnicas de propagação baseadas em meta-caminhos considerando as técnicas HetPathMine e HeteClass. Como pode ser observado pelos resultados apresentados no Capítulo 6, o método proposto apresenta resultados promissores para redes geradas sobre diferentes conjuntos de dados, sendo dois conjuntos de dados do mundo real e também redes geradas de forma sintética. Ao longo dos experimentos o método TCHN se mostrou sempre dentre os melhores, em termos de resultados, sendo em alguns casos superior às demais técnicas.

Além do método desenvolvido que é tema desta tese, ao longo deste doutorado foram feitos diversos outros trabalhos, estudos e colaborações gerando diversas contribuições apresentadas na Seção 1.1. Em especial, ainda no contexto de HIN, foi realizado o desenvolvimento de uma ferramenta para a geração de redes sintéticas chamada HNOC, uma das necessidades da área já apontada na literatura (ANGELOVA; KASNECI; WEIKUM, 2012). Tal ferramenta foi desenvolvida partindo da ferramenta BNOC de geração de redes bipartidas, tal ferramenta possui a vantagens de ter sido idealizada de forma modular de forma que sua extensão para redes heterogêneas gerais foi uma tarefa possível. Os detalhes mais importantes deste desenvolvimento são explicado no Capítulo 5.

## 1.1 Contribuições do trabalho

A seguir apresentamos as principais contribuições resultantes deste trabalho de doutorado:

- **Publicação do método TCHN**:

  Para a divulgação dos resultados da contribuição principal desta tese, que é o desenvolvimento do método de classificação transdutivo em redes heterogêneas de informação, foi escrito um artigo submetido para a revista JCST e se encontra atualmente em revisão, citado como:

  ROMANETTO, L. M.; LOPES, A. A.. Semi-supervised Classification in Heterogeneous Information Networks based on KL-divergence. **Journal of computer science and technology**.

  Submetido em revisão.

- **Publicação da ferramenta BNOC**:

  Para a divulgação da ferramenta HNOC, seus resultados foram divulgados como uma extensão da ferramenta BNOC no mesmo artigo na forma de anexo, na seguinte publicação:

  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

## 1.2 Outras atividades e colaborações

A seguir apresentamos outras atividades e colaborações desenvolvidas ao longo deste trabalho de doutorado:

- **Estágio no Google em buscas esportivas**:

  Durante período deste trabalho, a candidata realizou uma estágio de doutorado na empresa Google no Brasil, tento trabalho dentro da área de buscas esportivas. Neste estágio a candidata desenvolveu um projeto *full-stack* enfrentado diversos desafios de mundo real, o que veio a enriquecer bastante suas experiências e conhecimentos.

- **Publicação de estudos sobre tensores**:

  No inicio de seu doutorado a candidata realizou o estudo sobre **tensores** uma outra estrutura de representação de dados heterogêneos, sendo este estudo focado na detecção e visualização de padrões sociais, gerando a seguinte publicação:

  ROMANETTO, L. M.; Leao, A ; Dias, F ; NONATO, L. G. . Tensor Decomposition for Multi-way Time-Varying Data Visualization. **WVIS - SIBGRAPI 2017 - XXX Conference on Graphics, Patterns and Images, 2017**

  Tal publicação foi reconhecida no Workshop de Visualização do SIBGRAPI 2017 com o seguinte prêmio:

  **Best paper no WVIS 2017**

- **Colaboração para a ferramenta VisLattes**:

  Como parte dos estudos de decomposições tensoriais, a candidata realizou uma colaboração na área de visualização e detecção de padrões para decomposições matriciais fazendo parte da seguinte publicação:

  DIAS, M. D. S. S.; MANSOUR, M.; ROMANETTO, L.; OLIVEIRA, M. C.; NONATO, L. Vizlattes: a tool for relevance analysis from scientific co-authorship networks. In:

**Workshop on visual analytics, information visualization and scientific visualization**. [S.l.: s.n.], 2015.

## 1.3 Estrutura da Tese

O restante desta tese está organizado da seguinte forma:

- Capítulo 2 - Redes heterogêneas de informação: revisão e conceitos básicos: primeiramente é apresentada uma conceituação sobre redes heterogêneas de informação e alguns conceitos preliminares de classificação em redes. Em seguida, são apresentados o levantamento bibliográfico dos principais trabalhos encontrados na área com relevância para esta tese e alguns exemplos de redes heterogêneas de mundo real.

- Capítulo 3 - Classificação transdutiva: neste capítulo é apresentada a fundamentação teórica dos métodos de classificação transdutiva, os quais são então contextualizados para sua aplicação sobre redes heterogêneas de informação. Em seguida, são apresentados os principais métodos de tal área brevemente descritos, tais métodos são posteriormente utilizados na validação do método proposto nesta tese.

- Capítulo 4 - TCHN um novo método para a classificação transdutiva em HIN: este capítulo apresenta a principal contribuição desta tese, que é o desenvolvimento do método TCHN, mostrando diversos estudos e experimentos realizados para a otimização dos resultados de tal técnica com base em estratégias de propagação e inicialização dos elementos.

- Capítulo 5 - Geração de redes heterogêneas sintéticas: este capítulo brevemente a ferramenta HNOC gerada ao longo deste trabalho como colaboração com o grupo de pesquisa, e motiva pela necessidade surgida ao longo do estudo para a validação do método TCHN.

- Capítulo 6 - Experimentos: são apresentados neste capítulo os principais resultados obtidos neste trabalho. Mostrando a comparação do método TCHN com outros métodos significativos na área.

- Capítulo 7 - Conclusão: finalmente concluímos discutindo os principais resultados e contribuições deste trabalho, levantando os bons resultados e vantagens das técnicas desenvolvidas, e apontando alguns possíveis caminhos para desenvolvimentos ainda inexplorados neste trabalho.

CAPÍTULO

2

# REDES HETEROGÊNEAS DE INFORMAÇÃO: CONCEITOS BÁSICOS E TRABALHOS RELACIONADOS.

Redes heterogêneas de informação (HIN)[1] possuem por natureza uma alta flexibilidade na representação de diferentes tipos de informação, possibilitando seu uso em uma grande variedade de dados (SUN; HAN, 2012). Em biologia, HIN tem sido utilizada para representar cadeias alimentares de diferentes espécies de animais, como também, dados biomédicos contendo relação de doenças, genes, medicamentos e efeitos colaterais (HIMMELSTEIN; BARANZINI, 2015). Considerando a área de informática, redes heterogêneas são utilizadas para representar conexões entre computadores e outros dispositivos com diferentes sistemas operacionais e/ou protocolos (ABDULLA, 2012). Além disso, elas também são utilizadas em redes sociais, por exemplo formando redes bipartida de contratação separando empregados de empregadores.

Estes são apenas alguns exemplos de dados naturalmente modelados com redes heterogêneas, as quais podem ser consideradas uma das formas mais gerais de redes de informação. Ao contrário das redes homogêneas, HIN contemplam formas distintas de entidades, conexões e estrutura. Devido a essa flexibilidade, essas redes necessitam de uma atenção especial na modelagem e construção tanto dos dados quanto dos métodos usados em cada tarefa de aprendizado de máquina.

Esses aspectos das HIN fazem com que os métodos apresentados na literatura possuam conceitos complexos, que necessitam de uma descrição precisa dos elementos a fim de possibilitar a sua compreensão. Assim, este capítulo tem como objetivo apresentar os principais conceitos e definições para uma boa compreensão da área e dos trabalhos desenvolvidos nesta tese.

[1] do inglês *Heterogeneous Information Networks*

## 2.1 Definições básicas

Nesta seção, temos como objetivo descrever a notação essencial da área de classificação transdutiva em redes heterogêneas de informação, bem como os conceitos envolvidos na modelagem do método proposto. Ao longo deste trabalho, a representação dos elementos matemáticos foram definidas da seguinte forma: letras minúsculas para valores escalares ($x$), letras minúsculas em negrito para vetores ($\mathbf{x}$), letras maiúsculas em negrito para matrizes ($\mathbf{A}$) e com índices sub-escritos para seus elementos ($\mathbf{A}_{ij}$). Conjuntos de elementos serão denotados como letras maiúsculas em itálico ($S$), funções em equações como letras maiúsculas sem estilo (F) e funções de mapeamento com letras gregas ($\phi$). A descrição e notação usada para os elementos da rede serão descritos a seguir.

Partimos de um dos conceitos mais fundamentais desta teoria, que é a definição de uma rede de informação. Esta nos dá a base de como transformar um conjunto de dados em redes onde podemos aplicar os métodos de aprendizado de máquina.

**Definição 1** (Rede de Informação (SHI *et al.*, 2017))**.** Dado um conjunto de dados sintéticos ou do mundo real a ser modelado como uma rede de informação, identifica-se entidades a serem estudadas contidas nos dados, os quais formam o conjunto $X = \{x_1, x_2, \ldots, x_N\}$. Tais entidades são então mapeadas nos vértices da rede $\vartheta : X \to V$. No mesmo conjunto de dados são identificadas as relações/conexões entre as entidades de estudo, assim como, a importância ou peso de cada, estes dão origem às arestas (e seus pesos) da rede, e são representados por $E$ e $W$, respectivamente. Ao fim, uma rede de informação é definida por uma tripla $\mathscr{G} = (V, E, W)$ onde $V = \{v_1, v_2, \ldots, v_N\}$ representa o conjunto de vértices, $E = \{e_{i,j}, i, j \in 1, 2, \ldots, N\}$ o conjunto de arestas e $W = \{w_{i,j} = f(e_{i,j}), e_{i,j} \in E\}$ o conjunto de pesos das arestas, sendo $f$ uma função que atribui a cada aresta um peso. Os vértices e conexões de uma rede possuem também tipos que são gerados pelo mapeamento $\varphi : V \to O$ e $\varphi : E \to R$.

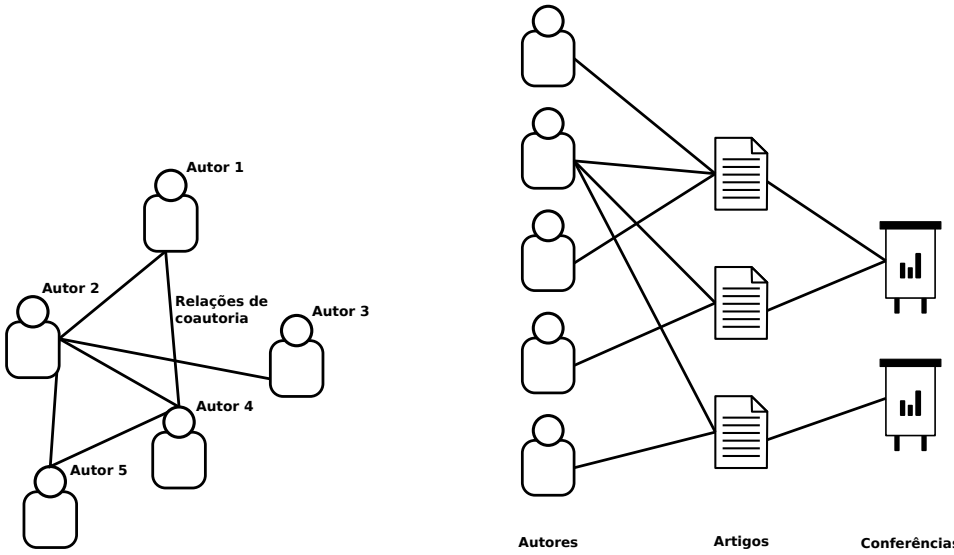
As entidades extraídas do conjunto de dados sob análise podem ser de um mesmo tipo ou de tipos distintos. Por exemplo, redes sociais de amizades, em sua grande maioria, são formadas por pessoas, as quais são *a priori* consideradas iguais dentro da rede, sendo geralmente representadas de uma mesma forma e dotadas de um mesmo conjunto de características. Por outro lado, para alguns conjuntos de dados, faz-se necessário distinguir diferentes tipos de entidades. Tal modelagem pode surgir como uma necessidade intrínseca dos dados, mas também, como uma possibilidade de prover o sistema de aprendizado com uma quantidade maior de informação e, por consequência, um aprendizado com maior embasamento.

Uma rede é caracterizada como homogêneas ou heterogêneas de acordo com a configuração das entidades dentro dela como descrito na Definição 2. Na Figura 1 são ilustradas os dois tipos de rede de informação, ambas representando a atividade de colaboração científica. A Figura 1a mostra uma rede homogênea de informação de conexões entre autores de textos acadêmicos que colaboram entre si; enquanto a Figura 1b apresenta uma rede heterogênea de

informação, incluindo entidades que representam os autores, os artigos publicados e conferências. Além disso, na Figura 1b, a relação de autoria e relação de local de publicação também são representadas, apresentando ligações entre autores com artigos e artigos com conferências.

**Definição 2** (Rede Homogênea / Heterogênea)**.** As redes de informação se dividem em dois grupos: redes homogêneas e heterogêneas. Uma rede é dita heterogênea se as entidades e/ou conexões representadas na rede são de mais de um tipo, ou seja, $|O| > 1$ e/ou $|R| > 1$ ; caso contrário, a rede é dita homogênea.

Figura 1 – Exemplo de redes homogênea e heterogênea de informação, baseadas em dados bibliográficos de colaboração cientifica.




(a) Rede homogênea de colaboração cientifica.

(b) Rede heterogênea de colaboração cientifica.

Fonte: Elaborada pelo autor.

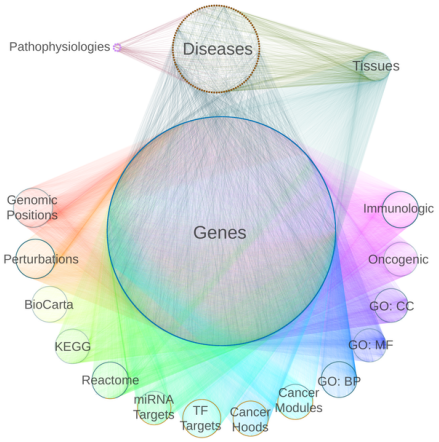
Chegamos então ao foco de estudo desta tese que é o conceito de HIN, que por completude é formalmente apresentado na Definição 3.

**Definição 3** (Rede Heterogênea de Informação (HIN))**.** Dado um conjunto de dados com $m$ tipos de objetos, para cada tipo $i$, tem-se o conjunto $X_i = \{x_{i1}, \ldots, x_{in_i}\}$ de objetos deste tipo, e $\mathscr{X} = \{X_1, X_2, \ldots, X_m\}$ o conjunto de todos os objetos, e entre estes objetos são estabelecidas relações $E = \{e_{ij,pq} | i, j \in \{1,2,\ldots,m\}, p \in \{1,2,\ldots n_i\}, q \in \{1,2,\ldots n_j\}\}$. Uma rede heterogênea de informação é formada pelo grafo $G = (V, E, W)$, onde $V = \bigcup_{i=1}^{m} X_i$ com $m > 1$, $E$ é o conjunto das relações entre dois objetos de $V$, e $W$ é o conjunto dos pesos das relações.

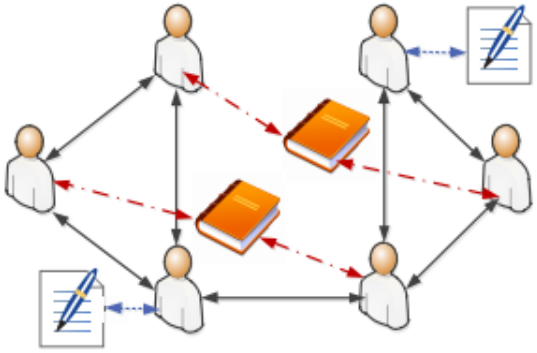
Assim redes de diferentes formatos estão inseridas na categoria de redes heterogêneas. Alguns exemplos são redes bipartidas, redes multi-relação e redes estrela ou multi-hub. As redes bipartidas são formadas por dois tipos de elementos, cuja conexão são apenas entre elementos de tipos distintos. Essas são largamente usadas em dados reais para modelar a iteração de dois

tipos de entidades, tais como documentos textuais e termos, pixeis e imagens, filmes e atores. As redes multi-relação são formas bastante uteis no estudo de redes sociais, onde os usuários se conectam e interagem de diferentes formas, uma vez que essas redes possuem apenas um tipo de elemento e múltiplos tipos de conexões. Já as redes multi-hub são redes com diversos tipos de entidades, onde algumas desempenham um papel central, e a maioria se conectam apenas com os tipos chamados *hubs*, os quais podem conectar com outros *hubs*. Um exemplo deste tipo de rede é mostrada na Figura 2a, que representa uma rede multi-hub baseada em dados biomédicos, onde podemos observar o papel central dos genes e das doenças na representação dos dados. Na figura 2, diferentes formatos de HIN são mostrados, os quais foram tirados de trabalhos encontrados na literatura de classificação transdutiva, como referenciado na figura.

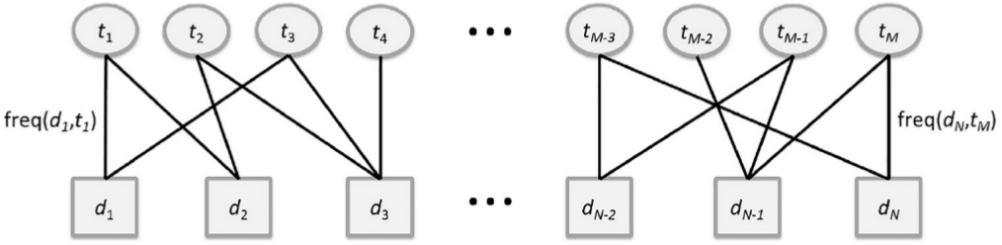
Figura 2 – Exemplo de HIN de diferentes formatos encontrados na literatura.




(a) Rede multi-hub construída com base em dados biomédicos. Fonte (HIMMELSTEIN; BARANZINI, 2015)



(b) Rede heterogênea construída com base em dados de colaboração científica. Fonte (SUN; HAN, 2013)




(c) Rede bipartida construída com base em dados textuais. Fonte (ROSSI; LOPES; REZENDE, 2017).

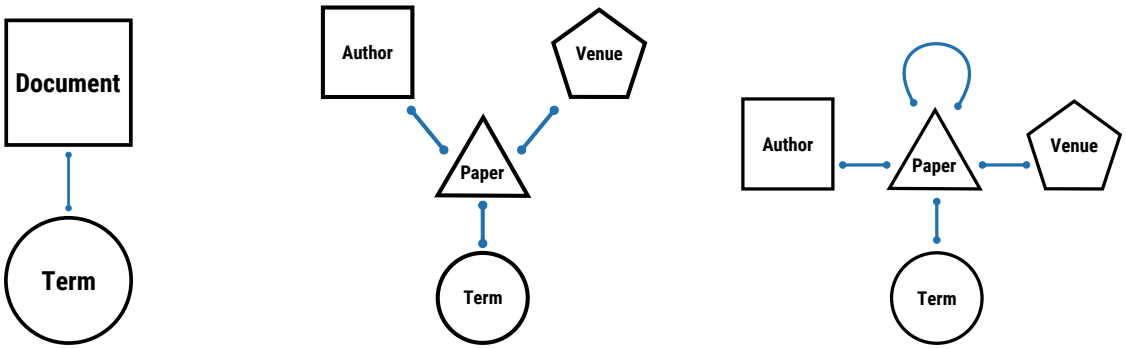
De acordo com os objetivos de pesquisa, um mesmo conjunto de dados pode dar origem a diferentes HINs, como por exemplo, um conjunto de dados bibliográficos pode ser modelado como uma rede bipartida de artigos e termos, ou então pode incluir os autores e revistas que se relacionam com os artigos, formando assim uma rede $k$-partida. A escolha das entidades e conexões a serem representadas pela HIN é essencial, podendo enriquecer a informação extraída dos dados, ou então apenas adicionar redundâncias. Assim, a caracterização da estrutura base de uma HIN é muito importante para a melhor representação e estudo dos dados. Tal caracterização

é feita pelo chamado esquema de rede descrita na Definição 4.

**Definição 4** (Esquema de Rede)**.** (LAO; COHEN, 2010) Um esquema de rede, denotado como $T_G = (O, R)$, é uma meta representação de uma HIN que define um modelo para a construção, restringindo as redes a serem construídas, e contendo objetos do tipo $O$ e relações do tipo $R$.

Na Figura 3, alguns exemplos de esquema para HIN baseados em dados bibliográficos são ilustrados. A Figura 3a apresenta a estrutura de uma rede bipartida entre termos e documentos; a Figura 3b mostra uma rede k-partida estrela, onde artigos se conectam a todos os outros tipos de elementos na rede; e por fim, a Figura 3c ilustra uma rede heterogênea onde existe conexões entre elementos de um mesmo tipo. Complementarmente, um exemplo de rede deste esquema, chamada instância, é ilustrado na Figura 4.

Figura 3 – Exemplos de esquemas de HIN.



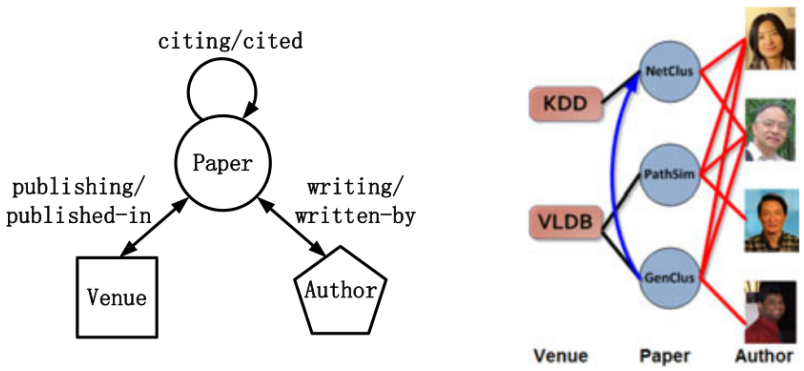

(a) Esquema de rede bipartida.

(b) Esquema de $k$-partida estrela.

(c) Esquema de rede heterogênea irrestrita.

Fonte: Elaborada pelo autor.

Figura 4 – Exemplo de um esquema de rede heterogênea de informação de dados bibliográficos.




(a) Esquema de rede.

(b) Instância do esquema de rede.

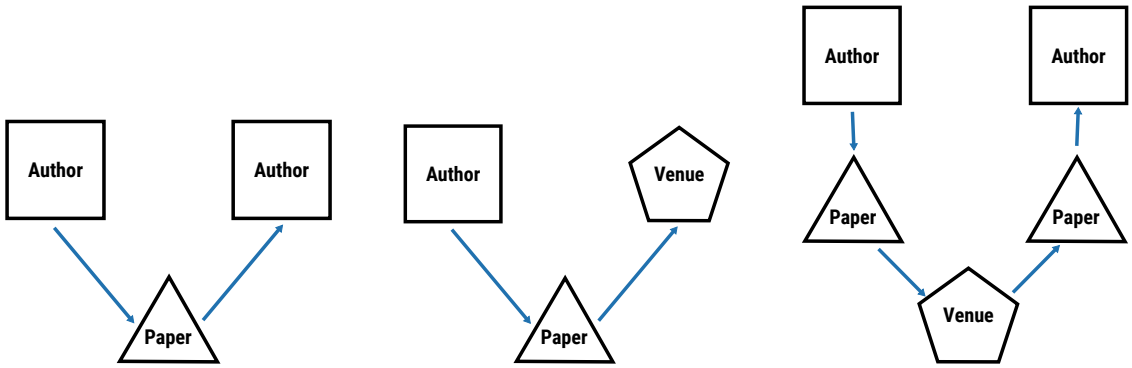
Fonte: Sun e Han (2012).

A modelagem do esquema da rede faz parte da etapa de pré-processamento de dados, que impacta diretamente nos resultados a serem obtidos ao estudar um conjunto de dados via HIN. Assim, o estudo dos esquemas de rede torna-se uma etapa importante na construção de uma boa metodologia, o que gerou, na literatura, diversos trabalhos que exploram e desenvolvem conceitos para melhor utilizar o esquema de rede e acoplar suas características na construção dos métodos (SUN *et al.*, 2013). A seguir são apresentados os principais conceitos construídos a partir de esquemas de rede que serão uteis no desenvolvimento desta tese.

**Definição 5** (Meta-caminhos)**.** (SUN *et al.*, 2011) Um meta-caminho $\mathscr{P}$ é um caminho definido sobre um esquema de rede $T_G = (O,R)$, e é representado como $O_1 \xrightarrow{R_1} O_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} O_{l+1}$, que define uma composição de relações $R = R_1 \circ R_2 \circ \cdots \circ R_{l+1}$ entre os objetos $O_1, O_2, \ldots, O_{l+1}$.

A figura 5 ilustra exemplos de meta-caminhos em uma HIN definida pelo esquema dado na Figura 3c. Como pode ser visto pelos meta-caminhos ilustrados, estes possuem a capacidade de representar relações complexas existente entre os elementos da rede, os quais não podem ser diretamente representados por uma aresta. Na Figura 5a, a relação de co-autoria é ilustrada, em que dois autores são autores de um mesmo artigo, enquanto a Figura 5b apresenta a relação de um autor que se relaciona com uma conferência através da publicação de artigo nesta, pelo meta-caminho ( autor - artigo - conferência ). Por fim, a Figura 5c mostra a relação de dois autores que publicam artigos em uma mesma conferência, o que pode indicar áreas de pesquisa próximas.

Figura 5 – Exemplo de meta-caminhos criados em rede heterogênea de informação de dados bibliográficos, em que A indica os autores, P indica os artigos e V indica uma conferência.




(a) A - P - A (b) A - P - V (c) A - P - V - P - A

Fonte: Elaborada pelo autor.

Os meta-caminhos são utilizados para explorar as conexões da HIN sem que necessariamente cada aresta envolvida seja tratada. Por exemplo, na Figura 5b, o meta-caminho produzido cria uma ligação direta entre autores e conferências considerando as relações com artigos apenas de maneira indireta. Em geral, os meta-caminhos são criados para representar relações com significado semântico. Neste contexto, diferentes meta caminhos são criados e ponderados de

acordo com métricas de relevância, tais como PathSim (SUN *et al.*, 2011), DPLRel (GUPTA; KUMAR; BHASKER, 2015) e HeteSim (SHI *et al.*, 2014). Estas construções serão vistas com detalhes no próximo capítulo, onde são descritas suas utilizações em diferentes métodos de classificação transdutiva.

Dado um meta caminho sobre uma HIN, ligando os objetos de tipos $O_1$ e $O_{l+1}$, podemos ver este meta caminhos como uma rede bipartida ou homogêneas (no caso onde $O_1 = O_{l+1}$), pois conecta dois tipos de elementos. Para dar peso a esta conexão, utiliza-se matrizes de pesos de caminhos descrita na Definição 6

**Definição 6** (Matriz de pesos de caminhos)**.** (GUPTA; KUMAR; BHASKER, 2015) Dado uma rede heterogênea $G$, e um tipo de meta caminho definido sobre esta $\mathscr{P} = (O_1 \xrightarrow{R_1} O_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} O_{l+1})$, a matriz de peso de caminho para $\mathscr{P}$ é definida como a composição $M = W_{O_1O_2} \times W_{O_2O_3} \times \cdots \times W_{O_lO_{l+1}}$, onde $W_{O_iO_j}$ é a matriz de adjacência entre os elementos dos tipos $O_i$ e $O_j$.

Considerando as matrizes de adjacência como matrizes binárias, cada entrada $M[x_i, y_j]$ da matriz de peso nos dá a quantidade de meta caminhos existentes entre os elementos $x_i$ e $y_j$ relacionados com esta entrada.

As redes heterogêneas trazem uma nova forma de tratar dados multi-tipos em redes, e vem atraindo crescente interesse em diversas tarefas de mineração de dados, como por exemplo, predição de links (SHAHREZA *et al.*, 2017; SHI *et al.*, 2012), agrupamento (YIN; HAN; YU, 2006; PIO *et al.*, 2018), e classificação (SUN; HAN, 2012; FALEIROS; ROSSI; LOPES, 2017). Neste trabalho, temos como foco estudar e desenvolver novas técnicas de classificação transdutiva sobre redes heterogêneas. A seguir são definidos sucintamente alguns conceitos essenciais na modelagem de um método de classificação.

A tarefa de classificação busca identificar as reais classes de um conjunto de amostras. Dada a complexidade de tipos em uma HIN é importante definir o exato escopo de uma classe neste contexto, como feito na Definição 7.

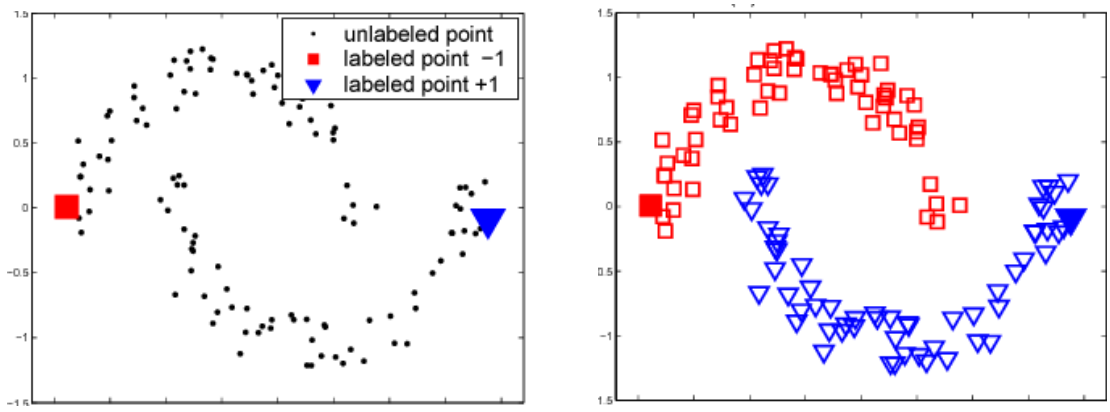
**Definição 7** (Classe)**.** Dada uma rede heterogênea de informação $G = (V, E, W)$, uma classe de $G$ é dada por um subconjunto $V'$ quando $V' \subset V$. Em um caso geral, $V'$ pode conter elementos de quaisquer tipos, ou seja, $V \subset \mathscr{X}$. Porém, quando a classificação é operada apenas em um tipo de objeto, como no caso do método HetPathMine (LUO *et al.*, 2014), temos a restrição adicional de $V' \subset X_t$, onde $t$ é o tipo de objeto classificado.

Especificamente, a classificação transdutiva busca predizer as classes para os elementos de uma rede com base em um subconjunto rotulado, sem gerar um modelo que generalize a classificação para novas amostras, como ilustrado na Figura 6. A Figura 6a apresenta as amostras onde há regiões densas bem definidas e alguns pontos com rótulos conhecidos, enquanto na

Figura 6b, o resultado buscado na classificação é ilustrado, onde regiões de amostras são classificadas como classes de acordo com as amostras pré-rotuladas.

Especificamente para o contexto de redes heterogêneas, que é o escopo deste trabalho, tem-se na Definição 8 a descrição formal para esta tarefa.

Figura 6 – Ilustração da tarefa de classificação transdutiva.




(a) Amostras usadas na classificação. (b) Objetivo final da classificação.

Fonte: Zhou *et al.* (2004).

**Definição 8** (Classificação Transdutiva de Redes Heterogênea de Informação)**.** Dada uma rede heterogênea de informação $G = (V, E, W)$ e um subconjunto de objetos rotulado como valores $\mathscr{Y}$ representando as classes dos objetos, a classificação transdutiva tem o objetivo de predizer a classe de todos os outros elementos não rotulados $V - V'$. Como ilustrado na Figura 6.

Assim temos a definição básica da tarefa de um método de classificação transdutiva. Para a construção de cada algoritmo de classificação transdutiva encontrado na literatura são usados diversos outros conceitos e hipóteses, os quais são detalhadamente descritos no Capítulo 3, onde são contextualizados dentro de sua grande área de aprendizado semissupervisionado.

## 2.2 Trabalhos relacionados

Muitos dos conjuntos de dados do mundo real são por natureza heterogêneos, contendo em sua estrutura multi-tipos de elementos e/ou conexões. Esse aspecto faz com que a representação por redes heterogêneas seja uma opção adequada para estes dados, quando comparada com a representação mais tradicional produzida por redes homogêneas. Além disso, com a evolução atual das técnicas baseadas em HIN, existe um interesse crescente por novas técnicas para diversas tarefas desta área. Nesta seção buscamos dar uma visão geral das técnicas encontradas na literatura em diferentes abordagem focadas na classificação transdutiva em HIN, e assim embasar a teoria proposta nesta tese dentro dos recentes desenvolvimentos da área.

### 2.2.1 Abordagens metodológicas

Inicialmente, as técnicas desenvolvidas de classificação em HIN tendiam a generalizar técnicas tradicionais de redes homogêneas para este contexto. Um exemplo disto é o trabalho de Ji *et al.* (2010), que propõe a técnica GNetMine de regularização baseada nas hipóteses de consistência local e global, estendendo o trabalho de Zhou *et al.* (2004). A GNetMine pode ser considerada umas das técnicas mais conhecidas e simples de classificação transdutiva em HIN. Apesar de ser uma técnica já bem consolidada, ela não explora a modelagem da função de custo na regularização, e apresenta resultados já não tão atrativos em vista das técnicas mais atuais. Outra técnica bastante similar, proposta por Bangcharoensap *et al.* (2016), parte da mesma base de regularização baseada no grafo. Porém, esta utiliza a medida de grau de intermediação (*betweenness*) das arestas para limitar a influência de arestas localizadas entre comunidades, o que a difere do método GNetMine, que limita a influência de vértices centrais.

Os meta caminhos também são muito utilizados, pois são capazes de capturar indiretamente relações de vizinhança reduzindo as operações iterativas de propagação e ao mesmo tempo podendo acoplar semântica em sua construção, como o caso da técnica HetPathMine proposta por Luo *et al.* (2014). Essa técnica busca classificar apenas um tipo de elemento na HIN, chamado tipo alvo, e constrói um conjunto de meta-caminhos cíclicos em torno do tipo alvo (ou seja, meta caminhos que se iniciam e terminam no tipo alvo). Constrói matrizes de relação do tipo alvo para o mesmo com base em cada meta-caminho. A classificação se faz com base nas matrizes de relação que são ponderadas por pesos, os quais são aprendidos por um processo irrestrito de aprendizado supervisionado.

Mais recentemente, a técnica HeteClass, proposta por Gupta, Kumar e Bhasker (2017), propõe um *framework* para a classificação transdutiva sobre um tipo alvo, buscando automatizar a criação de meta-caminhos e seus pesos através de técnicas de aprendizado baseada na medida de similaridade DPRel (GUPTA; KUMAR; BHASKER, 2015), que é capaz de otimizar os pesos para obter melhores resultados. Apesar de muitas vezes se mostrarem bastante eficientes, a restrição de operações no processo iterativo apenas no tipo alvo faz com que tal técnica forneça uma classificação apenas para este tipo, além de não utilizar dados pré-rotulados de outros tipos, o que pode ser desejável em alguns contextos.

Alguns trabalhos argumentam que um conjunto de classes relativos a um tipo não devem ser utilizados na classificação da rede por completo, argumentando que os rótulos de classe são semanticamente específicos de cada tipo (ANGELOVA; KASNECI; WEIKUM, 2012). Porém, em alguns casos, os conjuntos de rótulos coincidem, como é o caso do conhecido conjunto de dados *four-areas* da base DBLP [2] (LEY, 2002) de dados bibliográficos. Neste conjunto os dados possuem quatro classes relacionadas às áreas de pesquisa, as quais são usadas para rotular o tipo dos elementos: autor e conferência. Com isso, se estivermos interessados em obter a classificação dos autores, podemos utilizar as informações de classe de conferências, onde os

[2] https://dblp.uni-trier.de/

autores publicam, para enriquecer o aprendizado. Sendo assim, neste trabalho, consideramos que, quando o conjunto de rótulos de classe coincidem entre tipos, toda a informação disponível deve ser utilizada.

Outros dois tipos de técnicas encontradas são as baseadas em dados relacionais (ELMASRI; NAVATHE, 2010) e em *random walk* (SZUMMER; JAAKKOLA, 2002). No caso da proposta de Serafino, Pio e Ceci (2018), as redes heterogêneas são transformadas em uma base de dados relacional de forma que a classificação se dá pela extensão da técnica Mr-SBC, a qual explora o método de classificação naive-Bayes (MCCALLUM; NIGAM *et al.*, 1998) no contexto de dados multi-relacionais. Pio *et al.* (2018) projetaram recentemente outra técnica chamada HENPC, a qual explora a modelagem de redes heterogêneas como dados relacionais, focando no tratamento da classificação multi-tipos, ou seja, em possibilitar a classificação de mais de um tipo de vértice na rede heterogênea.

Abordagem estendida de técnicas tradicionais, nas técnicas baseadas em *random walk* o processo iterativo se faz baseado em técnicas estocásticas para a propagação dos rótulos. Angelova, Kasneci e Weikum (2012) e Angelova *et al.* (2009) propuseram a técnica Graffiti, que busca tratar o grafo como um mapa de cores, propagando uma cor (ou rótulo) pelo grafo entre dois vértices de acordo com sua mútua influência, realizando transições ou saltos entre vizinhos imediatos ou por vários níveis ou saltos aleatórios na rede.

### 2.2.2 Redes bipartidas

Uma das formas mais simples e também mais utilizadas de redes heterogêneas são as redes bipartidas (ROSSI *et al.*, 2014). Como mencionado anteriormente, redes bipartidas são formadas por dois tipos de elementos distintos ligados por arestas. As redes bipartidas são largamente utilizadas, pois são uma alternativa à representação atributo-valor (FALEIROS; ROSSI; LOPES, 2017), que é tradicionalmente empregada para a representação de dados. Essas redes usam técnicas bastante consolidadas de decomposição matricial como SVD (GOLUB; LOAN, 2012), PCA (JOLLIFFE, 2011) e NMF (LEE; SEUNG, 1999). A representação de dados via redes bipartidas possui diversas vantagens em relação à tradicional atributo-valor (FALEIROS; ROSSI; LOPES, 2017), pois contorna o problema de esparsidade dos dados, garantindo um melhor uso de memória. Além disso redes bipartidas também apresentam maior flexibilidade para a inserção de elementos ou estruturas topológicas dos dados; e demonstra melhores resultados na extração de padrões em relação à representação atributo-valor (BREVE *et al.*, 2012).

Na literatura, encontramos muitas técnicas que exploram as redes bipartidas e suas vantagem na representação de dados. Em Rossi, Lopes e Rezende (2016), um métodos de classificação transdutiva é proposto para a tarefa de classificação textual, neste trabalho, tanto os documentos rotulados quanto os não rotulados são utilizados para induzir a classe dos termos, e as classes induzidas dos termos são usadas para prever a classe dos documentos. Com esta base, a função de custo é modelada como a diferença quadrática dos valores ponderados. Já

em Faleiros, Rossi e Lopes (2017) a divergência de Kullback–Leibler é utilizada como função de similaridade para a construção da regularização, o que traz uma novidade em relação às outras técnicas. No desenvolvimento de técnicas de propagação em redes bipartidas encontramos diversas similaridades e bases teóricas nos modelos vetoriais, isso se deve ao fato de muitos dados vetoriais possuírem uma representação direta por redes bipartida, além dessa representação ser mais compacta, tratando assim, a esparsidade presente em muitos dados.

### 2.2.3 Aplicações

Além das diferentes abordagens metodológicas da área, é importante citar as aplicações das técnicas em áreas como: mineração de dados textuais, reposicionamento de fármacos, dentre outras. Dado que grande parte dos dados disponíveis tem o formato textual, diversas tarefas são operadas com a utilização de classificação transdutiva, como por exemplo a classificação de entidades presentes em uma base de dados textuais. Em diversos trabalhos na literatura, a base de dados textuais de colaboração científica DBLP é explorada na classificação dos autores presentes nos dados (JI *et al.*, 2010; BANGCHAROENSAP *et al.*, 2016; LUO *et al.*, 2014; GUPTA; KUMAR; BHASKER, 2017).

A classificação transdutiva em HIN também vem sendo aplicada no contexto biomédico no reposicionamento de fármacos, o qual é motivado pela necessidades do desenvolvimento de novos fármacos, que em geral, é um processo lento e custoso. Como uma alternativa ao desenvolvimento de novos fármacos do zero, o reposicionamento busca o estudo de fármacos já existentes aplicados a doenças ainda carentes de terapias. Utilizando grandes bases de dados biomédicas, os métodos computacionais buscam descobrir potenciais fármacos para o estudo clínico em novas terapias.

O método MINProp (HWANG; KUANG, 2010) utiliza uma propagação alternada de rótulos entre elementos do mesmo tipo e elementos de tipos distintos de acordo com o esquema da rede. Ao final, ranqueia os genes mais relacionados, a fim de obter estudos em potencial de diferentes doenças que possam ter terapias similares.

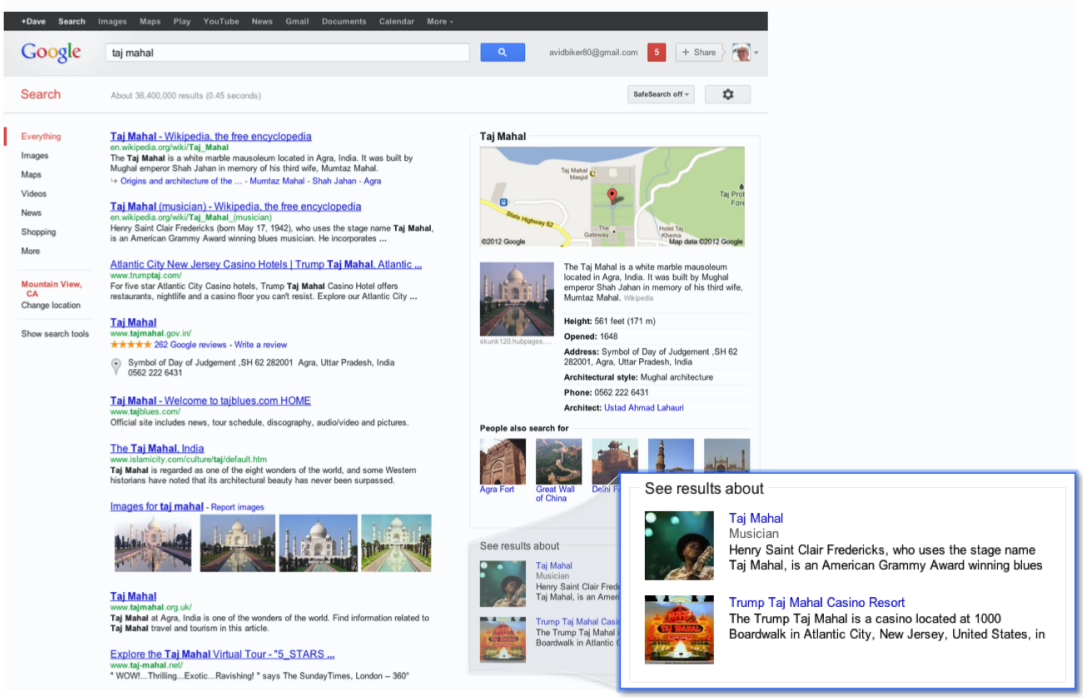
Já em Shahreza *et al.* (2017), o método Heter-LP integra diferentes bases de dados sobre doenças, fármacos e proteínas, na forma de uma HIN, e aplica uma propagação de rótulos para descobrir novas relações entre fármacos-doenças e entre fármacos-proteínas. A grande novidade deste método é o uso de toda a informação extraída das bases de dados e a propagação de rótulos distintos. Ambos os métodos trazem novidades na combinação de diferentes bases de dados para a busca de novas relações entre alvos e doenças. Por outro lado, elas são modelados com base nos pressupostos de consistência, trazendo apenas modificações nos dados e na ordem do processo iterativo.

### 2.2.4 Redes heterogêneas do mundo real

Com a crescente disponibilidade de dados gerados para representar os sistemas cada vez mais complexos presentes no mundo físico e virtual, como por exemplo: dados de Sensu populacional, sinais de equipamentos eletrônicos de monitoramento, base de dados de texto entre outras. Uma forma bastante tradicional para a representação de dados coletados em mundo real são as bases de dados relacionais, que permite definir conjuntos de entidades distintas e registrar diferentes relações com o uso de múltiplas tabelas. Exatamente por essa característica, os dados relacionais são facilmente transformáveis em redes heterogêneas, as quais podem extrair padrões latentes com o uso das técnicas de aprendizado de maquina baseadas em HIN. A base de dados DBLP é um exemplo de base de dados relacional, que transformada em HIN, é largamente utilizada como *baseline* para a validação de técnicas de classificação transdutiva.

Outra base de dados encontrada na forma de HIN é a IMDB [3], que contém dados relacionais sobre música, cinema, filmes, programas e comerciais para televisão, e jogos de computador [4] (PEMBERTON, 2008). Tal base de dados é composta de diversas tabelas que podem ser cruzadas com o uso de modelagem via HIN.

Figura 7 – Exemplo de uma busca no Google sobre 'taj mahal' que utiliza informações estruturadas provenientes da *Knowledge Graph*, produzindo um resultado mais completo que apresenta informações sobre entidades relacionadas ao tópico pesquisado, além do local geográfico.




Fonte: Singhal (2012).

Além da transformação de dados em formato tradicional para HIN, atualmente já existem

[3] IMDB vem termo em inglês *Internet Movie Database*.
[4] IMDB está disponível online em <https://www.imdb.com/>.

disponíveis dados construídos nesta forma. Um exemplo disto é a base de dados Knowledge Graph (KG) (SINGHAL, 2012), também conhecida como Google Knowledge Graph. Esta base foi criada com o objetivo de estruturar informações sobre objetos, pessoas e locais. Além disso, ela também serve como base na busca por pessoas, eventos esportivos e pontos turísticos, como no exemplo mostrado na Figura 7. Tal base de dados é baseada no *framework* de representação RDF [5] (ÖZSU, 2016), o qual segue o modelo que parte da tripla **< Subjeito; Propriedade; Objeto >**, onde cada tripla define uma relação entre o **Sujeito** e o **Objeto** descrito pela **Propriedade**. Cada **Sujeito** ou **Objeto** são entidades conhecidas e definidas na KG, e podem ser vistas como vértices da rede, sendo que as relações entre estas são definidas pelas **Propriedades**, que são equivalente às arestas. Assim, a KG assume uma estrutura de HIN, contendo entidades e relações de diferentes tipos.

## 2.3 Considerações finais

Neste capítulo foram apresentados os principais conceitos e definições de representação de dados em redes de informação, base para os estudos desenvolvidos ao longo desta tese. Em especial, o conceito de redes heterogêneas de informação é detalhadamente descrito, bem como diversas construções sobre este. Muitos dos conceitos apresentados aqui são revisitados e contextualizados para aplicações especificas aos longo do texto. Além dos conceitos básicos, é apresentado um levantamento bibliográfico dos métodos mais relevantes encontrados na literatura e aplicações de HIN. Finalmente, são apresentados alguns exemplos de HIN de mundo real, para ilustrar a importância de tal representação e seu atual crescimento.

Como pode ser visto nas discussões apresentadas neste capítulo, apesar da grande atratividade da área de classificação transdutivos em HIN, existe ainda um grande horizonte de possibilidades a ser explorado. problemas de crescente interesse que demandam novos e efetivos métodos.

[5] do inglês *Resource Description Framework*.

CAPÍTULO

# 3

# CLASSIFICAÇÃO TRANSDUTIVA

As técnicas de classificação em aprendizado de máquina podem ser divididas em três grandes grupos: **supervisionado**, **semissupervisionado** e **não supervisionado**. As técnicas supervisionadas se caracterizam por generalizar um conjunto de amostras de dados cuja classificação é conhecida *a priori*, gerando um modelo para a classificação de novas amostras. Este processo se faz em geral, baseado em características associadas às amostras com classes conhecidas.

Dentre redes baseadas em dados de mundo real, em muitos casos não existem naturalmente atributos ou características locais vinculadas aos elementos. Sendo que uma possível abordagem considerar as conexões da rede para a construção de atributos atrelados aos elementos. Porém, tal representação geralmente se torna demasiado custosa e esparsa conforme o número de elementos da rede cresce.

Desta forma, técnicas tradicionais de aprendizado de máquina como a SVM e Naive Bayes não são aplicáveis para tais dados. Além disso, em muitos casos a quantidade de amostras rotuladas do conjunto de dados são escassas, ao ponto de não serem suficientes para que técnicas de aprendizado supervisionado leve à construção de bons modelos de predição.

Por outro lado, o aprendizado semissupervisionado se dá quando a função de predição é aprendida com base nas amostras rotuladas e não rotuladas. Este é especialmente adequado quando as amostras possuem dados rotulados, porém estes não são suficientes para que técnicas supervisionadas levem a uma boa predição e, ao mesmo tempo, quando as amostras não rotuladas adicionam informações importantes, como informações de vizinhança e topologia da distribuição dos dados (WU; SCHÖLKOPF, 2007).

Já as técnicas de classificação não-supervisionado em geral são usadas diferentes modelagem para o particionamento dos dados, e atribuição de classes sem que se tenha o conhecimento da classificação de amostradas *a priori*, apenas considerando outras características dos dados, como : topologia, distribuições de características dentre outras.

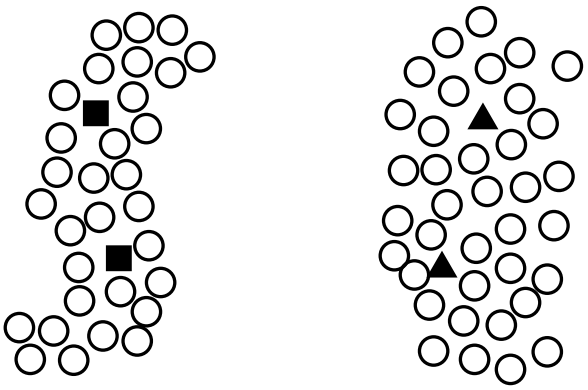
Neste trabalho, abordamos técnicas de classificação semissupervisionada, em especial técnicas transdutivas de propagação com modelagem em redes. Dentro deste conjunto de técnicas encontramos que o principal conceito que dá base ao aprendizado semissupervisionado é o conceito de consistência (ZHOU *et al.*, 2004; AMINI; USUNIER, 2015), que se baseia em duas premissas:

**Premissa 1.** Duas amostras $x_1$ e $x_2$ que são próximas em uma região densa de amostras, tendem a ter rótulos de classe similares ou iguais.

**Premissa 2.** Se duas amostras $x_1$ e $x_2$ estão no mesmo grupo, estas tendem a ter mesmo rótulo de classe.

Tais premissas buscam garantir a consistência do rótulo associado a cada vértice, mantendo a informação de grupos, de forma que grupos de vértices altamente conexos tenham essas informações o mais similares possível, enquanto grupos de vértices disjuntos tenham essas informações dissimilares, como ilustrado na Figura 8. Partindo desta ideia, os métodos de aprendizado semissupervisionado buscam formular um problema de otimização no qual o objetivo é maximizar a similaridade, ou minimizar a divergência, entre os rótulos resultantes da predição de acordo com as premissas citadas.

Figura 8 – Ilustração do conceito de consistência, em que elementos de grupos em regiões densas devem possuir rótulos próximos.



Fonte: Elaborada pelo autor.

## 3.1 Abordagem gráfica

Na construção de um método de aprendizado semissupervisionado, podemos citar em especial três tipos de abordagem: abordagem gráfica, métodos generativos e métodos de descriminantes (ZHOU *et al.*, 2004). Neste trabalho focamos apenas nos métodos gráficos, os quais

expressam a geometria dos dados construindo uma rede sobre todas as amostras disponíveis, rotuladas e não rotuladas. Com base na rede construída, os métodos gráficos buscam espalhar a informação de rótulos a fim de manter a consistência na predição o máximo possível.

A construção de uma rede com base nos dados se dá pela representação do conjunto de amostras $X = \{x_1, x_2, \ldots, x_n\}$, por vértices $V = \{v_1, v_2, \ldots, v_n\}$ do grafo $\mathscr{G}$, e pelas arestas $E$, que recebem pesos de acordo com a ligação, similaridade ou relevância das conexões ou iterações entre as amostras. Tais pesos são usualmente representados por uma matriz positiva $\mathbf{W} = [W_{ij}]$, chamada matriz de pesos, a qual tem valores não zero quando os vértices $i$ e $j$ são conectados. Alguns exemplos mais utilizados de construção da matriz de pesos são as seguintes:

- Matriz binária de iteração:

$$W_{ij} = \begin{cases} 1 & \text{se os vértices } i \text{ e } j \text{ são conectados por uma aresta;} \\ 0 & \text{caso contrário.} \end{cases} \tag{3.1}$$

  Também conhecida como matriz de adjacência, ela é um indicador binário de iteração entre vértices.

- Matriz de peso de interação:

$$W_{ij} = \begin{cases} c_{ij} & \text{se os vértices } i \text{ e } j \text{ tem interação não nula;} \\ 0 & \text{caso contrário.} \end{cases} \tag{3.2}$$

Construída com base no peso de interações entre os vértices, um exemplo bastante conhecido ocorre quando se usa a representação de textos via *bag of words - bow*. Nesse caso, cada documento é representado por um vetor cujas entradas são pesos dados a cada elemento da bow. Por exemplo, usando a frequência do termo (tf) no documento.

Matriz binária dos $k$ vizinhos mais próximos:

Quando os dados são dotados de uma representação local, ou seja um vetor de características $\mathbf{x}_i$, este pode ser usado na construção da rede por meio de medidas de distância. Com as medidas de distância, uma matriz de pesos binária é construída ligando um vértice a seus k vizinhos mais próximos:

$$W_{ij} = \begin{cases} 1 & \text{se os vértices } i \text{ é um dos } k \text{ vizinhos mais próximos do vértice } j; \\ 0 & \text{caso contrário.} \end{cases} \tag{3.3}$$

Matriz de similaridade gaussiana:

Outra possibilidade é a utilização de um kernel gaussiano para a construção da matriz de pesos como:

$$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}} \quad , \forall v_i, v_j \in V, \tag{3.4}$$

onde $\sigma$ é uma parâmetro dado.

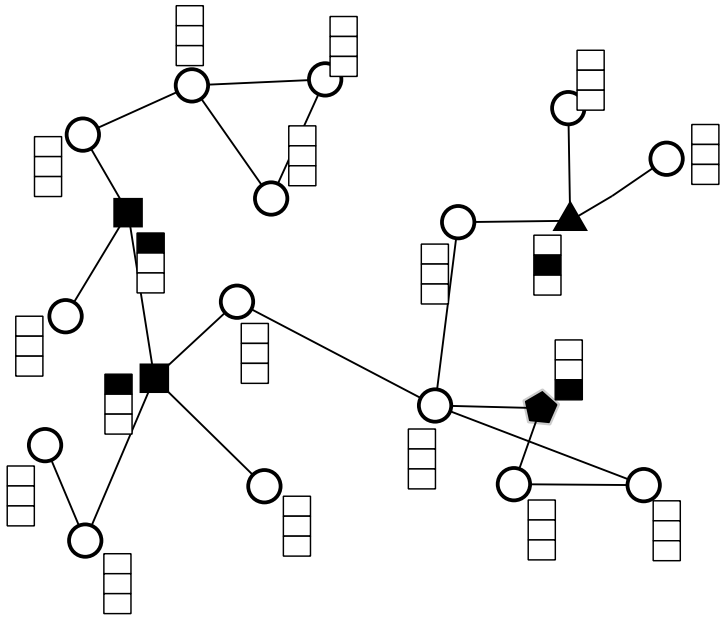
Em seguida, o grafo é formado de acordo com as associações de cada elementos $x_i$ com um vértice $v_i$, através de arestas $e_{i,j}$, que são definidas com base nas entrada da matriz $\mathbf{W}$ quando essas não são nulas. Por fim, obtém-se o grafo $\mathscr{G} = (V,E,W)$, onde $V = \{v_1, v_2, \ldots, v_n\}$ é o conjunto de todos os vértices, $E = \{e_{i,j}\}$ é o conjunto de todas as arestas e $W$ é a matriz de pesos.

A partir desse grafo, uma predição de rótulos para os elementos não rotulados é conduzida com base nos elementos rotulados, em que os rótulos sujeito às premissas de consistência são espalhados de acordo com a geometria do grafo. Assim, dado o conjunto de rótulos conhecidos $C = \{c_1, c_2, \ldots, c_K\}$, cada vértice pré-rotulado $v_i$ do grafo é associado a um vetor $\mathbf{y}_i = [y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}]$ para codificar informações de rótulos, de tal forma que:

$$y_i^{(k)} = \begin{cases} 1 & \text{se os vértices } i \text{ são rotulados com a classe } k; \\ 0 & \text{caso contrário.} \end{cases}$$

Da mesma forma, para a classificação dos elementos, cada vértice $v_i$ do grafo é associado a um vetor $\mathbf{f}_i = [f_i^{(1)}, f_i^{(2)}, \ldots, f_i^{(k)}]$, onde cada entrada $f_i^{(k)}$ define a medida de confiança do elemento $i$ pertencer à classe $k$. Este vetor dá origem à classificação final do método. A Figura 9 ilustra a representação do grafo com os índices dos vértices, arestas e os rótulos multidimensionais associados.

Figura 9 – Representação do grafo com os vértices, arestas e os rótulos multidimensionais associados.

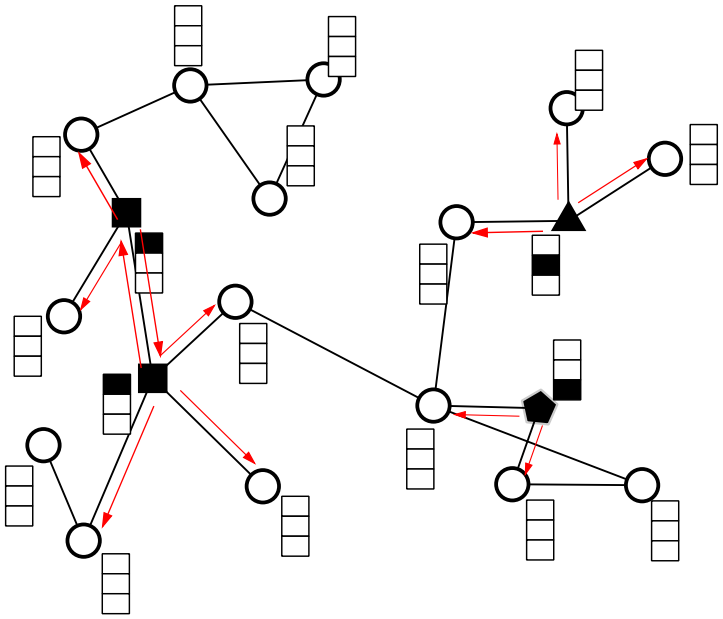


Fonte: Elaborada pelo autor.

A propagação da informação de rótulo é efetuada utilizando a estrutura do grafo. Geralmente, o método de propagação aplicado é o modelo chamado algoritmo de espalhamento (AMINI; USUNIER, 2015), o qual transmite a informação de rótulo de cada vértice a

seus vizinhos como ilustrado na Figura 10. Em suma, este algoritmo busca encontrar os rótulos $Y = (Y^l, Y^u)$, sendo $Y^l$ os vértices rotulados de acordo com a topologia do grafo definido pela matriz de pesos $\mathbf{W}$, de modo a manter a consistência com a informação real.

Figura 10 – Representação do grafo com os índices dos vértices, arestas e os rótulos multidimensionais associados.



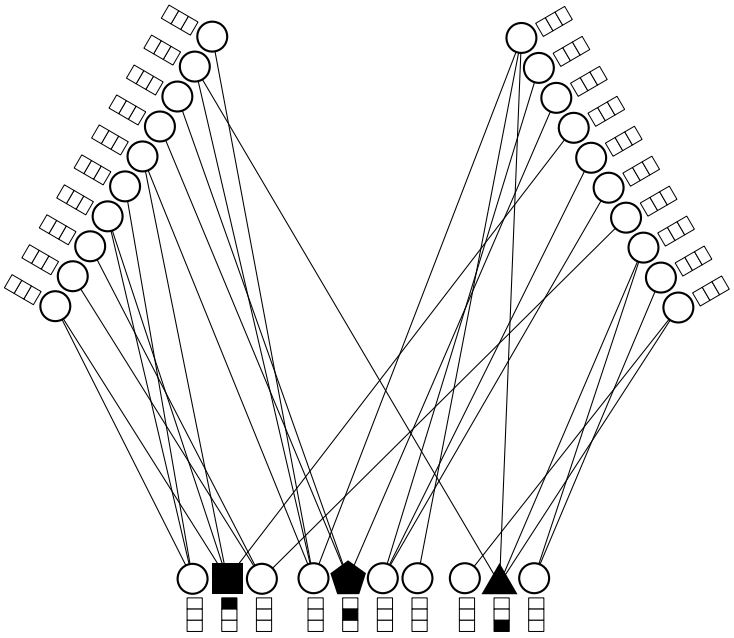
Fonte: Elaborada pelo autor.

Outra variação de métodos de propagação são os baseados em *random-walk* (SZUMMER; JAAKKOLA, 2002) ("passeio aleatório"), os quais são formulados como um modelo de classificação. Esses tem como objetivo solucionar um problema de otimização para a propagação probabilística dos rótulos no grafo. Em outras palavras, a propagação dos rótulos entre os vértices se dá por um modelo probabilístico, o qual é acoplado na função objetivo para a otimização, gerando a predição final.

No caso de redes heterogêneas, a complexidade dos diferentes tipos e camadas devem ser considerados na modelagem, adicionando, assim, vértices e arestas que possuem significados e interações diferentes. Sua construção agora é dotada de uma variedade maior de vértices que podem ou não receber rótulos multidimensionais. Neste contexto, os pesos das arestas são dados por um conjunto de matrizes $W = \{\mathbf{W}_{11}, \mathbf{W}_{12}, \ldots, \mathbf{W}_{mm}\}$, onde $m$ é o número de tipos presentes na rede. Cada matriz $\mathbf{W}_{ij} \in \mathscr{R}^{(n_i, n_j)}$ pode ser definida de acordo com a modelagem utilizada, considerando as possíveis matrizes de peso definidas anteriormente. A Figura 11 ilustra um exemplo simples de uma rede heterogênea, em que os vértices recebem rótulos multidimensionais.

A rede heterogênea e suas conexões fornecem informações relevantes para a tarefa de classificação. Em muitos casos, o objetivo final é classificar apenas um subconjunto de tipos, chamados tipos alvo. Por exemplo, em dados bibliográficos é possível efetuar diferentes

Figura 11 – Representação de uma rede heterogênea com os vértices, arestas e os rótulos multidimensionais associados.



Fonte: Elaborada pelo autor.

classificações devido a quantidade de informação disponível, tais como artigos publicados, palavras que ocorrem nesses artigos, revistas ou conferências de publicação, citações. No entanto, como objetivo final é a classificação de autores em áreas de pesquisa, a rede heterogênea construída não necessita rotular todos os tipos envolvidos em sua estrutura.

Os tipos não classificados na rede heterogênea podem ser utilizados na operação da propagação de rótulos por todos os tipos de dados, adicionando a semântica das conexões à modelagem por meio de estratégias de propagação ou termos de regularização. Como também, podem ser utilizados em meta-caminhos cíclicos que iniciam e terminam nos tipos alvo, operando a classificação sobre a rede homogênea, construída de tipo alvo para tipo alvo, conectada pelos meta caminhos.

Muitos trabalhos defendem que a informação de rótulos relacionados a um tipo não devem ser utilizados na classificação da rede por completo, argumentando que os rótulos de classe são semanticamente específicos de cada tipo, e por isso utilizam apenas a informação de rótulos do tipo alvo. Porém, quando os rótulos entre diferentes tipos coincidem, eles podem ser utilizados com sucesso para enriquecer a qualidade da classificação. Esse aspecto é mostrado na Seção 6.1, em que redes formadas com base no subconjunto de dados DBLP *four-areas* são geralmente modeladas com quatro tipo de vértices (autores, artigos, termos, conferências) e possui classificação disponível para autores e conferências. Nesses experimentos foi possível observar que as informações de rótulo das conferências é bastante útil na tarefa de classificação dos autores, aumentando a qualidade dos resultados.

## 3.2 Modelagem matemática

Formalmente, os métodos transdutivos buscam obter uma classificação para um subconjunto de dados não rotulados sem construir um modelo capaz de estimar o rótulo de novas amostras. Do ponto de vista matemático, os métodos de classificação transdutiva baseados na abordagem gráfica, buscam otimizar uma função de regularização. Tal função é construída com base nos vetores de confiança $\mathbf{f}_i$; nos vetores de classe $\mathbf{y}_i$, as quais são conhecidas *a priori*; e nos valores da matriz de peso, que representa o grafo. Em geral, seu equacionamento possui o formato mostrado na Equação 3.5, onde $\Omega$ e $\Omega'$ são funções de métrica entre os elementos que representam o grafo, e $\mu$ regula a influência dos termos. (ZHU; GOLDBERG, 2009; ZHOU *et al.*, 2004).

$$Q(\mathscr{G}) = \frac{1}{2} \sum_{e_{i,j} \in E} \mathbf{W}_{ij} \Omega(\mathbf{f}_i, \mathbf{f}_j) + \mu \sum_{v_i \in V^l} \Omega'(\mathbf{f}_i, \mathbf{y}_i) \tag{3.5}$$

O modelo de função objetivo, mostrado na Equação 3.5, tem como base as premissas de consistência, em que cada um de seus termos possui um significado próprio. O primeiro termo da equação é responsável por regular a premissa de que os rótulos de elementos ligados por arestas devam ser próximos, ou seja, vizinhos. Enquanto o segundo termo mantém a premissa de que a informação de rótulos preditos para elementos previamente rotulados devam ser próximos às informações originais.

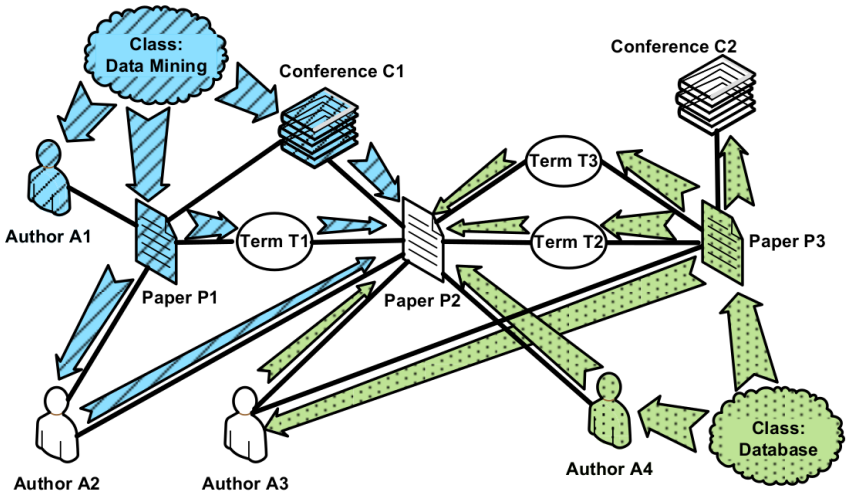
As funções de métricas, como similaridade alta entre vizinhos ou divergência baixa, são modeladas de acordo com características assumidas sobre a distribuição da informação de rótulo. Essas funções podem controlar alguns efeitos indesejados na propagação, como a alta influência de elementos centrais ou com alto grau de intermediação. Além disso, diversos métodos fazem uso de parâmetros para controlar a influência das diferentes camadas da rede heterogênea, bem como da combinação de meta-caminhos.

A seguir, são descritos os principais métodos de classificação transdutiva em HIN encontrados na literatura, no intuito de ilustrar as diferentes abordagens para a construção das redes e modelagem da função objetivo.

## 3.3 GNetMine

O método GNetMine, proposto por Ji *et al.* (2010), é um dos primeiros métodos de classificação transdutiva em HIN, o qual tem sido largamente utilizado na literatura. Este método busca estabelecer um *framework* de regularização com base gráfica, que utiliza as arestas da estrutura da HIN para propagar as classes de um subconjunto rotulado para elementos sem rótulo, como ilustrado na figura 12. Nesse método, todos os tipos de objetos são classificados, ou seja, temos ao final uma classificação para todos os elementos de $X = \bigcup_{i=1}^{m} X_i$.

Figura 12 – Propagação de rótulos em uma rede de dados bibliográficos.




Fonte: Ji *et al.* (2010).

Para operar a tarefa de classificação, uma matriz indicadora de classe é construída de acordo com a seguinte definição:

**Definição 9** (Matriz Indicadora de Classe)**.** Supondo que temos um número de classe igual a $K$. Para qualquer tipo de objeto $X_i, i \in \{1,\ldots,m\}$, a matriz indicadora de classe $\mathbf{F}_i = [\mathbf{f}_i^{(1)},\ldots,\mathbf{f}_i^{(K)}] \in \mathbb{R}^{n_i \times K}$, onde cada $\mathbf{f}_i^{(k)} = [f_{i1}^{(k)},\ldots,f_{in_i}^{(k)}]^T$ apresenta uma medida de confiança de que cada objeto $x_i p \in X_i$ pertence à classe $k$.

O objetivo do método é predizer a matriz indicadora de confiança, e, com base nesta, atribuir a cada objeto $x_{ip}$ de tipo $X_i$ uma classificação $c_{ip}$, encontrando o valor máximo da linha $p$ de $\mathbf{F}_i$, ou seja:

$$c_{ip} = \arg\max_{1 \le k \le K} f_{ip}^{(k)}. \tag{3.6}$$

Considerando os pares de elementos ligados de uma HIN, é possível considerar que cada tipo de aresta representa uma rede de relação $\mathscr{G}_{ij}$, a qual define a interação entre dois tipos de objetos dos conjuntos $X_i$ e $X_j$, sendo $i, j \in \{1,\ldots,m\}$, em que $i$ pode ser igual a $j$. Nesse contexto, uma matriz de relação $W_{ij}$ é definida sobre um grafo de relação $\mathscr{G}_{ij}$. Tal matriz nada mais é do que a matriz de pesos do grafo, sendo definida como:

$$W_{ij,pq} = \begin{cases} 1 & \text{, se os objetos } x_{ip} \text{ e } x_{jq} \text{ possuem uma relação;} \\ 0 & \text{, caso contrário.} \end{cases} \tag{3.7}$$

Na área de classificação transdutiva, a construção da função objetivo utiliza duas suposições básicas baseadas nas premissas de consistência:

1. O estimador de confiança de dois objetos $x_{ip}$ e $x_{jq}$ de uma classe $(k)$, $f_{ip}^{(k)}$ e $f_{jq}^{(k)}$, devem ser similares se $x_{ip}$ e $x_{jq}$ possuem alguma relação, i. e., o valor de $W_{ij,pq} > 0$;
2. O estimador de confiança $\mathbf{f}_i^{(k)}$ deve ser similar a $\mathbf{y}_i^{(k)}$.

Assim, a função objetivo é construída utilizando a métrica euclidiana com os vetores de confiança e indicadores de classe. Equação 3.8 define essa construção, onde $D_{ij,pp}$ é a soma da $p-$ésima linha da matriz $W_{ij}$, e $\lambda_{i,j}$ e $\alpha_i$ são parâmetros obtidos empiricamente. Vale mencionar que o termo $D_{ij,pp}$ é responsável por reduzir a influência dos chamados *hubs* ou nós, os quais possuem muitas ligações no grafo, enquanto os termos $\lambda_{i,j}$ e $\alpha_i$ tem como função regulam a influência das camadas e dos termos da equação.

$$\begin{aligned} J(\mathbf{F}_1,\ldots,\mathbf{F}_m) &= \sum_{i,j=1}^{m} \lambda_{i,j} \sum_{p=1}^{n_i}\sum_{q=1}^{n_j} W_{ij,pq}\left(\frac{1}{\sqrt{D_{ij,pp}}} f_{ip}^{(k)} - \frac{1}{\sqrt{D_{ji,qq}}} f_{jq}^{(k)}\right)^2 \\ &+ \sum_{i=1}^{m} \alpha_i (\mathbf{f}_i^{(k)} - \mathbf{y}_i^{(k)})^T (\mathbf{f}_i^{(k)} - \mathbf{y}_i^{(k)}) \end{aligned} \tag{3.8}$$

Em seguida, partindo da função de custo dada na Equação 3.8 utiliza-se o método dos gradientes descendentes na dedução do Algoritmo 1, derivando a equação em relação às variáveis e igualando este resultado a zero e chega-se por meio de manipulações algébricas ao algoritmo iterativo para a solução que otimiza a função de custo.

**Algoritmo 1** – Algoritmo GNetMine

**para** $\forall k \in \{1,\ldots,K\}, \forall i \in \{1,\ldots,m\}$ **faça**
    inicialize o estimador de confiança $\mathbf{f}_i^{(k)}(0) = \mathbf{y}_i^{(k)}$;
**fim para**
$t = 0$
**enquanto** não atingir o critério de convergência **faça**
    **para** $i \in \{1,\ldots,m\}$ **faça**
        $\mathbf{f}_i^{(k)}(t+1) = \frac{\sum_{j=1,j\neq i}^{(k)} \lambda_{ij}\mathbf{S}_{ij}\mathbf{f}_j^{(k)}(t) + 2\lambda_{ii}\mathbf{S}_{ii}\mathbf{f}_i^{(k)}(t) + \alpha_i \mathbf{y}_i^{(k)}}{\sum_{j=1,j\neq i}^{(k)} \lambda_{ij} + 2\lambda_{ii} + \alpha_i}$, com $\mathbf{S}_{ij} = \mathbf{D}_{ij}^{-\frac{1}{2}}\mathbf{W}_{ij}\mathbf{D}_{ji}^{-\frac{1}{2}}$.
        $t = t+1$
    **fim para**
**fim enquanto**
**para** $\forall i \in \{1,\ldots,m\}$ **faça**
    atribua $p-$ésimo objeto de tipo $X_i$ como da classe
    $c_{ip} = \arg\max_{1\leq k\leq K} f_{ip}^{(k)}$
**fim para**

Como mostrado no artigo que propõe o método, é possível demonstrar que o mesmo algoritmo pode ser obtido pela propagação simples dos rótulos entre vizinhos em uma rede, sendo ponderado pelo grau de cada vértice, e este resultado converge para a solução da função de custo deduzida acima.

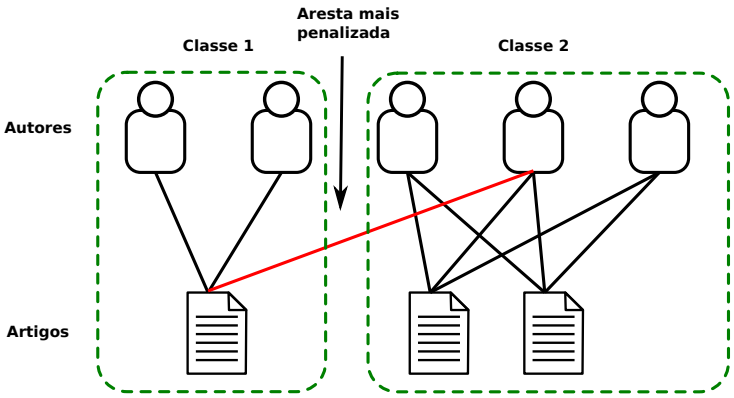
## 3.4 Regularização baseada em normalização com medida de intermediação de arestas

Similar ao método GNetMine, a formulação do método proposto por Bangcharoensap *et al.* (2016) parte das mesmas premissas de consistência base da área de classificação transdutiva. No entanto, este introduz mais uma premissa:

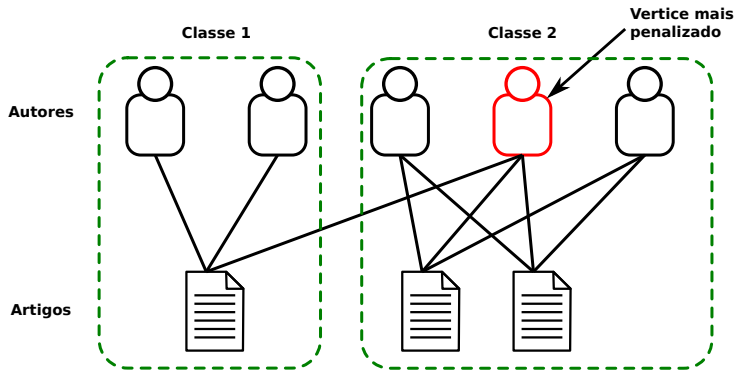
- A propagação de rótulos por arestas localizadas entre comunidades é menos desejável no processo de propagação. Assim, a medida de intermediação de arestas é uma medida mais adequada para a normalização dos elementos de regularização.

Bangcharoensap *et al.* (2016) argumentam que arestas que fazem pontes entre comunidades devem ser reguladas, evitando assim, a influência de rótulos entre comunidades distintas. Esse conceito difere do apresentado no método GNetMine, o qual utiliza a normalização baseada no grau dos vértices para evitar que estes dominem o processo de propagação. A Figura13 ilustra os elementos mais penalizados em uma rede para cada tipo de regularização utilizado.

Figura 13 – Ilustração dos elementos mais penalizados para cada tipo de regularização utilizada nos métodos apresentados.




(a) Exemplo de elementos penalizados pela regularização baseada na medida de intermediação de arestas.




(b) Exemplo de elementos penalizados pela regularização baseada no grau dos vértices.

Fonte: Elaborada pelo autor.

Tipicamente, a medida de intermediação é definida originalmente para redes homogêneas, sem considerar a possibilidade de haver mais um tipo de vértice da rede. No caso do método de Bangcharoensap *et al.* (2016), uma nova medida é proposta, a qual é baseada no conceito de intermediação de arestas para redes heterogêneas. A medida de centralidade para uma aresta $e = (v_{ip}, v_{iq})$, onde $v_{ip} \in V_i$ e $v_{iq} \in V_j$, é defina de acordo com a Equação 3.9, em que $EP(e)$ é o conjunto dos pontos extremos de $e$, ou seja, $\{v_i, v_j\}$; $\sigma(s,t|e)$ é o número de menores caminhos entre $s$ e $t$ passando por $e$; e $\sigma(s,t)$ é o total de caminhos mínimos entre $s$ e $t$.

$$C(e) = 1 + \sum_{s \in V_i \setminus EP(e)} \sum_{t \in V_j \setminus EP(e)} \frac{\sigma(s,t|e)}{\sigma(s,t)} \tag{3.9}$$

A exclusão dos vértices extremos é inspirada no trabalho de Freeman (1977), e tal estratégia leva a benefícios no contexto de classificação. Assim, com base na medida de intermediação proposta por Bangcharoensap *et al.* (2016), uma nova matriz de pesos normalizada é definida como dada na Equação 3.10, o que deu origem à função objetivo normalizada pelas arestas definida na Equação 3.12,

onde $\alpha_i$ e $\lambda_{ij}$ são parâmetros que controlam a influência dos termos, em que são dados por um processo definido pelos autores (BANGCHAROENSAP *et al.*, 2016).

$$\overline{W}_{ij,pq} = \frac{1}{C(e_{ij,pq})} W_{ij,pq} \tag{3.10}$$

$$J(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \sum_{t=1}^{m} E(\mathbf{F}_t) \tag{3.11}$$

$$E(\mathbf{F}_i^{(k)}) = \alpha_i \sum_{p=1}^{n_i} (f_{ip}^{(k)} - y_{ip}^{(k)})^2 + \sum_{j=1}^{m} \lambda_{ij} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \overline{W}_{ij,pq} (f_{ip}^{(k)} - f_{jq}^{(k)})^2 \tag{3.12}$$

Assim como no método GNetMine, utiliza-se o método dos gradientes descendentes na dedução do Algoritmo 2, derivando a Equação 3.12 em relação às variáveis e igualando este resultado a zero chega-se por meio de manipulações algébricas ao algoritmo iterativo para a solução que otimiza a função de custo. Além disso, da mesma forma o mesmo algoritmo pode ser obtido pela propagação entre vértices vizinhos agora ponderada pelo grau de intermediação das arestas.

**Algoritmo 2** – Algoritmo baseado em intermediação de arestas

**para** $\forall k \in \{1, \ldots, K\}, \forall i \in \{1, \ldots, m\}$ **faça**
    Inicialize o estimador de confiança $\mathbf{f}_i^{(k)}(0) = \mathbf{y}_i^{(k)}$;
**fim para**
**para** $\forall i, j,$ tal que $\mathbf{W}_{ij} \in W$ **faça**
    Normalize a matriz de pesos $\overline{W}_{ij,pq} = \frac{1}{C(e_{ij,pq})} W_{ij,pq}$
    Inicialize $\mathbf{D}_{ij} = \mathbf{O}^{(n_i, n_j)}$
**fim para**
**para** $\forall e_{ij,pq} \in E$ **faça**
    $\mathbf{D}_{ij,pp} = \mathbf{D}_{ij,pp} + \mathbf{W}_{ij,pq}$
**fim para**
**para** $\forall i \in \{1, \ldots, m\}$ **faça**
    $\mathbf{M}_i = \sum_{j=1, j \neq i}^{m} \lambda_{ij} \mathbf{D}_{ij} + \alpha_i \mathbf{I} + 2\lambda_{ii} \mathbf{D}_{ii}$
**fim para**
$t = 0$
**enquanto** não atingir o critério de convergência **faça**
    **para** $k \in \{1, \ldots, K\}$ **faça**
        **para** $i \in \{1, \ldots, m\}$ **faça**
            $\mathbf{f}_i^{(k)}(t+1) = \mathbf{M}_i^{-1} (\sum_{j=1, j \neq i}^{m} \lambda_{ij} \mathbf{W}_{ij} \mathbf{f}_j^{(k)}(t) + \alpha_i \mathbf{y}_i^{(k)} + 2\lambda_{ii} \mathbf{W}_{ii} \mathbf{f}_i^{(k)}(t))$
            $t = t + 1$
        **fim para**
    **fim para**
**fim enquanto**
**para** $\forall i \in \{1, \ldots, m\}$ **faça**
    Atribua $p-$ésimo objeto de tipo $X_i$ como da classe
    $c_{ip} = \arg\max_{1 \leq k \leq K} f_{ip}^{(k)}$
**fim para**

## 3.5 HetPathMine

Diferentemente dos métodos anteriores, o método HetPathMine (LUO *et al.*, 2014) opera a classificação sobre apenas um tipo de elemento dentro da rede heterogênea, o qual é chamado de tipo alvo. Esse método não utiliza a estrutura da HIN para efetuar a propagação dos rótulos em objetos que não são do tipo alvo. Ao invés disso, ele constrói uma rede homogêneas com base na combinação de meta-caminhos cíclicos, em que a função objetivo é computada com base nesta nova rede. Para a construção do método os autores se baseiam em duas suposições:

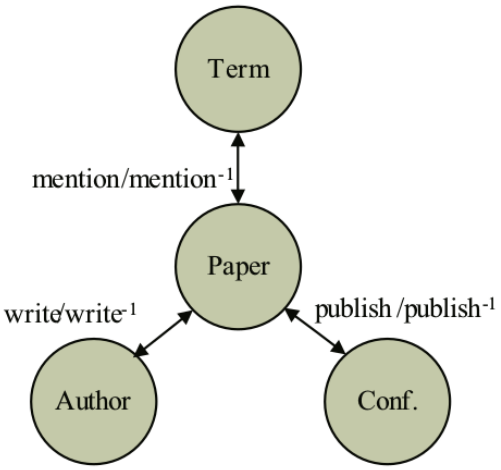
1. Uma rede resultante de um processo social normalmente possui muitos fatores de influência. De tal forma que objetos conexos possuem uma possibilidade maior de ter a mesma classe;
2. A classificação final deve ser próxima para elementos pré rotulados.

A primeira suposição suporta o uso de meta-caminhos na propagação de rótulos, substituindo relações diretas.

O HetPathMine também utiliza a matriz indicadora de classes $\mathbf{F}_t$, porém, estima apenas os valores para o tipo alvo. E para cada objeto $X_{ti}$ do tipo alvo, a classe é dada pelo máximo valor da linha $i$ em $\mathbf{F}_t$.

Na construção do método, são definidos $P$ meta-caminhos cíclicos, em que com o primeiro e o último tipo de objeto são o do tipo alvo. Por exemplo, uma rede que instancia a estrutura mostrada na Figura 14, em que o tipo alvo são autores, um meta-caminho cíclico possível é: ( autor $\xrightarrow{\text{escreveu}}$ artigo $\xrightarrow{\text{contém}}$ termo $\xrightarrow{\text{contido em}}$ artigo $\xrightarrow{\text{escrito por}}$ autor). Luo *et al.* (2014) defendem o uso de meta-caminhos, pois estes introduzem informações semânticas ao método, uma vez que estes foram construídos para representar relações significativas da HIN.

Figura 14 – Exemplo de uma estrutura de rede heterogênea baseada em dados bibliográficos.




Fonte: Luo *et al.* (2014).

Para cada meta-caminho $p \in 1, 2, \ldots, P$, uma matriz de similaridade $W^p \in \mathbb{R}^{n_t \times n_t}$ é construída, onde $n_t$ é o número de objetos do tipo alvo. A entrada $(i, j)$ da matriz de similaridade nos dá o número de

meta-caminhos do tipo $p$, que ligam o objeto $i$ ao objeto $j$. Tal matriz pode ser obtida de maneira direta pela multiplicação das matrizes de peso binárias **W**'s.

O conjunto de meta caminhos são então combinados por pesos $B = \{\beta_1, \beta_2, \ldots, \beta_P\}$, obtido pela otimização não restrita da função de custo dada na Equação 3.13, onde $n_l$ é o número de elementos rotulados do tipo alvo, $R$ é uma matriz de relação binária definida na Equação 3.14, e $s_p^{Pathsim}$ é a medida de similaridade Pathsim para o meta caminho $p$ (SUN *et al.*, 2011).

$$G(B) = \left\| \sum_{i=1}^{n_l} \sum_{j=1}^{n_l} \left( R(v_{t,i}, v_{t,j}) - \sum_{p=1}^{P} \beta_p s_p^{Pathsim}(v_{t,i}, v_{t,j}) \right) \right\| + \mu \left\| \sum_{p=1}^{P} \beta_p \right\|$$

$$R(v_{t,i}, v_{t,j}) = \begin{cases} 1 & \text{se os vértices } i \text{ e } j \text{ possuem o mesmo rótulo de classe;} \\ 0 & \text{caso contrário.} \end{cases} \tag{3.13}$$

Com base nestes elementos e nas suposições descritas, chega-se à seguinte função de custo:

$$\begin{aligned} J(F) &= \frac{1}{2} \sum_{p=1}^{P} \beta_p \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} W_{i,j}^p \left\| \frac{F_{ti}}{\sqrt{D_{ii}^p}} - \frac{F_{tj}}{\sqrt{D_{jj}^p}} \right\|^2 \\ &+ \lambda \sum_{i=0}^{n_t} \|F_{ti} - Y_{ti}\|^2, \end{aligned}$$

onde $D_{ii}^p$ é a soma da $i-$ésima linha da matriz $W^p$, que tem a função de reduzir a influência dos chamados *hubs* ou nós que possuem muitas ligações no grafo; e $beta_p$ são pesos atribuídos a cada um dos meta-caminhos, que são obtidos através de um processo de otimização descrito em (LUO *et al.*, 2014).

De forma análoga ao processo feito para técnicas de propagação geral, os algoritmos iterativos são deduzidos pelo método de gradientes descendentes. Porém, neste caso as variáveis envolvidas na modelagem do método são apenas os vetores de confiança dos vértices do tipo alvo. Para a dedução do algoritmo de propagação mostrado no Algoritmo 3, a função de custo é derivada com base nos vetores de confiança, considerando os pesos de meta-cominhos com escalares fixos.

**Algoritmo 3** – Algoritmo HetPathMine

**para** $\forall i \in \{1,\ldots,n_t\}$ do tipo alvo $t$ **faça**
    Inicialize o estimador de confiança $\mathbf{f}_{ti}(0) = \mathbf{y}_{ti}$;
**fim para**
**para** $\forall p \in \{1,\ldots,P\}$ **faça**
    Calcule a matriz de similaridade $W^{(p)}$ para o meta-caminho $p$;
**fim para**
$\tau = 0$
$t = 0$
**enquanto** não atingir o critério de convergência **faça**
    $\mathbf{F}_t(\tau+1) = (\sum_{p=1}^{P} \beta_p \mathbf{S}^p)\mathbf{F}_t(\tau) + \mu(\mathbf{F}_t(\tau) - \mathbf{Y}_t)$, com $\mathbf{S}^p = (\mathbf{D}^k)^{-\frac{1}{2}}\mathbf{W}^k(\mathbf{D}^k)^{-\frac{1}{2}}$
    $\tau = \tau + 1$
**fim enquanto**
**para** $\forall i \in \{1,\ldots,n_t\}$ **faça**
    Atribua $i-$ésimo objeto de tipo $X_t$ como da classe
    $c_{ti} = \arg\max_{1 \le k \le K} f_{ti}^{(k)}$
**fim para**

## 3.6 HeteClass

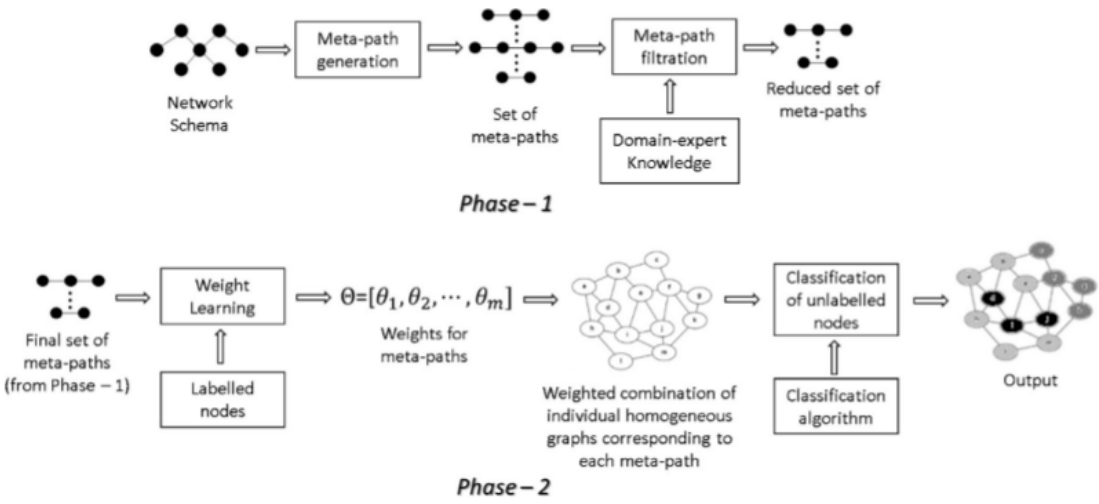
Assim como o HetPathMine, o método HeteClass, proposto por Gupta, Kumar e Bhasker (2017), é baseado em meta caminhos, utilizando estes para capturar a estrutura da HIN no processo de classificação. Além disso, os autores buscam estabelecer um modelo geral para a construção de métodos de classificação transdutiva em HIN baseados em meta caminhos.

Tal modelo geral se baseia em duas etapas principais, as quais são ilustradas na Figura 15. Na primeira etapa, diversos meta-caminhos são construídos automaticamente com base na HIN $\mathscr{G}$. Essa construção é efetuada respeitando algumas poucas restrições como o tamanho de caminho e a simetria, a qual é imposta devido à forma como a métrica de relevância DPRel é calculada (para mais detalhes consulte (GUPTA; KUMAR; BHASKER, 2017)). Em seguida, os meta-caminhos relevantes são selecionados, utilizando o conhecimento de especialistas da área, com o intuito de avaliar a semântica e relevância das relações presentes no meta-caminho.

Na segunda etapa, os pesos $\Theta = \theta_1, \theta_2, \ldots, \theta_P$ são otimizados para cada mate-caminho selecionado. Em seguida, todos meta-caminhos são combinados a partir dos pesos otimizados para resultando em uma nova estrutura de rede, a qual será utilizada para realizar a classificação. O processo de otimização de pesos é feito com base na função de custo dada na Equação 3.14, onde $V_t^l$ é o conjunto de elementos rotulados do tipo alvo $t$, $Sign(v_i, v_j)$ é uma função indicadora definida na Equação 3.14, e a função $Sim_{P_k}(v_i, v_j)$ baseada na métrica de relevância DPRel (GUPTA; KUMAR; BHASKER, 2015), que calcula a relação entre dois elementos.

$$L(\Theta) = \frac{1}{2} \sum_{v_i, v_j \in V_t^l, i \neq i} \left\| 1 - Sign(v_i, v_j) \sum_{k=1}^{K} \theta_k Sim_{P_k}(v_i, v_j) \right\|_2^2 + \frac{\lambda}{2} \|\Theta\|_2^2, \text{sujeito a } \theta_k \geq 0, \forall k = 1, \ldots, K$$

Figura 15 – Ilustração do fluxograma do modelo geral do método HetClass.




]

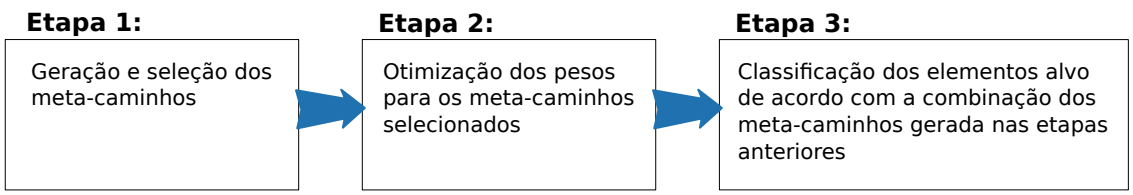
Fonte: Gupta, Kumar e Bhasker (2017).

$$Sign(v_i, v_j) = \begin{cases} 1 & \text{se os vértices } i \text{ e } j \text{ possuem o mesmo rótulo de classe, com } i \neq j; \\ -1 & \text{caso contrário.} \end{cases} \tag{3.14}$$

Após a obtenção dos pesos, o conjunto de meta-caminhos são combinados como mostrado na Equação 3.15, dando origem a uma rede homogênea sobre a qual é realizada a classificação transdutiva. De acordo com Gupta, Kumar e Bhasker (2017), a classificação foi realizada com a utilização do método PageRank Personalizado (HAVELIWALA, 2002). No entanto, os autores mencionam que a etapa de classificação pode ser realizada por qualquer método de classificação transdutiva para rede homogêneas.

$$Sim_{P_1,\ldots,P_K} = \frac{1}{\|\Theta\|_1} \sum_{k=1}^{K} \theta_k Sim_{P_k} \tag{3.15}$$

Figura 16 – Fluxograma que ilustra um modelo geral para a construção de técnicas de classificação transdutivas em HIN baseadas em meta-caminhos.




Fonte: Elaborada pelo autor.

No trabalho desenvolvido por Gupta, Kumar e Bhasker (2017), é possível detectar uma estrutura ainda mais geral para a construção de métodos de classificação transdutiva em HIN baseadas em meta-

caminhos. A Figura 16 ilustra o fluxograma dessa estrutura, a qual é composta por três etapas essenciais: construção e seleção dos meta-caminhos; combinação dos meta-caminhos por meio de pesos aprendidos com uso de medidas de relevância; e classificação transdutiva do tipo alvo com base na rede gerada. Em qualquer uma das etapas, pode-se intercalar diferentes métodos para as tarefas. Na primeira etapa, diferentes técnicas de escolha e seleção de meta-caminhos podem ser usadas, assim como, na segunda etapa, diferentes métricas de relevância. Por fim, o método de classificação pode ser escolhido de acordo com a aplicação final.

## 3.7 Considerações finais

Neste capítulo, são apresentados os principais fundamentos de aprendizado de maquina utilizado ao longo do desenvolvimento desta tese. Em especial, apresentamos uma breve contextualização do aprendizado transdutivo dentro da área, e motivações para seus uso quando o conjunto de amostras pré-rotuladas não são suficientes para a generalização base de técnicas supervisionadas. Em seguida, são apresentados os principais conceitos e modelagem das técnicas de classificação transdutiva com modelagem gráfica.

Finalmente, são apresentadas diferentes técnicas relevantes dentre da área, que mostram diferentes modelagens para métodos de classificação transdutiva sobre HIN, utilizando tanto modelagens acoplando métricas baseadas na topologia da rede quanto utilizando construções sobre o esquema. Apesar de os métodos apresentados serem bastante relevantes na área, estes exploram apenas algumas das possibilidades e certamente existe ainda um vasto horizonte de estudo. Neste contexto, esse trabalho visa explorar uma nova modelagem baseada na divergência de KL. No Capítulo 4 são apresentados os estudos e desenvolvimentos para a proposta desta nova técnica.

CAPÍTULO

4

# TCHN UM NOVO MÉTODO PARA A CLASSIFICAÇÃO TRANSDUTIVA EM HIN

A área de pesquisa em redes heterogêneas de informação é uma área bastante recente, tendo grande parte de seus avanços sido desenvolvidos na década atual. O próprio conceito de redes heterogêneas de informação foi apresentado por Sun *et al.* (2009), e, desde então, essa área vem crescendo cada vez mais e ganhando atenção em diversas tarefas de aprendizado de máquina, dentre elas, classificação de dados. Apesar do crescente avanço, quando comparadas com as pesquisas existentes em redes homogêneas ou mesmo redes bipartidas que são um caso particular de redes heterogêneas, existem ainda diversas abordagens inexploradas.

Neste trabalho foi o desenvolvido o algoritmo TCHN [1], que dentre outras motivações, busca atender tal demanda de novas abordagens dentro de sua área. Tal método se baseia na divergência de Kullback-Leibler (KL) com medida de similaridade dentre os vetores de informação. O uso desta medida é inspirado pelo trabalho de Faleiros, Rossi e Lopes (2017), onde se argumenta que tal medida é mais adequada para e modelagem da função de custo que busca regularizar a distribuição da informação nos vetores de informação quando esta se assemelha a uma distribuição estatística, como discutido na Seção 3.2.

O método TCHN é um método de classificação transdutiva em redes heterogêneas de informação, que generaliza o método TPBG [2] (FALEIROS; ROSSI; LOPES, 2017) que se dedica a redes bipartida. Dessa forma, este capítulo descreve o método TCHN proposto, abordando alguns dos principais aspectos da modelagem de métodos transdutivos de classificação, o que inclui estratégias de propagação e modelagem da função objetivo.

[1] baseado na sigla do inglês para *Transductive Classification in Heterogeneous Information Network*.
[2] baseado na sigla do inglês para *Transductive Propagation in Bipartite Graph*.

## 4.1 Notação

Nesta seção, a notação utilizada para o desenvolvimento do método TCHN é apresentada detalhadamente. Apesar de muitos dos elementos utilizados já terem sido apresentados no Capítulo 3, a Tabela 1 mostra os principais elementos do método para facilitar a compreensão do mesmo.

Tabela 1 – Notação adotada para os principais elementos utilizados no método TCHN proposto.

| Notação | Descrição |
|---|---|
| $m$ | número de tipos de elementos na base de dados |
| $n_i$ | número de elementos do tipo $i$ |
| $X_i = \{x_{i1}, \ldots, x_{in_i}\}$ | conjunto dos elementos do tipo $i$ |
| $X = \cup_{i=1}^{m} X_i$ | todos os elementos |
| $V_i = \{v_{i1}, \ldots, v_{in_i}\}$ | vértices associados aos elementos do tipo $i$ |
| $V = \cup_{i=1}^{m} V_i$ | todos os vértices |
| $E = \{e_{ij,pq}\}$ | conjunto das arestas do grafo, onde $i, j = \{1, \ldots, m\}$ são tipos de elementos, $p = \{1, \ldots, n_i\}$ e $q = \{1, \ldots, n_j\}$ |
| $e_{ij,pq}$ | aresta que liga o vértice $v_{ip}$ ao vértice $v_{jq}$ |
| $W_{ij,pq}$ | valor não negativo associado ao vértice $e_{ij,pq}$ |
| $\mathscr{G} = (V, E, W)$ | grafo heterogêneo que representa os elementos e relações na base de dados |
| $\mathbb{C} = \{c_1, c_2, \ldots, c_K\}$ | conjunto das $K$ possíveis classes |
| $V^l \subset V$ | conjunto dos vértices associados a elementos pré rotulados da base de dados |
| $V^u \subset V$ | conjunto dos vértices associados a elementos não rotulados |
| $V = V^l \cup V^u$ | |
| $Y_i = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in_i}\}$ | conjunto de vetores $l$ dimensionais, onde para cada elementos $v_{ip} \in V^l$ rotulado como da classe $c_k$, o valor $k-$ésima entrada de $Y_ip$ é igual a 1 e as outras entradas são iguais a 0, para os elementos em $V^u$ todos os valores são nulos, porém não são relevantes pois não entram no equacionamento |
| $\mathbf{f}_{ip}$ | vetor de informação do vértice $v_{ip}$ |
| $\mathbf{c}_{ij,pq}$ | vetor de informação da aresta $e_{ij,pq}$ |
| $T = \{i_1, i_2, \ldots, i_t\}$ | conjuntos de tipo de elementos |

Fonte: Elaborada pelo autor.

## 4.2 Modelagem matemática

Como discutido no Capítulo 3, os métodos de classificação transdutiva baseiam-se nas premissas de consistência e são modelados com base no modelo matemático apresentado na Seção 3.2, de acordo com a Equação 3.5. Tal modelo depende da escolha de uma medida para sua aplicação. Tipicamente, a medida Euclideana é adotada para estabelecer a relação entre os vetores de informação de classe e os elementos da rede.

Neste trabalho, também foi utilizado o mesmo modelo para o equacionamento matemático na construção do algoritmo proposto. Porém, na escolha da função de medida, foi utilizada a divergência de Kullback-Leibler (KL), baseando esta escolha nos resultados apresentados por Faleiros, Rossi e Lopes (2017) para redes bipartidas. A medida de divergência KL, dada na Equação 4.1, tem como objetivo quantificar o quão bem a distribuição de probabilidades $p$ se aproxima da distribuição $q$. A ideia de que os vetores de informação de classe possuem uma distribuição similar à uma distribuição de probabilidades, faz com que elementos com mesma classe tendam a possuir baixa divergência. Isso motivou o uso da medida de divergência de Kullback-Leibler (KL), um vez que ela apresenta melhores características nesse contexto.

$$D_{KL}(p\|q) = \sum_i p_i log_2 \frac{p_i}{q_i} \tag{4.1}$$

Além da utilização de uma medida distinta da usualmente encontrada na literatura (FALEIROS; ROSSI; LOPES, 2017), um vetor de informação é associado a cada aresta da rede, com o objetivo de contornar possíveis divergência semânticas de elementos, os quais podem possuir mais de uma classe com significado distinto. Por exemplo, em conjuntos de dados textuais bibliográficos, que são utilizados neste trabalho, é comum se encontrar como um dos tipos de entidades os termos que ocorrem no texto, os quais podem estar em mais de uma classe, assumindo diferentes significados. Especificamente, dado um exemplo mais concreto, é possível encontrar o termo "rosa"que se relacionar tanto a "flor"ou a "cor", as quais possuem significados distintos.

O vetor de informação associado às arestas garante que um mesmo vértice da rede seja capaz de propagar diferentes informações de classes a seus vizinhos (FALEIROS; ROSSI; LOPES, 2017). Assim, o método TCHN associa vetores de informação a cada um dos vértices $\mathbf{f}_{ip}$ e arestas $\mathbf{c}_{ij,pq}$ da rede. Em seguida, busca-se minimizar a divergência entre os vetores resultante do produto Hadamard $\mathbf{f}_{ip} \odot \mathbf{f}_{jq}$ e o vetor $\mathbf{c}_{ij,pq}$. De tal forma que, quanto maior o peso da aresta $W_{ij,pq}$, que liga dois vértices da rede $v_{i,p}$ e $v_{j,q}$, maior deve ser a concordância entre as informações contidas nos vetores.

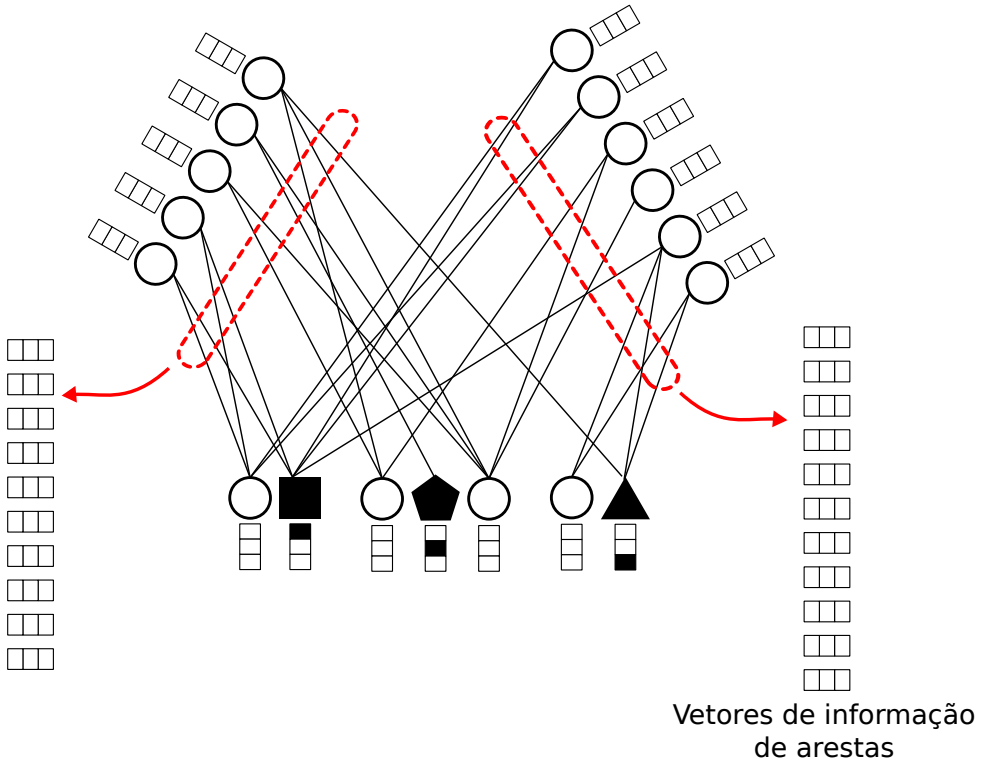
Assim, a Equação 4.2 é obtida utilizando a medida de divergência KL para medir a similaridade entre os vetores. A Figura 17 mostra um exemplo simples de rede heterogênea, onde os vetores de informação são apontados junto aos elementos da rede, os quais são associados pelo modelo.

$$D_{KL}(W_{ij,pq}\mathbf{c}_{ij.pq}\|\mathbf{f}_{ip} \odot \mathbf{f}_j q) = \sum_{k=1}^{K} W_{ij,pq}\mathbf{c}_{ij,pq}^{(k)} \log_2 \frac{W_{ij,pq}\mathbf{c}_{ij,pq}^{(k)}}{\mathbf{f}_{ip}^{(k)} \odot \mathbf{f}_{jq}^{(k)}} \tag{4.2}$$

A otimização se dá pela minimização da divergência KL, que, por simplicidade, é transformado em um problema de maximização pela seguinte manipulação algébrica:

$$\begin{aligned} minD_{KL}(W_{ij,pq}\mathbf{c}_{ij,pq}\|\mathbf{f}_{ip} \odot \mathbf{f}_j q) &= -maxD_{KL}(W_{ij,pq}\mathbf{c}_{ij,pq}\|\mathbf{f}_{ip} \odot \mathbf{f}_j q) \\ &= max \sum_{k=1}^{K} W_{ij,pq}\mathbf{c}_{ij,pq}^{(k)} \log_2 \frac{\mathbf{f}_{ip}^{(k)} \odot \mathbf{f}_{jq}^{(k)}}{W_{ij,pq}\mathbf{c}_{ij,pq}^{(k)}}. \end{aligned}$$

Figura 17 – Representação de uma rede heterogênea com os vértices, arestas e os rótulos multidimensionais associados de acordo com a modelagem do método TCHN.




Fonte: Elaborada pelo autor.

Com base nesta construção, tem-se então, na Equação 4.3, o primeiro termo do modelo de regularização adotado, onde $R$ é o termo regularizador definido na Equação 4.4, o qual é aplicado sobre cada vértice $v_i$; $\alpha_i$ controla a concentração no vetor informações $\mathbf{f}_{ip}$, que pode ser definido para cada tipo $i$; e $T_A$ é um subconjunto de tipos onde a regularização é desejada.

$$Q_{\mathscr{G}}(\mathbf{F},\mathbf{C})^{(I)} = \sum_{e_{ij,pq}\in E}\left(W_{ij,pq}\mathbf{c}^{(k)}_{ij,pq}\log_2\frac{\mathbf{f}^{(k)}_{ip}\odot\mathbf{f}^{(k)}_{jq}}{\mathbf{c}^{(k)}_{ij,pq}}\right)+\sum_{v_{ip}\in V_i|i\in T_A}R(\mathbf{f}_{ip},\alpha_i) \tag{4.3}$$

$$R(\mathbf{f}_{ip},\alpha_i) = (\alpha_i-\mathbf{f}_{ip})\log\mathbf{f}_{ip}+\mathbf{f}_{ip}(\log\mathbf{f}_{ip}-1) \tag{4.4}$$

O segundo termo do modelo de regularização, conhecido como o termo de ajuste, busca a similaridade entre os valores reais e a classificação para os termos pré rotulado. Este termo é construído de maneira análoga ao primeiro, buscando minimizar a divergência entre os vetores informação $\mathbf{f}_{ip}$ e o de rótulo pré conhecido $\mathbf{y}_{ip}$. Este também utiliza a medida de divergência KL, chegando, assim, no termo dado na Equação 4.5.

$$Q_{\mathscr{G}}(\mathbf{F},\mathbf{C})^{(II)} = \sum_{v_{ip}\in V^l}\mathbf{y}_{ip}\log\frac{\mathbf{f}_{ip}}{\mathbf{y}_{ip}} \tag{4.5}$$

Combinando os dois termos construídos, tem-se a função de custo a ser maximizada, como definida na Equação 4.6.

$$Q_{\mathscr{G}}(\mathbf{F},\mathbf{C}) = \sum_{e_{ij,pq}\in E} W_{ij,pq}\mathbf{c}_{ij,pq}\log\frac{\mathbf{f}_{ip}\bigodot\mathbf{f}_{jq}}{\mathbf{c}_{ij,pq}} + \sum_{v_{ip}\in V_i|i\in A} R(\mathbf{f}_{ip},\alpha_i) + \sum_{v_{ip}\in V^l}\mathbf{y}_{ip}\log\frac{\mathbf{f}_{ip}}{\mathbf{y}_{ip}} \tag{4.6}$$

Em muitos casos, os tipos de elementos são separados em tipos alvo $T_A$ e tipos secundários $T_S$, mesmo em métodos que não se baseiam em meta-caminhos, onde todos os elementos da rede são usados na propagação. No trabalho de Faleiros, Rossi e Lopes (2017), redes bipartidas são focadas na classificação de dados textuais, em que apenas dois tipos de elementos são usados: documentos e termos. Neste caso, os documentos são considerados o tipo alvo, enquanto termos são do tipo secundário. Estes dois termos são tratados de forma distinta ao longo da modelagem do algoritmo, quando referentes às regularizações impostas a seus elementos associados. De forma análoga, para a dedução da fórmula iterativa para o algoritmo proposto nesta tese, os vértices são tratados de formas distinta de acordo com seu tipo.

A seguir são apresentadas as deduções utilizadas para a otimização da função de custo de acordo com cada vetor de informação associado aos vértices e arestas.

- Dedução da equação de atualização para o vetor de informação da aresta $\mathbf{c}_{\zeta\xi,\rho\sigma}$:

  Isolamos os termos contendo $\mathbf{c}_{\zeta\xi,\rho\sigma}$ e adicionamos o multiplicador de Lagrange apropriado, resultados em:

$$Q_{\mathscr{G}[\mathbf{c}_{\zeta\xi,\rho\sigma}]} = \left(W_{\zeta\xi,\rho\sigma}\mathbf{c}_{\zeta\xi,\rho\sigma}\log\frac{\mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma}}{\mathbf{c}_{\zeta\xi,\rho\sigma}}\right) + \lambda\left(\sum_{k=1}^{K}[\mathbf{c}_{\zeta\xi,\rho\sigma}]^{(k)} - 1\right). \tag{4.7}$$

  Impondo $\sum_{k=1}^{K}[\mathbf{c}_{\zeta\xi,\rho\sigma}]^{(k)} = 1$ podemos eliminar o multiplicador, obtendo, assim:

$$\begin{aligned}\frac{\partial}{\partial\mathbf{c}_{\zeta\xi,\rho\sigma}}(Q(\mathbf{F},\mathbf{C})) &= \frac{\partial}{\partial\mathbf{c}_{\zeta\xi,\rho\sigma}}\left(W_{\zeta\xi,\rho\sigma}\mathbf{c}_{\zeta\xi,\rho\sigma}\log\frac{\mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma}}{C_{\zeta\xi,\rho\sigma}}\right)\\ &= \left(W_{\zeta\xi,\rho\sigma}\log\frac{\mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma}}{\mathbf{c}_{\zeta\xi,\rho\sigma}} + W_{\zeta\xi,\rho\sigma}\right)\\ &= W_{\zeta\xi,\rho\sigma}\left(\log\mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma} - \log\mathbf{c}_{\zeta\xi,\rho\sigma} + 1\right).\end{aligned}$$

  Logo temos que:

$$\frac{\partial}{\partial\mathbf{c}_{\zeta\xi,\rho\sigma}}(Q(\mathbf{F},\mathbf{C})) = W_{\zeta\xi,\rho\sigma}\left(\log\mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma} - \log\mathbf{c}_{\zeta\xi,\rho\sigma} + 1\right), \tag{4.8}$$

  sujeito à $\sum_{k=1}^{K}[\mathbf{c}_{\zeta\xi,\rho\sigma}]^{(k)} = 1$.

  Ajustando o resultado dessa derivada igual a zero, obtêm-se a equação de atualização dada na Equação 4.9.

$$\mathbf{c}_{\zeta\xi,\rho\sigma} \propto \mathbf{f}_{\zeta\rho}\bigodot\mathbf{f}_{\xi\sigma} \tag{4.9}$$

- Dedução da equação de atualização para o vetor de informação de vértice de tipo alvo $\mathbf{f}_{\zeta\rho}$ para $\zeta \in T_A$:

  Dado que todos os tipos em $T_A$ conterão a função de regularização, não adicionaremos aqui um multiplicador de Lagrange. Desta forma, isolando os termos contendo $\mathbf{f}_{\zeta\rho}$, e obtemos:

$$Q_{\mathscr{G}[\mathbf{f}_{\zeta\rho}]} = \sum_{e_{\zeta j,\rho q}\in\varepsilon} W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q}\log\mathbf{f}_{\zeta\rho} + R(\mathbf{f}_{\zeta\rho},\alpha_\zeta) + \left[\mathbf{y}_{\zeta\rho}\log\frac{\mathbf{f}_{\zeta\rho}}{\mathbf{y}_{\zeta\rho}}\right]_{v_{\zeta\rho}\in V^l}. \tag{4.10}$$

  Impondo $\mathbf{f}_{\zeta\rho} = \mathbf{y}_{\zeta\rho}$ para $v_{\zeta\rho} \in V^l$, podemos eliminar o últimos termo da equação. Obtemos, assim:

$$\begin{aligned}
\frac{\partial}{\partial\mathbf{f}_{\zeta\rho}}(Q(\mathbf{F},\mathbf{C})) &= \sum_{e_{\zeta j,\rho q}\in\varepsilon}\frac{\partial}{\partial\mathbf{f}_{\zeta\rho}}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q}\log\mathbf{f}_{\zeta\rho} + \frac{\partial}{\partial\mathbf{f}_{\zeta\rho}}R(\mathbf{f}_{\zeta\rho},\alpha_\zeta)\\
&= \frac{1}{\mathbf{f}_{\zeta\rho}}\sum_{e_{\zeta j,\rho q}\in\varepsilon}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q} + \left(\frac{\alpha_\zeta-\mathbf{f}_{\zeta\rho}}{\mathbf{f}_{\zeta\rho}} - \log\mathbf{f}_{\zeta\rho} + (\log\mathbf{f}_{\zeta\rho}-1) + \frac{\mathbf{f}_{\zeta\rho}}{\mathbf{f}_{\zeta\rho}}\right)\\
&= \frac{1}{\mathbf{f}_{\zeta\rho}}\left(\sum_{e_{\zeta j,\rho q}\in\varepsilon}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q} + \alpha_\zeta - \mathbf{f}_{\zeta\rho}\right).
\end{aligned}$$

  Logo temos que:

$$\frac{\partial}{\partial\mathbf{f}_{\zeta\rho}}(Q(\mathbf{F},\mathbf{C})) = \frac{1}{\mathbf{f}_{\zeta\rho}}\left(\sum_{e_{\zeta j,\rho q}\in\varepsilon}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q} + \alpha_\zeta - \mathbf{f}_{\zeta\rho}\right). \tag{4.11}$$

  Ajustando o resultado dessa derivada igual a zero, obtêm-se a equação de atualização dada na Equação 4.12.

$$\mathbf{f}_{\zeta\rho} = \alpha_\zeta + \sum_{e_{\zeta j,\rho q}\in E}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q} \tag{4.12}$$

- Dedução da equação de atualização para o vetor de informação de vértice de tipo secundário $\mathbf{f}_{\zeta\rho}$ para $\zeta \in T_S$:

  Dado que todos os tipos em $T_S$ não conterão a função de regularização, adicionaremos aqui um multiplicador de Lagrange. Desta forma isolando os termos contendo $\mathbf{f}_{\zeta\rho}$, e obtemos:

$$Q_{\mathscr{G}[\mathbf{f}_{\zeta\rho}]} = \sum_{e_{\zeta j,\rho q}\in E} W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q}\log\mathbf{f}_{\zeta\rho} + \left[\mathbf{y}_{\zeta\rho}\log\frac{\mathbf{f}_{\zeta\rho}}{\mathbf{y}_{\zeta\rho}}\right]_{v_{\zeta\rho}\in V^l} + \lambda\left(\sum_{v_{\zeta,p}\in V_\zeta}\mathbf{f}_{\zeta,p} - 1\right). \tag{4.13}$$

  Analogamente, derivando $Q_{\mathscr{G}[\mathbf{f}_{\zeta\rho}]}$, impondo $\sum_{v_{\zeta,p}\in V_\zeta}\mathbf{f}_{\zeta,p} = 1$ e ajustando o resultado dessa derivada igual a zero, obtêm-se a equação de atualização dada na Equação 4.14.

$$\mathbf{f}_{\zeta\rho} = \sum_{e_{\zeta j,\rho q}\in E}W_{\zeta j,\rho q}\mathbf{c}_{\zeta j,\rho q} \tag{4.14}$$

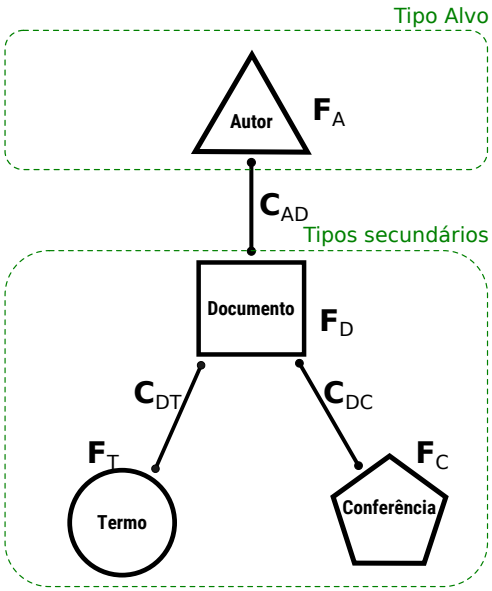
Ao final temos as seguintes equações de atualização para os vetores de informação associados aos elementos da rede, os quais serão utilizados na construção do algoritmo proposto:

- $\mathbf{c}_{\zeta\xi,\rho\sigma} \propto \mathbf{f}_{\zeta\rho} \odot \mathbf{f}_{\xi\sigma}$, sujeito a $\sum_{k=1}^{K} [\mathbf{c}^{(k)}_{\zeta\xi,\rho\sigma} = 1$;
- $\mathbf{f}_{\zeta\rho} = \alpha_{\zeta} + \sum_{e_{\zeta j,\rho q} \in E} W_{\zeta j,\rho q} \mathbf{c}_{\zeta j,\rho q}$ para $\zeta \in T_A$;
- $\mathbf{f}_{\zeta\rho} = \sum_{e_{\zeta j,\rho q} \in E} W_{\zeta j,\rho q} \mathbf{c}_{\zeta j,\rho q}$ para $\zeta \in T_S$, sujeito a $\sum_{v_{\zeta,p} \in V_{\zeta}} \mathbf{f}_{\zeta,p} = 1$.

## 4.3 Estratégias de propagação

Pela construção dos vetores de informação feita na Seção 4.2, cada vértice e aresta da rede associa-se aos vetores multidimensionais, em que equações de atualização otimizam a função objetivo para o modelo proposto. A Figura 18 ilustra, com o uso de um esquema de rede, os vetores associados a cada elemento da rede e suas distinções destacada.

Figura 18 – Exemplo dos elementos associados a uma HIN pelo modelo proposto, utilizando sua estrutura para ilustrar o vetores multidimensionais e a características dos elementos. Na figura, os conjuntos de tipos, que são considerados alvo ou secundários, são destacadas pelas caixas em verde pontilhadas. A cada tipo de aresta são destacadas as matrizes de informação $\mathbf{C}_{ij}$, e a cada tipo de vértice, as matrizes de informação $\mathbf{F}_i$.




Fonte: Elaborada pelo autor.

É importante notar que existe uma dependência cíclica nas equações de atualização, em que os vetores multidimensionais associados às arestas dependem dos vetores associados aos vértices, enquanto que os vetores multidimensionais associados aos vértices dependem dos vetores associados às arestas. Assim, para que o valor de um vértice seja atualizado, é necessário antes atualizar o valor nas aresta e vice versa.

Além disso, os vértices são divididos em dois grupos de acordo com seus tipos, os vértices alvo e os vértices secundários. Com isso, diferentes estratégias de propagação podem ser adotadas, considerando

tais dependências para atualização, assim como, prioridades nas atualizações de informação de acordo com o tipo de vértice.

Uma abordagem mais simples seria a iteração temporal dos elementos, partindo de uma inicialização dos elementos e tratando igualmente todos os vetores de informação, de tal forma que:

$$[\mathbf{F}(t+1), \mathbf{C}(t+1)] = F([\mathbf{F}(t), \mathbf{C}(t)]), \tag{4.15}$$

onde $F$ é uma função das equações de atualização dos elementos.

Outros trabalhos utilizam significados e características dos elementos para adotar estratégias de propagação, como é caso do método TPBG proposto por Faleiros, Rossi e Lopes (2017). Nessa proposta, os documentos, que são os elementos de tipo alvo, recebem precedência na propagação através de um processo chamado propagação local. Assim, os vetores de informação são utilizados para a atualização dos elementos de tipo alvo até se atingir um estado de convergência. Em seguida, após a atualização dos vetores de informação dos documentos, a propagação global é conduzida, em que esses vetores são usados para a atualização dos vetores de informação dos termos, que são do tipo secundários.

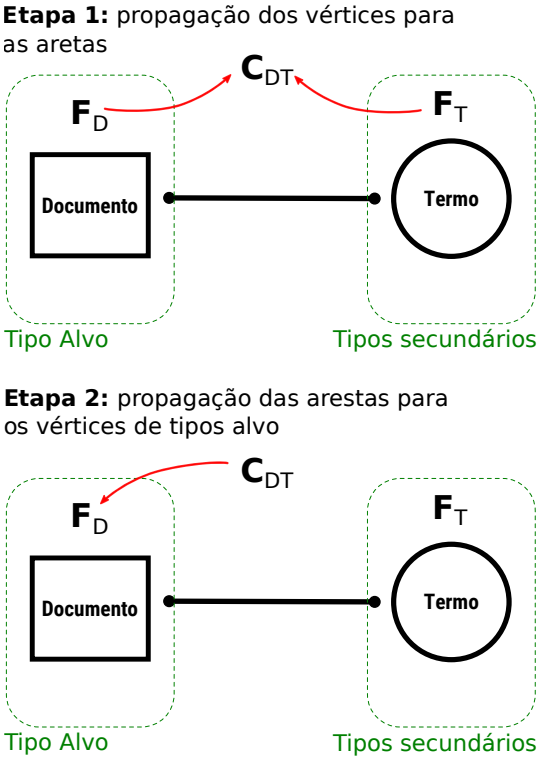
A Figura 19 busca ilustrar esquematicamente a estratégia de propagação adotada no método chamado TPBG (FALEIROS; ROSSI; LOPES, 2017), em que a Figura 19a ilustra a propagação local, enquanto a Figura 19b mostra a propagação global.

Para a construção do algoritmo proposto nesta tese, vamos estudar duas estratégias e analisar suas convergências e acurácias, tanto local, considerando a convergência dos vetores de informação para os elementos de tipo alvo, quanto para a convergência global, considerando todos os elementos da rede. Em ambos os caso, consideraremos também o número de operações de propagação para vértices e arestas, que como será visto, varia de acordo com cada estratégia adotada.

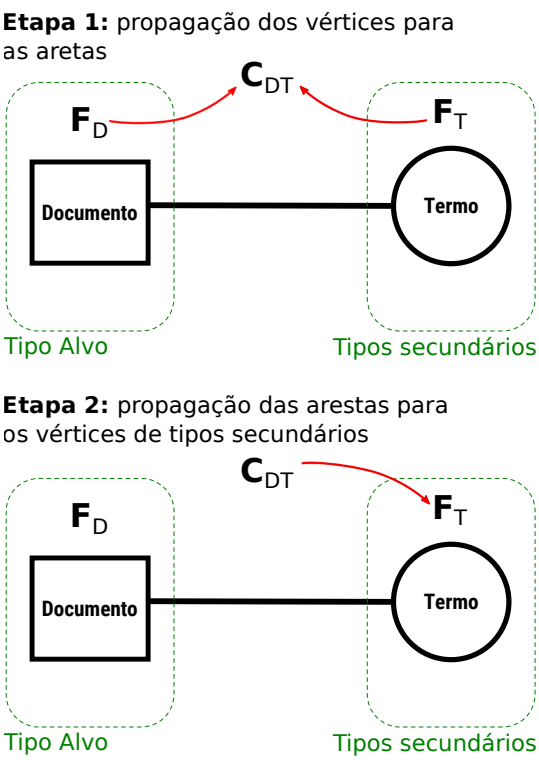
Na primeira estratégia apresentada, todos os elementos são tratados de forma uniforme, ou seja, todos os vértices são atualizados ao mesmo tempo. Sendo assim, a propagação se divide em dois passos: propagação de informação dos vértices para as arestas; e das arestas para os vértices. Para tal estudo, utilizamos um rede sintética gerada com o uso da ferramenta HNOC, desenvolvida ao longo deste projeto de doutorado e descrita no Capítulo 5. Tal rede foi gerada com base no esquema mostrado na Figura 18, contendo 200 autores, 1000 documentos, 20 conferências e 1000 termos. Esta foi construída contendo 4 comunidades distribuídas uniformemente para todos os tipos de vértices, e o tipo 'autor' recebeu rótulos para estas 4 comunidades em todos os elementos, os quais foram utilizados em proporções de 4% e 8% para os experimentos desta seção.

Dada a dependência cíclica de atualização entre os vetores de informação de vértices e arestas, apenas um dos conjuntos precisa ser inicializado, uma vez que os valores pré rotulados são diretamente colocado nos vértices. Sendo assim, partindo de um conjunto $\mathbf{F}(0)$ de $\mathbf{F}$'s, inicializados com os rótulos para o elementos pré-rotulados e aleatoriamente para os elementos não rotulados, a propagação se dá dos vértices para as arestas, das arestas para os vértices e assim sucessivamente até a convergência, como ilustrado na Figura 20.

Figura 19 – Ilustração esquemática da estratégia de propagação do método TPBG. Na figura as setas nas arestas indicam o sentido de propagação da informação. Na propagação local, Figura 19a, a informação é propagada de $\mathbf{F}_T$ e $\mathbf{F}_D$ para $\mathbf{C}_{DT}$ e, em seguida, de $\mathbf{C}_{DT}$ para $\mathbf{F}_D$. Já na propagação global, Figura 19b, a informação é propagada de $\mathbf{F}_T$ e $\mathbf{F}_D$ para $\mathbf{C}_{DT}$ e, em seguida, de $\mathbf{C}_{DT}$ para $\mathbf{F}_T$.



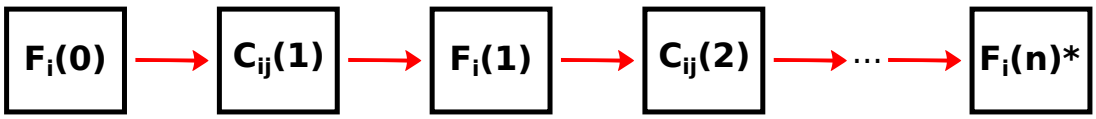

(a) Propagação local para o algoritmo TPBG.




(b) Propagação global para o algoritmo TPBG.

Fonte: Elaborada pelo autor.

Figura 20 – Estratégia de propagação, onde todos os vértices são atualizados ao mesmo tempo.

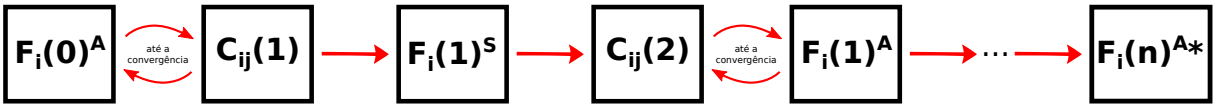



Fonte: Elaborada pelo autor.

Na segunda estratégia apresentada, os elementos são divididos entre tipos alvo e tipos secundário, os quais são tratados de formas distintas. Inicialmente, os vértices dos tipos alvo serão atualizados até que se atinja um convergência local, e, em sequência, esta informação é propagada para os outros elementos de tipos secundários. Sendo assim, a propagação se divide em duas etapas: a propagação de informação cíclica dos vértices de tipos alvo e arestas adjacentes até que se atinja uma convergência; e a propagação de informação simples das arestas para os elementos de tipo secundários. A Figura 21 ilustra essa estratégia

que foi desenvolvida para redes heterogêneas gerais, sendo esta uma extensão da estratégia adotada pelo algoritmo TPBG para redes bipartidas.

Figura 21 – Estratégia de propagação onde elementos são divididos entre tipos alvo e tipos secundário tratados de forma distinta.

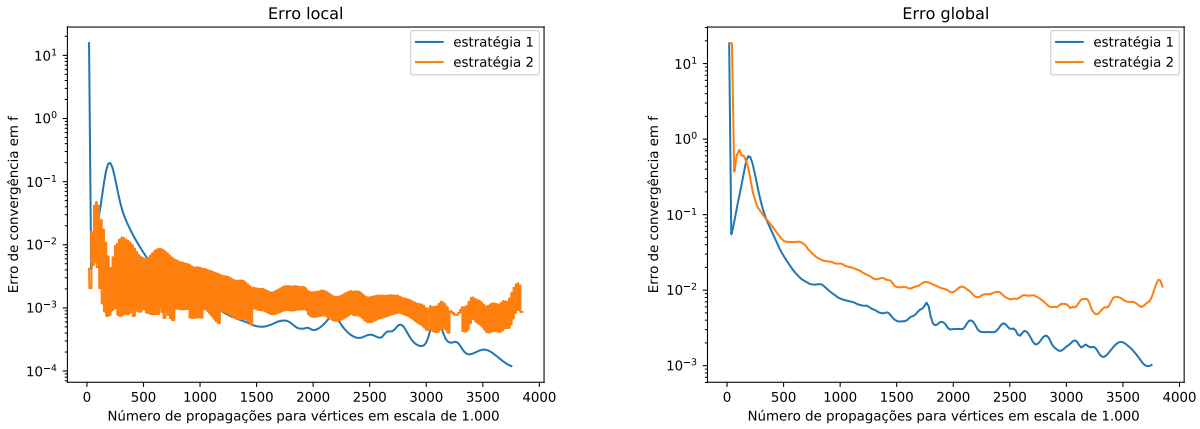



Fonte: Elaborada pelo autor.

Desta forma, considerando estas duas estratégias, obtivemos os resultados mostrados nas Figuras 22, 23 e 24. Tais resultados foram obtidos com uma inicialização unitária dos vetores de informação, e impondo critérios de parada em 200 iterações na propagação global e 30 iterações para a propagação local. Para os resultados de acurácia, cada estratégia foi rodada 50 vezes com seleção aleatória dos vértices pré-rotulados de forma uniforme.

As Figuras 22 e 23 mostram o erro de convergência local e global para as duas estratégias estudas. Tais resultados foram obtidos em uma rodada do método, considerando as imposições descritas acima. Como pode ser visto, ambas as estratégias se comportam de maneira semelhante quanto a convergência, tendo uma leve superioridade para a primeira estratégia. Apesar desses resultados serem obtidos de apenas uma rodada dos métodos, e terem objetivo apenas ilustrativo, este perfil de convergência foi observado em todas os experimentos deste projeto.

Figura 22 – Resultados do erro local e global para as duas possíveis estratégias de propagação estudadas, considerando um conjunto de dados com 4% dos autores pré-rotulados.

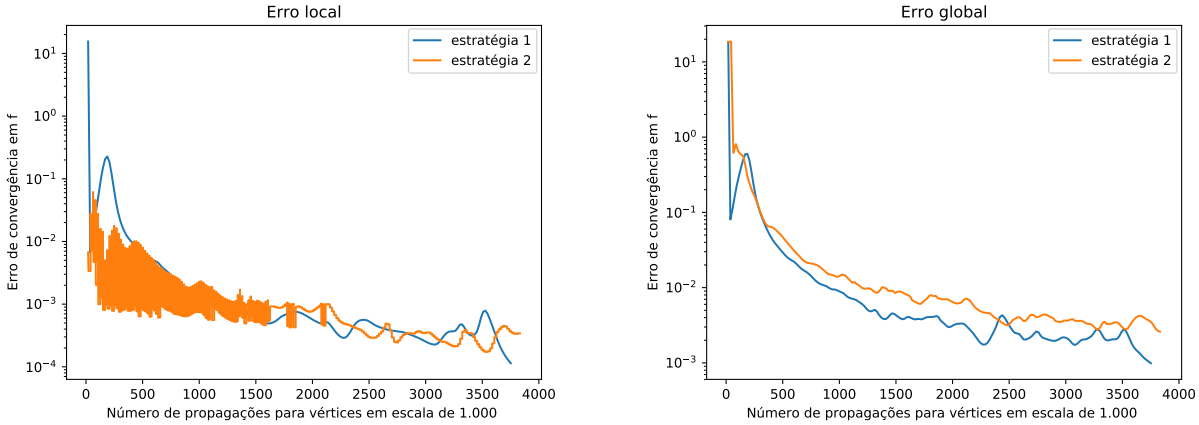



(a) Erro local.

(b) Erro global.

Fonte: Dados da pesquisa.

Figura 23 – Resultados do erro local e global para as duas possíveis estratégias de propagação estudadas, considerando um conjunto de dados com 8% dos autores pré-rotulados.



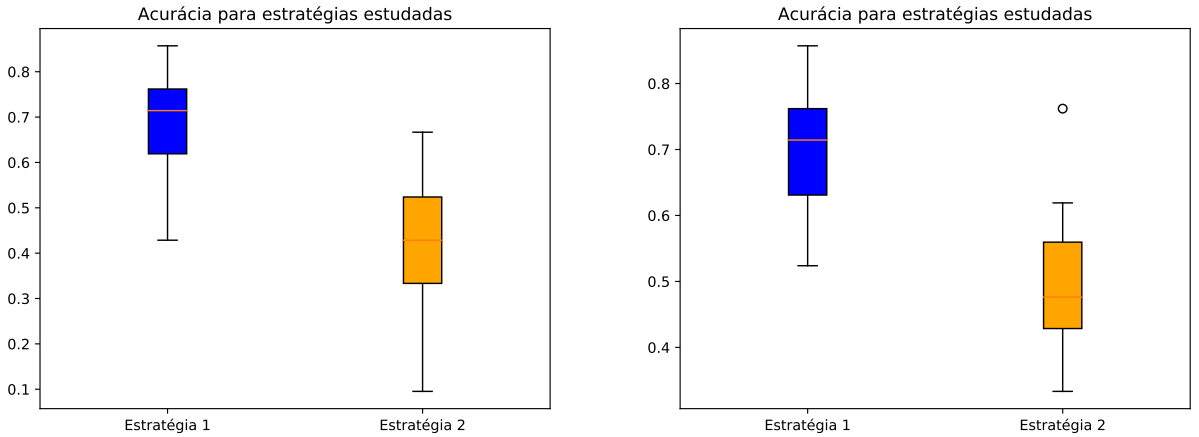

(a) Erro local.

(b) Erro global.

Fonte: Dados da pesquisa.

Do ponto de vista de convergência, ambas as estratégias demonstraram equivalência. Porém, como mostrado na Figura 24, as acurácias observadas para cada estratégia são muito distintas. É possível observar a superioridade da estratégia 1 em relação à estratégia 2, onde a acurácia em ambos os experimentos foi maior. Resultados similares foram observados para redes a partir de três camadas, e, por isso, optamos em adotar tal estratégia na proposta de algoritmo. Vale citar que, apesar da estratégia se mostrar superior em nossos experimentos, podem haver cenários onde a situação se inverta. No entanto, consideramos que um estudo mais aprofundado está fora do escopo desta tese.

Figura 24 – Resultados de acurácia para as duas possíveis estratégias de propagação estudadas, considerando um conjunto de dados com 4% e 8% dos autores pré-rotulados.




(a) Acurácia obtida considerando 4% dos autores pré-rotulados.

(b) Acurácia obtida considerando 8% dos autores pré-rotulados.

Fonte: Dados da pesquisa.

## 4.4 Inicialização

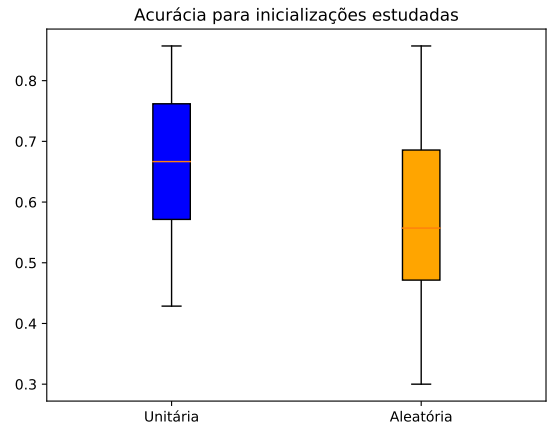
Outro ponto importante a ser avaliado na definição do algoritmo é o modo de inicialização dos vetores de informação. Como apresentado na Seção 4.2, os vetores de informação dos vértices e arestas, denotados por $\mathbf{F}_{i|i\in T}$ e $\mathbf{C}_{ij|i,j\in T}$, possuem uma dependência cíclica em suas atualizações. Além disso, dado que os elementos pré-rotulados são vértices, os vetores de informação associados a estes devem ser inicializados e são suficientes para que o algoritmo seja bem definido.

Nesse contexto, incluímos duas estratégias de inicialização, em que a primeira é a inicialização aleatória das entradas dos vetores de informação, enquanto, a segunda, é a inicialização nula dos vetores. Considerando a segunda abordagem que é a mais utilizada, os valores das entradas de $\mathbf{f}_{ip}$ seriam nulos, exceto para o vetores associados a vértices pré-rotulados, onde os valores são dados como $\mathbf{f}_{ip}^{(k)} = \{1 \text{ se } v_{ip} \text{ é da classe } i_k; 0 \text{ caso contrário}\}$.

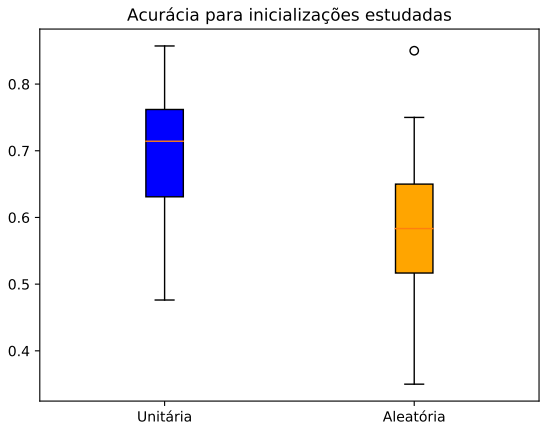
Porém, tal estratégia de inicialização não é valida para o método proposto, uma vez que este é um método multiplicativo. Isso pode ser observado nas fórmulas de atualização dos vetores de informação de arestas, em que sua atualização é a função da multiplicação dos valores de informação dos vértices. Assim, a inicialização nula das entradas levaria, em poucas iterações, à convergência a zero, não produzindo uma classificação.

Para evitar esse problema, adotamos uma forma não convencional de inicialização, o que leva em consideração a característica multiplicativa do método proposto. Sabendo que a inicialização nula é utilizada em métodos aditivos onde zero é o elemento neutro, o elemento neutro da multiplicação foi utilizado para inicialização, ou seja, o um. Neste caso, todos os valores das entradas dos vetores de informação são inicializados com o valor um exceto o vetores associados a vértices pré-rotulados.

Figura 25 – Resultados obtidos de acurácia para as duas possíveis inicializações dos vetores de informação para o método proposto, considerando um conjunto de dados com 4% e 8% dos autores pré-rotulados.




(a) Acurácia obtida considerando 4% dos autores pré-rotulados.




(b) Acurácia obtida considerando 8% dos autores pré-rotulados.

Fonte: Dados da pesquisa.

Resumindo, foram estudadas estas duas estratégias de inicialização, ou seja, a aleatória e a

unitária. Como na estratégia de análise empregada na Seção 4.3, a mesma rede sintética descrita foi usada para avaliar o desempenho das diferentes formas de inicialização dos vetores de informação. Nesses experimentos, a estratégia de propagação que demonstrou melhores resultados na Seção 4.3 foi aplicada. Assim, executou-se por 50 vezes o método, impondo critérios de parada em 50 iterações de propagação, e considerando o mesmo conjunto de dados com 4% e 8% dos autores pré-rotulados.

Como pode ser observado nos resultados mostrados na Figura 25, a estratégia de inicialização unitária mostrou melhores resultados em relação à estratégia aleatória. Na Figura 25a, que apresenta a acurácia obtida considerando 4% dos autores pré-rotulados, vemos que a inicialização unitária apresenta uma distribuição superior à da inicialização aleatória. Nesse caso, a inicialização unitária apresenta uma acurácia média de 66% mostra visivelmente uma variância menor quando comprada com a inicialização aleatória, a qual resultou em uma média de 57%. Analogamente, na Figura 25b, que mostra a acurácia para o conjunto de dados com 8% dos autores pré-rotulados, vemos também melhores resultados para a inicialização unitária com acurácia média de 70%, enquanto a inicialização aleatória resulta em 59%.

Considerando tais resultados, optamos pela utilização da estratégia de inicialização unitária no algoritmo proposto neste trabalho.

## 4.5 Algoritmo TCHN

Com base nos estudos de estratégias de propagação e inicialização, os quais foram apresentados nas Seções 4.3 e 4.4, chega-se à definição do Algoritmo 4 base para o método chamado TCHN.

Para compreender o funcionamento do processo iterativo executado pelo método, este pode ser descrito resumidamente da seguinte forma: o algoritmo proposto utiliza da dependência cíclica na atualização dos vetores de informação entre arestas e vértices. Partindo de uma inicialização dos vetores de confiança associados aos vértices, a cada iteração realiza inicialmente a atualização dos vetores de informação de todas as arestas $\mathbf{c}_{ij,pq}$. Em seguida utiliza os valores atualizados nas arestas para atualizar os valores dos vetores de informação dos vértices com base nos vetores das arestas adjacentes a este, aplicando as normalizações decorrentes das restrições impostas sobre a função de custo. Ao final, após a convergência dos vetores de informação ou após o número máximo de iterações toleradas, utiliza dos vetores de confiança dos vértices para estimar sua classe. Atribuindo a cada vértice a classe associada à maior entrada no vetor de confiança.

**Algoritmo 4** – Algoritmo TCHN

**Entrada:**
$G$ : rede heterogênea;
$X^l$ : conjunto de vértices rotulados;
$Y^l$ : rótulos do conjunto de vértices pré-rotulados;
$\alpha_i$ : parâmetros de concentração;
$T_A, T_S$ : conjuntos de tipos alvo e secundários;
$IMAX$ : número máximo de iterações.
**Saída:**
$Y$ : rótulos atribuídos a cada vértice de tipos alvo.
Inicializa os vetores $\mathbf{f}_{ip}, \forall i \in T, p \in n_i$;
**enquanto** não atingir a convergência **ou** número de iterações menor que $IMAX$ **faça**
    **para** toda aresta $e_{ij,pq} \in E$ **faça**
        $\mathbf{c}_{ij,pq} \leftarrow \frac{\mathbf{f}_{ip} \odot \mathbf{f}_{jq}}{\sum_k (\mathbf{f}_{ip} \odot \mathbf{f}_{jq})_k}$
    **fim para**
    **para** todo vértice $v_{ip} \in V$ **faça**
        **se** $v_{ip} \in V^l$ **então**
            $\mathbf{f}_{ip} \leftarrow \mathbf{y}_{ip}$
        **senão**
            **se** $i \in T_A$ **então**
                $\mathbf{f}_{ip} \leftarrow \alpha_i$
            **fim se**
            **para** toda aresta $e_{ij,pq} \in \varepsilon$ incidente em $v_{ip}$ **faça**
                $\mathbf{f}_{ip} \leftarrow \mathbf{f}_{ip} + W_{ij,pq}\mathbf{c}_{ij,pq}$
            **fim para**
        **fim se**
    **fim para**
    **para** todo vértice $v_{ip} \in V$ tal que $i \in T_S$ **faça**
        $\mathbf{f}_{ip} \leftarrow \frac{\mathbf{f}_{ip}}{\sum_{v_{jp} \in V_p} \mathbf{f}_{jp}}$
    **fim para**
**fim enquanto**
**para** todo vértice $v_{ip} \in V^u$ tal que $i \in T_A$ **faça**
    $\{\mathbf{y}_{ip}^{(k)} = 1 | k = argmax_{k=1}^{l} f_{ip}^{(k)}\}$
**fim para**

## 4.6 Análise de complexidade

A cada iteração de propagação, o Algoritmo 4 inicialmente opera e propaga os valores dos vetores de informação dos vértices para os vetores de informação das arestas, calculando, assim, $\mathbf{c}_{ij,pq}$ para cada aresta e cada classe. Tais operações possuem uma complexidade máxima de $O(T|E|K)$, onde $T$ é o número máximo de iterações para a convergência no processo de propagação, $|E|$ é o número de arestas na rede, e $K$ o número de classes no conjunto de dados.

O número de arestas da rede pode ser considerado um dos valores que mais afetam a complexidade do algoritmo, o qual, apesar de muito raro, pode chegar a uma ordem quadrática do número de vértices

das camadas em redes densas, próximas de completas. Porém, considerando redes esparsas, podemos assumir $|E| \sim |V|$.

Em seguida, o algoritmo opera propagando a informação das arestas para os vértices. Considerando um rede esparsa com grau médio dos vértices e um valor limitado a uma constante, tal operação possui complexidade de $O(TnK)$, onde $n$ é o número total de vértices da rede para todos os tipos. Assim, a complexidade total do algoritmo TCHN para a classificação de um conjunto dado é da ordem de:

$$O(T(|E|+n)K) \tag{4.16}$$

Utilizando uma estrutura de dados otimizada para o tratamento de arestas o algoritmo proposto tem seu custo computacional bastante atrativo. Além disso, este pode ser facilmente paralelizado dados que a atualização de cada camada é independentes umas das outras, sendo necessário apenas algum cuidado no acesso multo de dados.

## 4.7 Considerações finais

Neste capítulo o método TCHN é apresentado em detalhes. Inicialmente são apresentados os elementos associados aos vértices e arestas da rede e a modelagem matemática, mostrando algumas ilustrações para a melhor compreensão do desenvolvimento e todas as manipulações realizadas na dedução das equações de atualização dos vetores de informação, os quais embasam o algoritmo iterativo. Em seguida, são mostrados alguns estudos realizados ao longo do desenvolvimento do método para otimizar os resultados deste. Finalmente, o algoritmo iterativo é apresentado bem como sua analise de complexidade.

Apesar da evidente contribuição do método aqui proposto, vale ressaltar que algum existem ainda pontos de potencial desenvolvimento a serem estudados. Como por exemplo, neste trabalho não foi ponderado a propagação entre diferentes tipos de vértices e arestas, considerando assim a propagação entre os tipos de forma uniforme. Uma possível melhoria aqui é a ponderação da propagação onde as relações possuírem maior relevância para classificação. Outro ponto passível de ser explorado é a utilização de métodos de pré-processamento de informações textuais como o caso de *embeddings* para generalizar os conceitos expressados no texto, e assim melhor relacionar diferentes tópicos.

CAPÍTULO

5

# GERAÇÃO DE REDES HETEROGÊNEAS SINTÉTICAS

Uma das grandes dificuldades para o desenvolvimento de novas técnicas em redes heterogêneas é o levantamento de conjuntos de dados de qualidade e com características conhecidas, de modo que técnicas em desenvolvimento possam ser testadas em contextos controlados. Desta forma como já apontado em (ANGELOVA; KASNECI; WEIKUM, 2012), existe na literatura de HIN a necessidade por ferramentas para a geração de redes heterogêneas sintéticas para auxiliar o estudo e desenvolvimento de técnicas de classificação e detecção de comunidades.

Como o objetivo de preencher tal lacuna da área, ao longo deste trabalho de doutorado foi desenvolvido uma extensão da ferramenta de BNOC[1] (VALEJO *et al.*, 2019) de geração de redes bipartidas com sobreposição de comunidades. Como será visto a seguir uma rede heterogênea pode ser vista como uma composição de $m$ redes bipartidas ou homogêneas (no caso onde os dois tipos ligados serem o mesmo). Desta forma, as operações modulares base da ferramenta BNOC são utilizadas na construção de redes heterogêneas sintéticas com base em um conjunto de parâmetros. Nas próximas seções são descritas as operações básicas, bem como a construção do algoritmo chamado HNOC[2].

## 5.1 BNOC

A Ferramenta BNOC desenvolvida no grupo de pesquisa LABIC teve como motivação principal auxiliar a analise de redes complexas em diversas tarefas, provendo recursos com grande potencial de utilização, capaz de gerar redes bipartidas com diversas características de forma controlada. Resumidamente, a ferramenta BNOC recebe uma série de parâmetros descritos na Tabela 2 gerando uma rede bipartida com base na distribuição binomial negativa, para maiores detalhes sobre geração de rede dada pela BNOC recomendamos a leitura de (VALEJO *et al.*, 2019).

Na Figura 26 são mostradas duas redes bipartidas geradas pela ferramenta BNOC.

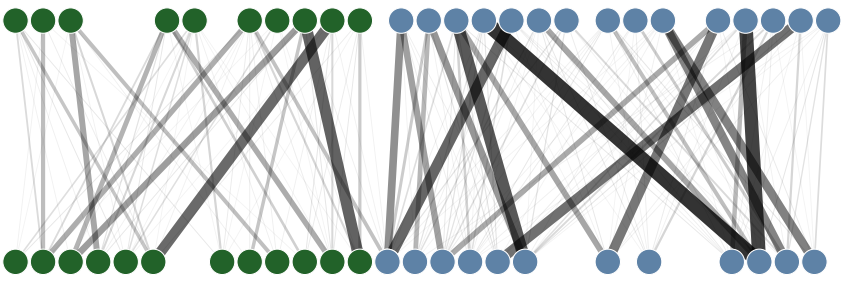
[1] sigla do inglês para *Bipartite Network with Overlapping Community*
[2] sigla do inglês para *Heterogeneous Network with Overlapping Community*

Tabela 2 – Descrição do parâmetros recebidos pela BNOC.

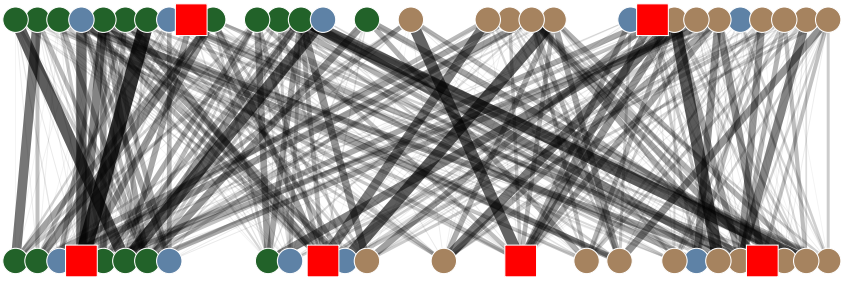
| Parâmetro | Domínio | Default | Descrição |
|---|---|---|---|
| -v, --vertices | $[1, \|V\|] \subseteq \mathbb{Z}$ | $[10, 20]$ | Number of vertices for each layer |
| -c, --communities | $[1, \|V\|] \subseteq \mathbb{Z}$ | $[2, 2]$ | Number of communities for each layer |
| -p0, --probability | $(0, \|V\|] \subseteq \mathbb{R}$ | $[0.3, 0.7]$ | Probabilities for vertices in layer 0 for each community |
| -p1, --probability1 | $(0, \|V\|] \subseteq \mathbb{R}$ | $[0.3, 0.7]$ | Probabilities for vertices in layer 1 for each community |
| -b, --balanced | $\{0, 1\}$ | 0 | Boolean balancing flag that suppresses -p parameter |
| -x, --overlap | $[0, \|V_0\|] \subseteq \mathbb{Z}$ | 1 | Number of overlapping vertices in $V_1$ |
| -y, --overlap | $[0, \|V_1\|] \subseteq \mathbb{Z}$ | 1 | Number of overlapping vertices in $V_2$ |
| -z, --noverlap | $[0, \|c\|] \subseteq \mathbb{Z}$ | 2 | Number of overlapping communities |
| -d, --dispersion | $\mathbb{R}_+$ | 0.1 | Dispersion of negative binomial distribution |
| -s, --success | $\mathbb{R}_+$ | 1 | Probability of success |
| -n, --noise | $(0, 1] \subseteq \mathbb{R}$ | 0.01 | Noise |
| -l, --normalize | $\{0, 1\}$ | 0 | 0-1 scale weight individually to unit norm |
| -u, --unweighted | $\{0, 1\}$ | 0 | Unweighted bipartite networks |

Fonte: Elaborada pelo autor.

Figura 26 – Exemplos de redes bipartidas geradas pela ferramenta BNOC. Nas figuras a densidade das linhas representam seus pesos, nos vértices quadrados vermelhos representam vértices de sobreposição de comunidades, e círculos coloridos vértices de não sobreposição de uma só comunidade.



(a) $v = [25, 15]$, $c = [2, 2]$, $d = 0.5$, $b$



(b) $v = [25, 30]$, $c = [2, 2]$, $d = 0.8$, $x = 4$, $y = 2$, $z = 2$, $n = 0.2$, $b$

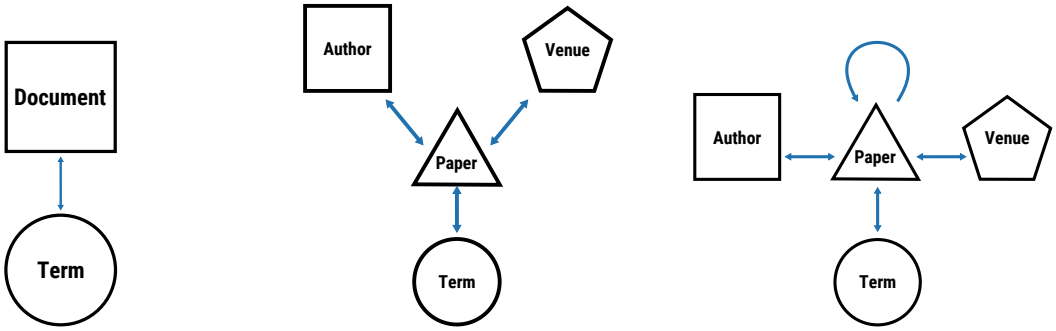
Fonte: Elaborada pelo autor.

## 5.2 HNOC

Considerando como generalizações naturais de redes bipartidas as redes k-partidas e heterogêneas, utilizamos as operações implementadas na ferramenta BNOC para implementar uma ferramenta de geração de redes heterogêneas gerais. Esse trabalho foi motivado pelo fato de que, assim como na área de redes bipartidas, encontramos na literatura de redes heterogêneas a carência por ferramentas de *benchmarking* para auxiliar o estudo de desenvolvimento nas áreas de classificação e detecção de comunidades.

Dado o número de tipos e conexões em uma HIN, sua estrutura é potencialmente complexa de se tratra e também visualizar. Uma ferramenta util para compreender e explorar as HINs são os esquema de rede, que dão uma visão geral dos tipos e estruturas das conecções contidas na rede. Como já descrito, um esquema de rede um meta-modelo para a rede, onde as entidades são mapeadas para seus tipos e as conexões são mapeadas como conexões entre tipos. Assim, o esquema para uma HIN $\mathscr{G}$, que é denotado por $T_{\mathscr{G}}(A,R)$ , é uma rede direcionada definida sobre arestas de tipos $A$ com arestas de relação em $R$, obtidas pelas funções de mapeamento $\varphi : V \to A$ and $\psi : E \to R$, respectivamente. Na Figura 27 são ilustrados alguns exemplos de esquemas para diferentes tipo de redes heterogêneas, onde fica claro a variedade de estruturas possíveis.

Figura 27 – Esquemas de rede com diferentes configurações para um mesmo conjunto de dados bibliográficos.




(a) Esquema de uma rede bipartida.

(b) Esquema de uma rede k-partida.

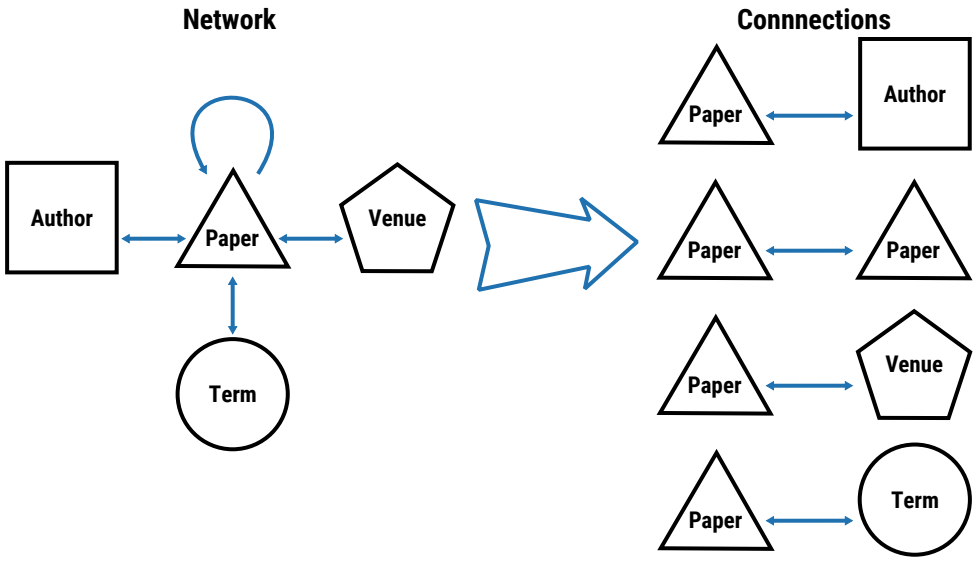
(c) Esquema de uma rede heterogênea.

Fonte: Elaborada pelo autor.

Um esquema representa os $m$ tipos de entidades e $r$ tipos de conexões. Cada tipo de conexão é composto por três partes: dois tipos extremos de entidades, e a definição do sua conexão com seu significado, dado que podem existir mais de um tipo de conexões entre o mesmo par de entidades, como o caso de rede multi-relações (YANG *et al.*, 2012). Com isso, uma HIN pode ser vista como uma composição de $r$ redes bipartidas (ou homogêneas, no caso de os dois extremos serem de mesmo tipo), como é ilustrado na Figura 28.

Considerando tal interpretação de uma HIN, foi implementada uma extensão da ferramenta BNOC para a geração sintética de instâncias de redes heterogêneas a partir de um esquema com o uso dos passos modulares disponíveis pela ferramenta BNOC. A construção se dá pela iteração sobre cada tipo de elemento, gerando as comunidades baseadas no modelo distribuição de probabilidades com proposto pela BNOC, e em seguida iterando sobre os tipos de conexões gerando as arestas entre as camadas de

Figura 28 – Visualização de um HIN como $r$ redes bipartidas/homogêneas.




Fonte: Elaborada pelo autor.

Tabela 3 – Descrição do parâmetros adicionais incluídos na extensão HNOC.

| Parâmetro | Domínio | Default | Descrição |
|---|---|---|---|
| -m, –layers | $\mathbb{N}$ | 3 | Número de tipos |
| -v, –vertices | $(0, \lvert V \rvert] \in \mathbb{N}$ | $[10, 10, 10]$ | Número de vértices de cada tipo |
| -e, –schema | $\mathbb{N}$ | *null* | Lista de pares de vértices, para tipos de conexões |
| -p, –probabilities | $(0, \lvert V \rvert] \in \mathbb{R}$ | *null* | Probabilidade de cada comunidade para cada camada |
| -b, –balanced | $0, 1$ | $[0, 0, 0]$ | booliano que indica distribuição uniforme de comunidades |
| -x, –overlap | $(0, \lvert V \rvert] \in \mathbb{N}$ | $[0, 0, 0]$ | Número de vértices de sobreposição para cada camada |
| -z, –noverlap | $(0, \lvert c \rvert] \in \mathbb{N}$ | $[2, 2, 2]$ | Número de comunidades sobrepostas para cada camada |

Fonte: Elaborada pelo autor.

acordo com o esquema dado como parâmetro. Resumidamente, esse processo é executado em duas partes baseadas nos passos da BNOC, como descrito a seguir:

- Parte 1: Itera um $m$ laço executando os passos 1 e 2 da BNOC, construindo cada camada $i$ com vértices $V_i$, atribuindo as comunidades para cada vértice considerando as propriedades de sobreposição desejadas.

- Parte 2: Itera um laço sobre todos os tipos de conexões dados por pares de tipos de vértices do esquema da rede, executando os passos 3, 4 e 5 da BNOC sobre cada camada, estabelecendo as conexões das arestas, construindo os pesos, densidade e ruido na rede de acordo com os parâmetros de entrada.

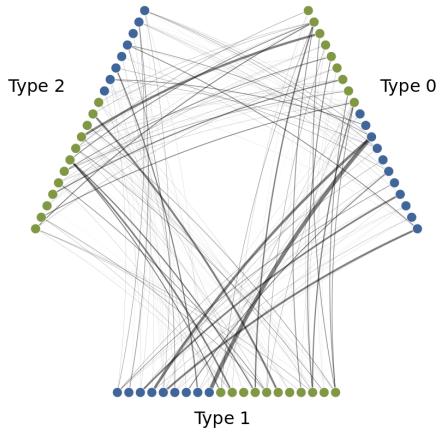
Dado que nas redes heterogêneas temos diversas camadas e tipos de conexões, alguns parâmetros da BNOC foram alteradas e novos parâmetros foram inseridos para que fosse possível estender tal ferramenta. A lista de parâmetros modificados e adicionados é apresentada na Tabela 3

Na Figura 29 são mostradas algumas redes obtidas como resultado da ferramenta HNOC desenvolvida, com a imposição de algumas combinações de parametros. Nas figuras são mostrados os parâmetros não *non-default* sobre cada rede gerada, os parâmetros *default* foram omitidos nas legendas e são mostrado na Tabela 2 e 3). A rede mostrada na Figura 29a, é uma rede de 3 camadas, com uma estrutura $k$-partida, construída para não ter sobreposições entre comunidades, já a rede mostrada na Figura 29b é uma rede mais densa contendo sobreposição de comunidades. Apesar de em ambos os casos a probabilidade ser dada como balanceada, podemos observar nas redes que as comunidades apresentam variações de tamanha, tal característica se dá pelo fato de que a distribuição não garante os tamanhos das comunidades, mas sim a média do comportamento de redes geradas pelo modelo.

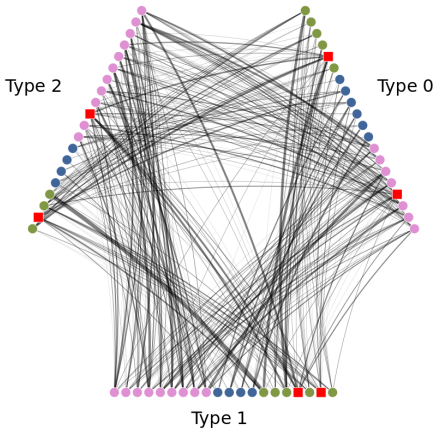
A HIN mostrada na Figura 29c instancia uma rede de 4 camadas com estrutura $k$-partida, topologia densa e sobreposição entre comunidades. Finalmente, a Figura 29d mostra uma rede com estrutura heterogênea, onde os vértices do tipo 1 possuem conexões entre eles mesmos.

Assim como a ferramenta BNOC, a extensão desenvolvida possui um grande potencial de utilização ao gerar redes sintéticas e auxiliar o desenvolvimento e validação de novas técnicas nas mais diversas tarefas da área de estudo de HIN. Dado que a ferramenta HNOC é baseada diretamente da implementação da BNOC, herda dessa suas principais características, como sua flexibilidade e robustez para a geração de redes de referência, capaz de gerar uma variedade de rede com diferentes características topológicas em tempos razoáveis. Para mais detalhes das ferramentas BNOC e HNOC, recomendamos a leitura de (VALEJO *et al.*, 2019).

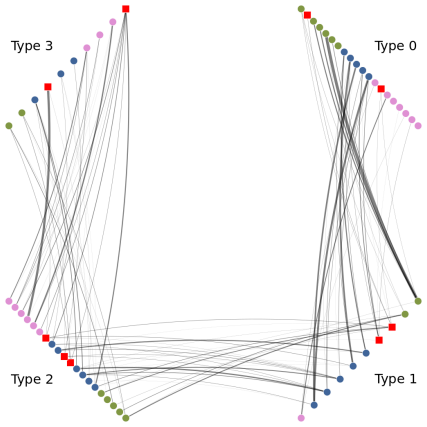
Figura 29 – Exemplos de redes heterogêneas com diferentes estruturas topologicas e propriedades geradas pela ferramenta HNOC. Nas figuras a densidade das linhas representam seus pesos, nos vértices quadrados vermelhos representam vértices de sobreposição de comunidades, e círculos coloridos vértices de não sobreposição de uma só comunidade. O layout que representa a rede é baseado na técnica PolyViz apresentada em (USLU; MEHLER, 2018)



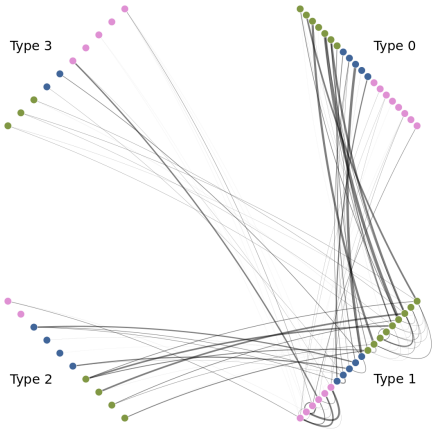

(a) `v=[20,20,20], c=2, e =[(0,1), (1,2), (2,0)]`




(b) `v=[20,20,20], c=3, e=[ (0,1), (1,2), (2,0) ], x=[2,2,2], d=0.8`




(c) `m=4, v=[20,10,20,10], c=3, e=[ (0,1), (1,2), (2,3) ], x=[2,2,3,2], z=[2,2,2,2]`




(d) `m=4, v=[20,20,10,10], c=3, e=[ (0,1), (1,2), (1,3), (1,1) ]`

Fonte: Elaborada pelo autor.

CAPÍTULO

# 6

# EXPERIMENTOS

O método proposto foi avaliado considerando-se o cenário de classificação transdutiva em redes heterogêneas, foram utilizados como comparação três método encontrados na literatura que consideramos como o estado da arte nesta linha de pesquisa, os quais são: GNetMine, HetPathMine e HeteClass, todos descritos no Capítulo 3. Para a avaliação, utilizamos quatro conjuntos de dados, dois de mundo real: DBLP e Flickr Fashion 10000; e dois sintéticos gerados com o uso da biblioteca HNOC descrita no Capítulo 5.

Neste capitulo inicialmente descrevemos os conjunto de dados utilizados na Seção 6.1. Na Seção 6.2 é descrita a modelagem dos experimentos e medida de qualidade adotada para a avaliação neste trabalho. Finalmente, na Seção 6.3 são apresentados os resultados comparativos do método proposto e métodos comparados mostrando os bons resultados obtidos pelo trabalho conforme metodologia adotada.

## 6.1 Conjuntos de dados

Como já discutido no Capítulo 2 a representação de dados por HIN é naturalmente expressiva e capaz de representar a natureza heterogênea e complexa de muitos dados de mundo-real. Ao mesmo tempo, tal complexidade faz da modelagem dos dados por HIN uma tarefa importante para uma melhor compreensão dos dados tanto do ponto de vista humano quanto para obtenção de melhores resultados em métodos de aprendizado de máquina.

Nesta seção buscamos descrever detalhadamente os conjuntos de dados utilizados, bem como a modelagem das entidades e conexões que são base para o esquema que define as HIN classificadas pelo método TCHN.

### *6.1.1 DBLP*

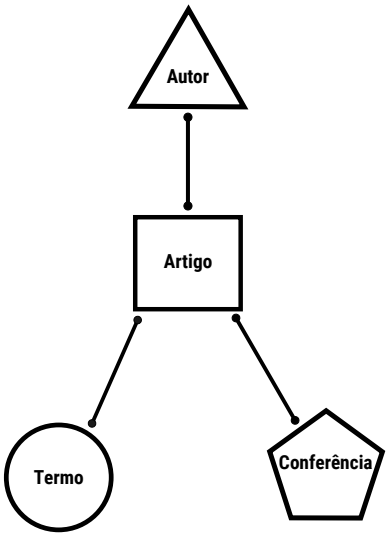
Neste trabalho foi utilizado a bem conhecida base de dados chamada DBLP [1] de dados bibliográficos, largamente utilizada em técnicas de classificação transdutiva de HIN. A base de dados DBLP em si, está publicamente disponível na forma de um repositório bibliográfico de ciência da computação, contendo

[1] https://dblp.uni-trier.de/

dados sobre publicações com seus conteúdos textuais, autores relacionados dentre outras informações. Desta forma, existem por si uma natureza heterogênea neste conjunto de dados, podendo ser representado com diversas entidades e relações entre estas.

Foram extraídas da base de dados entidades dos tipos: **autor**, **conferência**, **artigo** e **termo**. A coocorrência de um par de entidades em uma entrada de dado foi considerada como conexão, de forma que a HIN foi modelada com base no seguinte conjunto de entidades e conexões extraídas levando ao esquema de rede mostrada na figura 30.

Figura 30 – Esquema de rede utilizada baseda na base de dados DBLP.




Fonte: Elaborada pelo autor.

Nesta modelagem, foi extraído um subconjunto dos dados conhecido como *four-areas dataset*, o qual contém 18 conferências em 4 áreas: *Data Mining*, *Database*, *Information Retrieval* e *Machine Learning*. As conferências relacionadas em cada área são listadas a seguir:

- **Data Mining**: KDD, ICDM, SDM e PAKDD;
- **Database**: SIGMOD, VLDB, ICDE, PODS e EDBT;
- **Information Retrieval**: SIGIR, ECIR, WSDM e WWW, CIKM;
- **Machine Learning**: NIPS, ICML, AAAI e IJCAI.

Dentro deste subconjunto, foram selecionado com base nos autores que mais publicaram. Dos títulos dos artigos foram removidas as *stopwords*, e realizado o processo de *stemming* com auxílio da biblioteca **NLTK** em Python.

Com base nos dados obtidos pelo processamento descrito acima foi a HIN, a qual teve seus vértices rótulos de acordo dentro das quatro classes dadas pelas áreas de pesquisa descritas acima, tais rotulos foram obtidos com base os dados disponibilizados por Ji *et al.* (2010) [2]. Foram construídas

[2] http://web.cs.ucla.edu/ yzsun/data/

duas HINs contendo os 100 e 500 autores com maior número de publicações, os quais chamaremos de *DBLP_top*100 e *DBLP_top*500. Na Tabela 4 são descritas as propriedades numéricas das duas HIN construidas com base na DBLP e utilizadas neste trabalho.

Tabela 4 – Descrição do número de vértices e arestas nas HIN construídas baseadas na DBLP.

| | | *DBLP_top*100 | *DBLP_top*500 |
|---|---|---|---|
| **vértices** | autor | 100 | 500 |
| | artigo | 4.747 | 11.107 |
| | conferência | 20 | 20 |
| | termo | 4.399 | 7.322 |
| **arestas** | artigo ↔ autor | 5.901 | 15.414 |
| | artigo ↔ conferência | 4.747 | 11.107 |
| | artigo ↔ termo | 36.866 | 87.104 |
| **vértices rotulados** | autor | 43 | 194 |
| | conferência | 20 | 20 |

Fonte: Dados da pesquisa.

Para o método GNetMine e o método proposto TCHIN, as HINs foram utilizadas como modeladas considerando todos os tipos de relações, ou seja: (artigo $\longleftrightarrow$ autor), (artigo $\longleftrightarrow$ conferência) e (artigo $\longleftrightarrow$ termo). Já para os método baseados em meta-caminhos HetPathMine e HeteClass, foram utilizados quatro meta-caminhos, seguindo a metodologia do artigo de referência. Na Tabela 5 são descritos os meta-caminhos utilizados nos experimentos deste trabalho para as HIN construídas com base na DBLP.

Tabela 5 – Meta-caminhos utilizados nos métodos HetPathMine e HeteClass para as HINs basedas no conjunto de dados DBLP

| **Meta-Caminho** |
|---|
| autor(A) $\longrightarrow$ artigo(P) $\longrightarrow$ autor(A) |
| autor(A) $\longrightarrow$ artigo(P) $\longrightarrow$ conferência(C) $\longrightarrow$ artigo(P) $\longrightarrow$ autor(A) |
| autor(A) $\longrightarrow$ artigo(P) $\longrightarrow$ termo(T) $\longrightarrow$ artigo(P) $\longrightarrow$ autor(A) |
| autor(A) $\longrightarrow$ artigo(P) $\longrightarrow$ autor(A) $\longrightarrow$ artigo(P) $\longrightarrow$ autor(A) |

Fonte: Elaborada pelo autor.

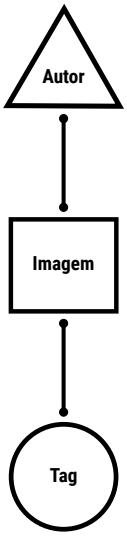
### 6.1.2 *Flickr Fashion* 10.000

O segundo conjunto de dados de mundo real utilizado foi o conjunto Flickr fashion 10.000 proposto por Loni *et al.* (2014), por via de um rastreador de imagens na web. A base de dados publicamente encontrada contem mais de 32000 imagens coletadas da plataforma de compartilhamento de imagens Flickr, considerando apenas imagens com alguma relação com temas de vestimenta e estilo. Dentro desta base, as imagens são distribuídas entre um total de X classes. Além das classe, as imagens são extraídas em conjunto com diversos dados contextuais bem como informações sobre seus autores e suas conexões nesta rede de compartilhamento.

A natureza heterogênea deste conjunto de dados, pode ser observada dadas as diversas informações sobre as imagens disponíveis, como por exemplo: seus autores, localização geográfica, temática dentre outras características. Sendo assim, uma base bastante atrativa para a utilização neste trabalho.

Foram extraídas dasta base entidades dos tipos: **imagem**, **autor** e ***tag***, onde a **imagem** é o conteúdo principal dos dados disponíveis na web sobre esta base de dados, o **autor** é uma informação coletada em conjunto com a imagem e a ***tag*** é obtida também através das informações sociais coletadas em conjunto com a imagem. As conexões consideradas são as obtidas naturalmente pela relação entre imagem $\longleftrightarrow$ autor e imagem $\longleftrightarrow$ *tag*, levando assim ao esquema de rede ilustrado na Figura 31.

Figura 31 – Esquema de rede utilizada baseada na base de dados Flickr fashion 10.000.




Fonte: Elaborada pelo autor.

Nesta modelagem, foi extraído um subconjunto dos dados contidos nas 20 classes com o maior número de imagens, onde se encontram um total de aproximadamente 4 mil imagens. Deste conjunto de dados foram criadas duas HINs uma contendo todas as imagens e outras contendo um subconjunto de 2 mil imagens selecionadas aleatoriamente, tais HIN foram chamadas $Flickr\_4K$ e $Flickr\_2K$, respectivamente. Na Tabela 6 são descritas as propriedades numéricas das duas HIN construídas com base na Flickr Fashion 10.000 e utilizadas neste trabalho.

Tabela 6 – Descrição do número de vértices e arestas nas HIN construídas baseadas no conjunto de dados Flickr Fashion 10.000.

| | | $Flickr\_4K$ | $Flickr\_2K$ |
|---|---|---|---|
| **vértices** | imagem | 3.895 | 2.000 |
| | autor | 721 | 464 |
| | *tag* | 9.122 | 6.220 |
| **arestas** | imagem $\leftrightarrow$ autor | 3.942 | 2.028 |
| | imagem $\leftrightarrow$ *tag* | 69.304 | 35.501 |
| **vértices rotulados** | imagem | 3.895 | 2.000 |

Fonte: Dados da pesquisa.

Assim como feito para o caso da DBLP, nos métodos de propagação geral GNetMine e TCHIN, as HINs foram utilizadas como modeladas considerando todos os tipos de relações, ou seja: (imagem $\longleftrightarrow$ autor) e (imagem $\longleftrightarrow$ tag). Já para os método baseados em meta-caminhos HetPathMine e HeteClass, foram utilizados quatro meta-caminhos, seguindo a metodologia do artigo de referência. Na Tabela 7 são descritos os meta-caminhos utilizados nos experimentos deste trabalho para as HIN construídas com base na Flickr fashion 10.000.

Tabela 7 – Meta-caminhos utilizados nos métodos HetPathMine e HeteClass para as HINs basedas no conjunto de dados Flickr Fashion 10.000

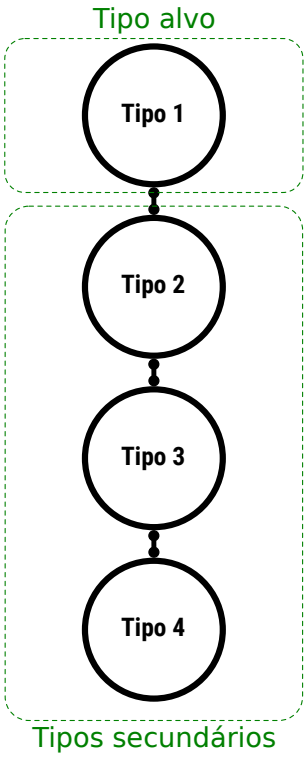
| **Meta-paths** |
|---|
| imagem(I) $\rightarrow$ autor(A) $\rightarrow$ imagem(I) |
| imagem(I) $\rightarrow$ *tag*(T) $\rightarrow$ imagem(I) |
| imagem(I) $\rightarrow$ autor(A) $\rightarrow$ imagem(I) $\rightarrow$ autor(A) $\rightarrow$ imagem(I) |
| imagem(I) $\rightarrow$ *tag*(T) $\rightarrow$ imagem(I) $\rightarrow$ *tag*(T) $\rightarrow$ imagem(I) |

### 6.1.3 Base de dados sintética

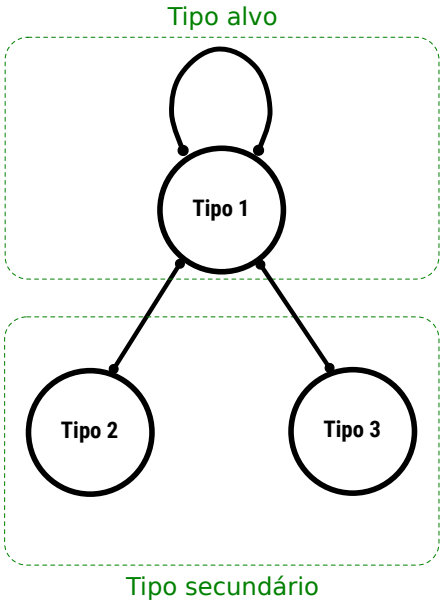
Além dos dois conjunto de dados de mundo real descritos nas seções anteriores, os resultados do método TCHN e os outros métodos foram comparados utilizando duas HINs sintéticas construídas com o auxílio da ferramenta HNOC descrita no Capítulo 5. Para melhor aproveitar a capacidade de geração de redes diversas da ferramenta, foram construídas duas redes com estruturas distintas das de mundo real já utilizadas, buscando explorar novas estruturas de rede como redes de muitas camadas e com conexões entre vértices de um mesmo tipo.

Na Figura 32 são mostrados os dois esquemas de rede utilizados para a geração das redes sintéticas. Nas Tabelas 8 e 9 são mostrados os principais parâmetros utilizados na HNOC, sendo estes a quantidade de vétices para cada tipo de vértice e a distribuição de probabilidades dos vertices dentre as classes. Além destes parametros foram utilizados os seguintes valores para os demais: $-dispersion = 0.15$, $-mu = 0.15$ e $-noise = 0.3$, cujos valores resultam na geração de redes esparsas com ruido moderado.

Figura 32 – Dois esquemas de rede utilizados para a geração das redes sintéticas com auxilio da ferramenta HNOC.




(a) Esquema de uma rede com múltiplas camadas.




(b) Esquema de uma rede com conexão entre vértices de um mesmo tipo.

Fonte: Elaborada pelo autor.

Tabela 8 – Distribuição de vértices e probabilidades utilizadas para a construção da primeira rede sintética utilizada neste trabalho.

| | Tipo 1 | Tipo 2 | Tipo 3 | Tipo 4 |
|---|---|---|---|---|
| Número de vértices | 200 | 500 | 1000 | 20 |
| Probabilidades para a classe 1 | 0.4 | 0.5 | 0.4 | 0.2 |
| Probabilidades para a classe 2 | 0.1 | 0.1 | 0.2 | 0.2 |
| Probabilidades para a classe 3 | 0.2 | 0.1 | 0.1 | 0.2 |
| Probabilidades para a classe 4 | 0.1 | 0.2 | 0.2 | 0.2 |
| Probabilidades para a classe 5 | 0.2 | 0.1 | 0.1 | 0.2 |

Tabela 9 – Distribuição de vértices e probabilidades utilizadas para a construção da segunda rede sintética utilizada neste trabalho.

| | Tipo 1 | Tipo 2 | Tipo 3 |
|---|---|---|---|
| Número de vértices | 500 | 1000 | 1000 |
| Probabilidades para a classe 1 | 0.5 | 0.54 | 0.45 |
| Probabilidades para a classe 2 | 0.46 | 0.4 | 0.5 |
| Probabilidades para a classe 3 | 0.04 | 0.06 | 0.05 |

## 6.2 Modelagem dos experimentos

Na validação do método TCHN, os quatro conjuntos de dados descritos na seção anterior foram utilizados, considerando-se diferentes porções de elementos rotulados como inicialização para os métodos de propagação transdutiva comparados, os quais são: GNetMine, HetPathMine e HeteClass. Como medida de avaliação utilizamos a acurácia, calculada sobre uma porção $\alpha$, em geral de 50%, pre-definida dentre os elementos que se possui classificação dos dados, porém tais classificações não são utilizadas como entrada dos métodos. Ou seja, a cada execução de cada um dos métodos considerando um conjunto de dados ao qual se tem conhecimento *a priori* de classe sobre uma porção $\alpha$ dos vértices, seleciona-se um conjunto $V^v$ contendo uma porção $\frac{\alpha}{2}$ para a validação, e um conjunto $V^l$ contendo uma porção $\rho$ do qual as classes são dadas como entrada para os métodos de classificação transdutiva em HIN aqui considerados, em todos os cenários os vértices em cada conjunto são selecionado aleatoriamente de forma proporcional à distribuição de classes.

Para melhor avaliar o comportamento médio dos métodos, considerando a aleatoriedade dos conjuntos iniciais, foram realizados 10 execuções para cada experimento e suas médias foram calculadas para a obtenção dos resultados.

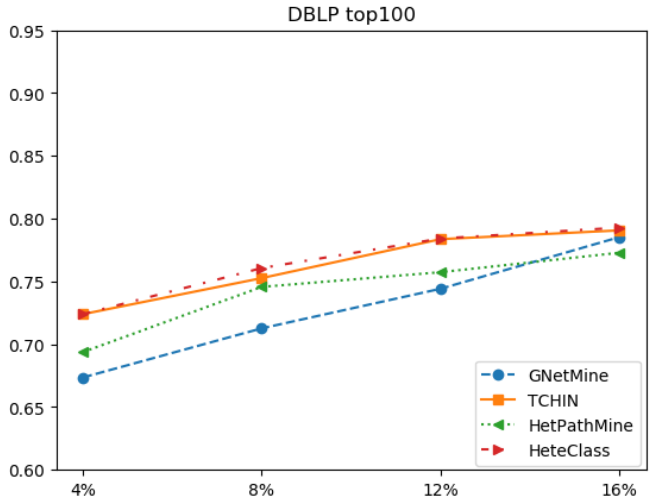
## 6.3 Resultados

Nesta seção são apresentados os resultados de comparação do método TCHN com outros métodos selecionados. Todos os métodos foram testados com redes heterogêneas baseadas nos conjuntos de dados DBLP, Flirck Fashion 10.000 e dados sintéticos criados com o uso da ferramenta HNOC.

### 6.3.1 Resultados obtidos para *DBLP_100* e *DBLP_500* utilizando rótulos de um tipo de vértice

Para os dois casos dos experimentos mostrados nesta subseção foram utilizados rótulos apenas nos vértices do tipo autor em diferentes proporções como entrada para a propagação. Nas Figuras 33 e 34 são apresentados os resultados de acurácia obtidos as HINs basedas na DBLP comparando os diferentes métodos. Como pode ser observado nestas figuras os resultados obtidos pelo método TCNH, proposto neste trabalho, se apresentam sempre dentre os melhor na comparação com métodos representativos da área. Apenas sendo superado pelo método HeteClass e mesmo neste caso por uma diferença pequena. Os mesmos resultados podem ser avaliados numericamente nas tabelas 10 e 11.

Figura 33 – Acurácia obtida para a base de dados DBLP top 100.



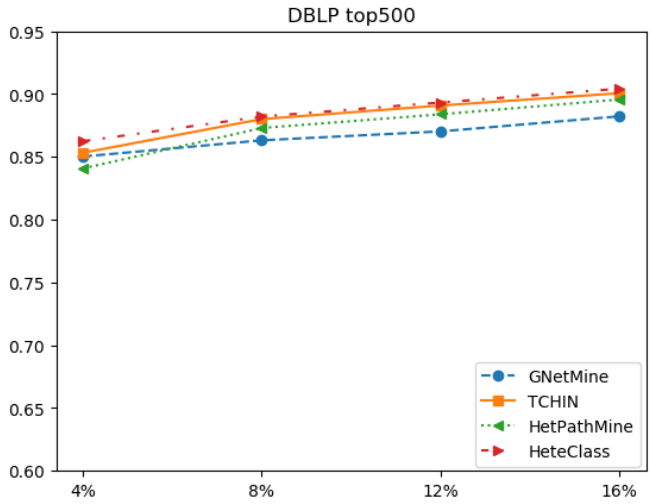

Fonte: Elaborada pelo autor.

Tabela 10 – Resultados obtidos para subconjunto da base de dados DBLP contendo os 100 autores com maior produção, sendo destes 36 autores rotulados entre as quatro áreas.

| **Método \ Rotulados** | 4% | 8% | 12% | 16% |
|---|---|---|---|---|
| GNetMine | 0.6736 | 0.7127 | 0.7442 | 0.7853 |
| TCHN | 0.724 | 0.7527 | 0.7837 | 0.7907 |
| HetPathMine | 0.6938 | 0.7457 | 0.7574 | 0.7729 |
| HeteClass | 0.724 | 0.7607 | 0.7842 | 0.7929 |

Fonte: Dados da pesquisa.

Figura 34 – Acurácia obtida para a base de dados DBLP top 500.




Fonte: Elaborada pelo autor.

Tabela 11 – Resultados obtidos para subconjunto da base de dados DBLP contendo os 500 autores com maior produção, sendo destes 175 autores rotulados entre as quatro áreas.

| **Método \ Rotulados** | 4% | 8% | 12% | 16% |
|---|---|---|---|---|
| GNetMine | 0.8503 | 0.8632 | 0.8703 | 0.8823 |
| TCHN | 0.8533 | 0.8801 | 0.8909 | 0.9007 |
| HetPathMine | 0.8408 | 0.873 | 0.8839 | 0.8957 |
| HeteClass | 0.8625 | 0.8821 | 0.8933 | 0.9043 |

Fonte: Dados da pesquisa.

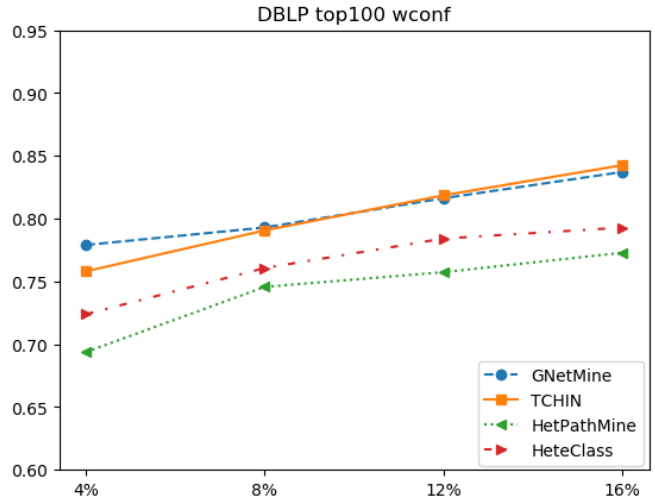
Para a HIN *DBLP*_100 os resultados entre o método HeteClass se aproximaram, bastante sendo ambos sempre os melhores e na maioria dos resultados apresentando uma superioridade de mais de $2-3\%$ de acurácia. Já para a HIN *DBLP*_500 os resultados obtidos são mais próximos entre todos os métodos.

Como já mencionado no Capítulo 2, os métodos de propagação baseados em meta-caminhos não consideram rótulos de vértices que não estão no conjunto alvo, por exemplo na DBLP apenas são capazes de considerar os rótulos dos autores e não dos vértices de conferência. Assim, não aproveita de informações potencialmente valiosas dos dados. Por outro lado, tais informações podem ser utilizadas livremente pelos métodos de propagação geral pela rede, como o GNetMine e TCHN, nestes todos os tipos de elementos da rede recebem um vetor de informação onde se considera seu rotulo conhecido *a priori*. Para explorar tal característica do método TCHN, foi utilizado um HIN construída com o conjunto de dados DBLP com rótulos associados para autores e conferencias, e assim mostrar as vantagens deste uso na classificação. Dados que em ambos os tipos de vértices (autores e conferências) os rótulos coincidem, pois são rótulos das quatro áreas de pesquisa, não existe perda de semântica dos rótulos se fazendo a propagação entre tipos.

### 6.3.2 Resultados obtidos para para *DBLP*_100 utilizando rótulos de mais de um tipo de vértice

Para os resultados mostrados nesta subseção, sobre a HIN *DBLP*_100 foram considerados rótulos em diferentes porcentagens para os vértices do tipo autor e para o tipo conferência foram considerados os rótulos de quatro vértices escolhidos aleatoriamente, sendo um de cada classe. Na Figura 35 são mostrados os resultados obtidos de acurácia, os quais também são mostrados numericamente na Tabela 12. É possível observar pelos resultados obtidos que a informação adicional dos rótulos de conferências pode aumentar a acurácia dos métodos de propagação geral de forma significativa em até oito pontos percentuais, fazendo com que tais métodos superassem os métodos baseados em meta-caminhos. Neste cenário, os resultados obtidos pelos método proposto TCHN se mantêm entre os melhores sendo em superior em alguns pontos.

Figura 35 – Acurácia para a base de dados DBLP top 100.




Fonte: Elaborada pelo autor.

Tabela 12 – Resultados obtidos para subconjunto da base de dados DBLP contendo os 100 autores com maior produção, sendo destes tem-se variáveis números de autores mais 4 conferências rotulados entre as quatro áreas.

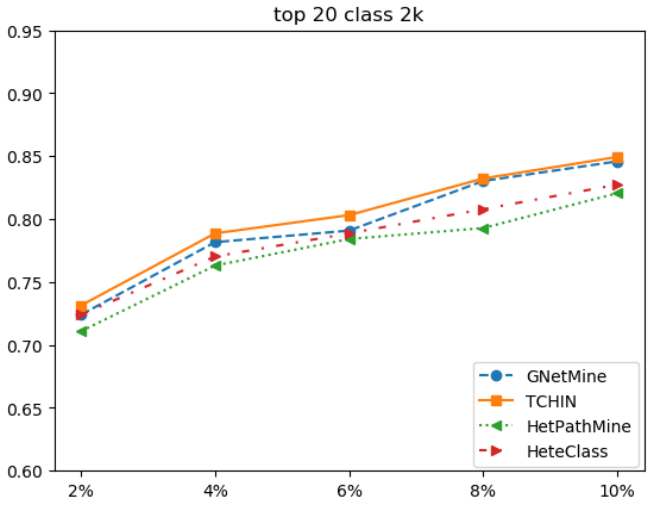
| **Método \ Rotulados** | 4% | 8% | 12% | 16% |
|---|---|---|---|---|
| GNetMine | 0.7791 | 0.793 | 0.8163 | 0.8372 |
| TCHN | 0.7581 | 0.7907 | 0.8186 | 0.8426 |
| HetPathMine | 0.6938 | 0.7457 | 0.7574 | 0.7729 |
| HeteClass | 0.724 | 0.7607 | 0.7842 | 0.7929 |

Fonte: Dados da pesquisa.

### 6.3.3 Resultados obtidos para $Flickr\_4K$ e $Flickr\_4K$ utilizando rótulos de um tipo de vértice

Considerando as HINs construídas com base no conjunto de dados Flirk Flashion 10,000, foram realizados experimentos com diferentes proporções de imagens rotulados dadas como entrada nos métodos comparados. Nas figuras 36 e 37 são mostrados os resultados obtidos de acurácia para as duas HINs $Flickr\_4K$ e $Flickr\_2K$, os mesmos resultados são apresentados numericamente nas tabelas 13 e 14. Como pode ser observado pelo resultados, para estas entradas a acurácia dos métodos de propagação TCHN e GNetMine superam os métodos baseados em meta-caminhos. Além disso para estas entradas o método TCHN é o que produz melhores resultados, confirmando assim seu grande potencial de aplicação.

Figura 36 – Acurácia obtida para subconjunto da base de dados Flirk Flashion 10.000 com as 20 maiores classes, com 2 mil imagens escolhidas aleatoriamente.



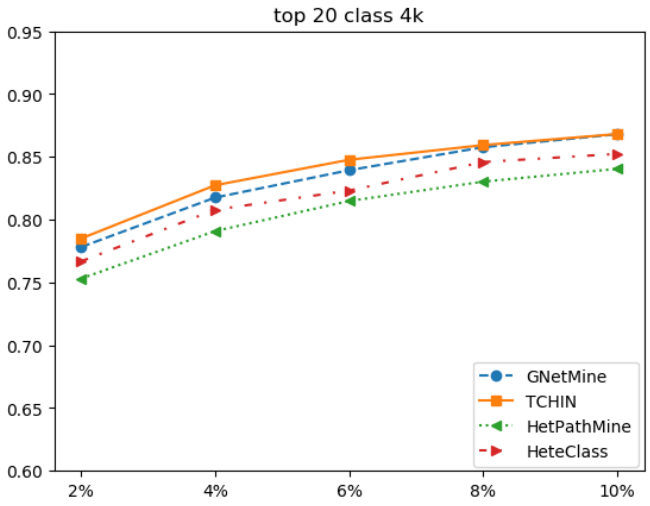

Fonte: Elaborada pelo autor.

Tabela 13 – Flirk Flashion 10.000.

| **Método \ Rotulados** | 2% | 4% | 6% | 8% | 10% |
|---|---|---|---|---|---|
| GNetMine | 0.724 | 0.7817 | 0.7909 | 0.8303 | 0.8459 |
| TCHN | 0.7314 | 0.7887 | 0.8032 | 0.8323 | 0.8494 |
| HetPathMine | 0.7109 | 0.7631 | 0.7842 | 0.7929 | 0.8206 |
| HeteClass | 0.724 | 0.7703 | 0.7887 | 0.8079 | 0.8274 |

Fonte: Dados da pesquisa.

Figura 37 – Acurácia obtida para subconjunto da base de dados Flirk Flashion 10.000 com as 20 maiores classes.




Fonte: Elaborada pelo autor.

Tabela 14 – Resultados obtidos para subconjunto da base de dados Flirk Flashion 10.000 com as 20 maiores classes.

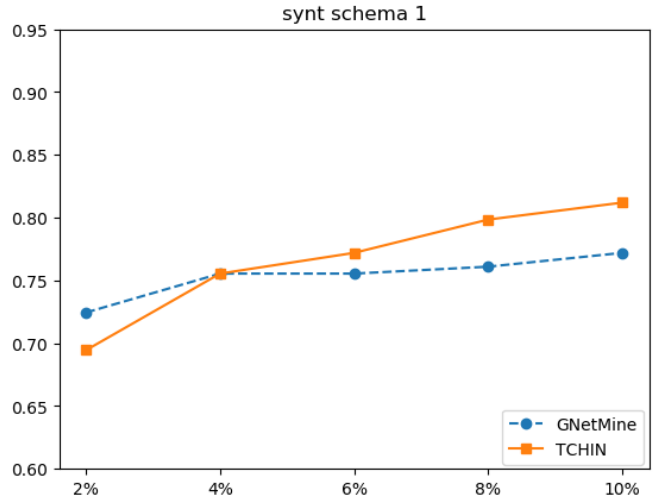
| **Método \ Rotulados** | 2% | 4% | 6% | 8% | 10% |
|---|---|---|---|---|---|
| GNetMine | 0.7782 | 0.8176 | 0.8394 | 0.8577 | 0.868 |
| TCHN | 0.785 | 0.8274 | 0.8477 | 0.8594 | 0.8682 |
| HetPathMine | 0.7531 | 0.7909 | 0.8149 | 0.8303 | 0.8405 |
| HeteClass | 0.7668 | 0.8079 | 0.8232 | 0.8459 | 0.8524 |

Fonte: Dados da pesquisa.

### 6.3.4 Resultados obtidos para HINs sintéticas

Para as duas redes sintéticas geradas com a ferramenta HNOC, dados que os meta-caminhos são gerados com base em significados semânticos o que não se tem em tais redes, o método TCHN foi comparado apenas com o método GNetMine. Nas figuras 38 e 39 são mostrados os resultados obtidos de acurácia para tais métodos sobre as redes sintéticas descritas anteriormente, os mesmos resultados são apresentados de forma numérica nas tabelas 15 e 16.

Figura 38 – Acurácia obtida para a base de dados sintética k-partida.



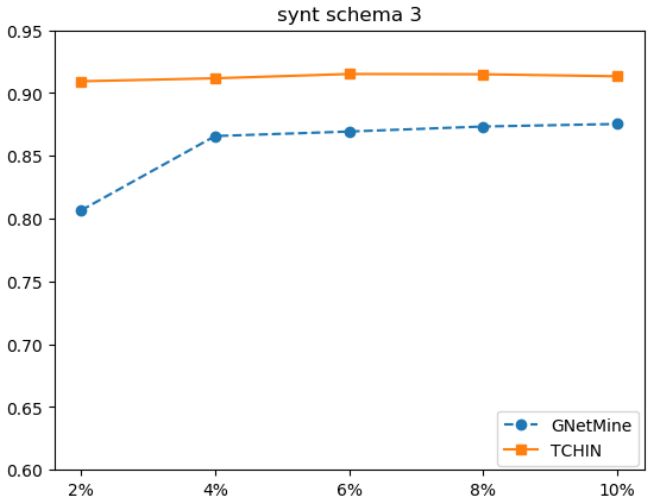

Fonte: Elaborada pelo autor.

Tabela 15 – Resultados obtidos para a base de dados sintética k-partida com quatro camadas de tamanhos $(200 \times 500 \times 1000 \times 20)$.

| **Método \ Rotulados** | 2% | 4% | 6% | 8% | 10% |
|---|---|---|---|---|---|
| GNetMine | 0.7245 | 0.7555 | 0.7555 | 0.761 | 0.772 |
| TCHN | 0.6945 | 0.7555 | 0.772 | 0.7985 | 0.812 |

Fonte: Dados da pesquisa.

Figura 39 – Acurácia obtida para a base de dados sintética heterogênea.




Fonte: Elaborada pelo autor.

Tabela 16 – Resultados obtidos para a base de dados sintética heterogênea três tipos de tamanhos ($500 \times 1000 \times 1000$).

| **Método \ Rotulados** | 2% | 4% | 6% | 8% | 10% |
|---|---|---|---|---|---|
| GNetMine | 0.8066 | 0.8658 | 0.8694 | 0.8734 | 0.8754 |
| TCHN | 0.9094 | 0.9118 | 0.9152 | 0.915 | 0.9134 |

Fonte: Dados da pesquisa.

Como pode ser observado, o método TCHN obteve para tais rede resultados de acurácia bastante superiores em comparação ao método GNetMIne. Acreditamos que o método proposto se mostra superior para estes casos devido ao fato de a modelagem do método ser baseada na suposição de que a distribuição do rótulo se aproximar de uma distribuição de probabilidades, assim rede onde esta suposição é verdadeira o método tem maior potencial de ser bem sucedido. Desta forma, como as redes sintéticas são geradas com base em uma distribuição de probabilidade, suas distribuições de rótulos tende a favorecer o desempenho do método TCHN.

CAPÍTULO

# 7

# CONCLUSÃO

Cada vez mais utilizadas para a modelagem grandes conjunto de dados em diversos domínios de aplicação, as redes heterogêneas de informação vem se confirmando como estruturas relevantes para uma representação mais precisa de dados cada vez mais complexos encontrados (PEMBERTON, 2008; SINGHAL, 2012; SHI *et al.*, 2017). Tal representação se mostra capaz de extrair e organizar várias entidades e relações presente nos dados, formando redes com esquemas complexos que podem conter várias camadas de vértices ou arestas. Assim o estudos das estruturas e desenvolvimento de uma sólida fundamentação teórica e desenvolvimento de técnicas com tal base se faz cada dia mais necessária.

De fato, a área de pesquisa em redes heterogêneas de informação(HIN) vem crescendo devido sua capacidade de representar dados de mundo real, bem como os bons resultados em tarefas de aprendizado de máquina. Neste contexto, neste trabalho foi proposto o método TCHN de classificação transdutiva de HIN. Tal método tem como diferencial a utilização da divergência KL como medida de similaridade para a regularização da propagação de informação pelos vetores de informação. Esta modelagem tem como motivação o fato de tal métrica ser mais apropriada para a regularização de distribuições de probabilidade, considerando que a distribuição de informação na rede deve se comporta de tal maneira (FALEIROS; ROSSI; LOPES, 2017).

Como apresentado no Capítulo 6, experimentos comprovam que o método TCHN produz resultados comparáveis ou até mesmo superiores aos métodos representativos da área, confirmando assim sua efetividade para a classificação em diversos cenários. Além disso, a complexidade deduzida para o método TCHN para redes esparsas mostra-se bastante atrativa para a aplicação em dados de mundo real, que como já discutido possuem naturalmente características heterogêneas.

Vale a pena observar que, os métodos analisados apresentaram variações em seus comportamentos de acordo com características e estruturas das redes fornecidas nos experimentos. Por exemplo, os métodos baseados em meta-caminhos apresentaram uma melhor acurácia para redes geradas com base no conjunto de dados DBLP usando apenas vértices do tipo autor como elementos rotulado de entrada. Por outro lado, quando utilizados rótulos de conferências para os métodos de propagação geral, este superaram os resultados dos métodos baseados em meta-caminhos. Já para o o conjunto de dados Fashion Flirck os métodos baseados em meta-caminhos apresentaram uma performance inferior, mostrando assim uma

sensibilidade à estrutura da rede de entrada.

Além disso, nos experimentos utilizando redes sintéticas foi possível observar uma superioridade do método TCHN sobre GNetMine. A observação de tais dependências dos resultados, motiva um estudo mais aprofundados do impacto das características das redes sobre o resultado dos métodos, onde podemos citar como características de maior potencial: a esparsidade da rede, esquema e outras medidas. Assim, um estudo neste nível pode ser capaz de recomendar melhores métodos de acordo com cada tipo de HIN, considerando acurácia e custos computacionais tolerados.

Além do desenvolvimento do método TCHN, como parte das demandas da área que impactaram neste trabalho, foi desenvolvida uma ferramenta de geração de redes heterogêneas sintéticas, camada HNOC. Tal ferramenta foi desenvolvida em parceria com outros pesquisadores do laboratório de pesquisa LABIC, sendo implementada com base na ferramenta BNOC de geração de redes bipartidas. A ferramenta HNOC é capaz de gerar rede heterogêneas sintéticas com diferentes topologias e dispersão baseadas em um conjunto de parâmetros, de forma fácil e não custosa. Esta se mostrou bastante útil para a validação do método TCHN, pois com seu uso, foi possível a comparação das técnicas em redes com diferentes características com um custo bastante reduzido se comparado com o possível custo de levantamento de redes semelhantes com base em dados reais.

Resumidamente, neste trabalho foram realizados estudos sobre as construção de redes heterogêneas e a classificação transdutiva de dados utilizando tal representação. Cuja ambas as áreas apresentam diversas demandas de desenvolvimento. Podemos citar como principais contribuições desta tese:

- O desenvolvimento do método TCHN de classificação transdutiva de redes heterogêneas de informação;
- O desenvolvimento da ferramenta HNOC para a geração de redes heterogêneas sintéticas.
- E as publicações correspondentes, já citadas.

## 7.1 Limitações e Trabalhos Futuros

Como levantado em diversos pontos do texto, apesar das contribuições deste trabalho, existem ainda diversos temas a serem explorados dentro da área de pesquisa desta tese em redes heterogêneas de informação. Dentro da modelagem de HIN existe ainda uma grande demanda por métricas significativas que expressem de maneira intuitiva características relacionadas ao esquema que define uma HIN, como por exemplo a importância das camadas de vértices e arestas. Um estudo mais aprofundado e desenvolvimento de boas métricas pode levar a formas automáticas para a modelagem de esquemas de rede, ou reestruturação de redes já existentes, por exemplo eliminando camadas redundantes, ou então sugerir um pré-processamento do conteúdo quando constatada a presença excessiva de ruído, o que pode ser feito com o uso de *embeddings* ou outras ferramentas. Ao mesmo tempo, tais métricas também podem ser utilizadas na seleção de meta-caminhos para métodos baseados nestes.

Como mostrado nos resultados o método TCHN se mostra com desempenho entre os melhores se comparado com métodos relevantes da área de classificação transdutiva de HIN, e alguns caso superando

os outros métodos. Dentre os casos onde o método TCHN se mostra superior está o caso de redes heterogêneas sintéticas, geradas pela ferramenta HNOC. Para melhor levantar as características de rede para as quais o método mostra melhor desempenho, um estudo aprofundados utilizando uma variedade maior de redes sintética pode fornecer respostas interessantes e perspicácia de novas direções de melhoria para o método.

Uma possível melhoria já levantada no texto é o uso de pesos que regularizem a propagação de informação entre as camadas, tipos que possuam melhores características para otimizar a classificação podem receber um maior peso ao propagar seus rótulos aos seus vizinhos, por exemplos no caso das HIN construídas com base no conjunto de dados DBLP, a camada de vértices do tipo conferência possuem sua classificação bastante direta pela área de pesquisa e seus conjuntos de vizinhos disjuntos de forma a possuir grande potencial para segregação das classe. Estas são apenas algumas possibilidades de melhoraria e desenvolvimento dentro de uma área recente e com grande crescimento.

# REFERÊNCIAS

ABDULLA, M. **On the fundamentals of stochastic spatial modeling and analysis of wireless networks and its impact to channel losses**. Tese (Doutorado) — Concordia University, 2012. Citado na página 31.

AMINI, M.-R.; USUNIER, N. **Learning with Partially Labeled and Interdependent Data**. [S.l.]: Springer, 2015. Citado nas páginas 46 e 48.

ANGELOVA, R.; KASNECI, G.; SUCHANEK, F. M.; WEIKUM, G. Graffiti: node labeling in heterogeneous networks. In: ACM. **Proceedings of the 18th international conference on World wide web**. [S.l.], 2009. p. 1087–1088. Citado na página 40.

ANGELOVA, R.; KASNECI, G.; WEIKUM, G. Graffiti: graph-based classification in heterogeneous networks. **World Wide Web**, Springer, v. 15, n. 2, p. 139–170, 2012. Citado nas páginas 27, 39, 40 e 79.

BANGCHAROENSAP, P.; MURATA, T.; KOBAYASHI, H.; SHIMIZU, N. Transductive classification on heterogeneous information networks with edge betweenness-based normalization. In: ACM. **Proceedings of the Ninth ACM International Conference on Web Search and Data Mining**. [S.l.], 2016. p. 437–446. Citado nas páginas 39, 41, 54, 55 e 56.

BREVE, F.; ZHAO, L.; QUILES, M.; PEDRYCZ, W.; LIU, J. Particle competition and cooperation in networks for semi-supervised learning. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 24, n. 9, p. 1686–1698, 2012. Citado na página 40.

ELMASRI, R.; NAVATHE, S. **Fundamentals of database systems**. [S.l.]: Addison-Wesley Publishing Company, 2010. Citado na página 40.

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 25, 27, 37, 40, 41, 63, 65, 67, 70 e 99.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, JSTOR, p. 35–41, 1977. Citado na página 55.

GOLUB, G. H.; LOAN, C. F. V. **Matrix computations**. [S.l.]: JHU press, 2012. v. 3. Citado na página 40.

GUPTA, M.; KUMAR, P.; BHASKER, B. A new relevance measure for heterogeneous networks. In: SPRINGER. **International Conference on Big Data Analytics and Knowledge Discovery**. [S.l.], 2015. p. 165–177. Citado nas páginas 37, 39 e 59.

____. Heteclass: A meta-path based framework for transductive classification of objects in heterogeneous information networks. **Expert Systems with Applications**, Elsevier, v. 68, p. 106–122, 2017. Citado nas páginas 39, 41, 59 e 60.

HAVELIWALA, T. H. Topic-sensitive pagerank. In: ACM. **Proceedings of the 11th international conference on World Wide Web**. [S.l.], 2002. p. 517–526. Citado na página 60.

HIMMELSTEIN, D. S.; BARANZINI, S. E. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. **PLoS computational biology**, Public Library of Science, v. 11, n. 7, p. e1004259, 2015. Citado nas páginas 31 e 34.

HWANG, T.; KUANG, R. A heterogeneous label propagation algorithm for disease gene discovery. In: SIAM. **Proceedings of the 2010 SIAM International Conference on Data Mining**. [S.l.], 2010. p. 583–594. Citado nas páginas 26 e 41.

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 nas páginas 39, 41, 51, 52 e 86.

JOLLIFFE, I. **Principal component analysis**. [S.l.]: Springer, 2011. Citado na página 40.

LAO, N.; COHEN, W. W. Relational retrieval using a combination of path-constrained random walks. **Machine learning**, Springer, v. 81, n. 1, p. 53–67, 2010. Citado na página 35.

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, 1999. Citado na página 40.

LEY, M. The dblp computer science bibliography: Evolution, research issues, perspectives. In: SPRINGER. **International symposium on string processing and information retrieval**. [S.l.], 2002. p. 1–10. Citado na página 39.

LONI, B.; CHEUNG, L. Y.; RIEGLER, M.; BOZZON, A.; GOTTLIEB, L.; LARSON, M. Fashion 10000: an enriched social image dataset for fashion and clothing. In: ACM. **Proceedings of the 5th ACM Multimedia Systems Conference**. [S.l.], 2014. p. 41–46. Citado na página 87.

LUO, C.; GUAN, R.; WANG, Z.; LIN, C. Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks. In: SPRINGER. **European Conference on Information Retrieval**. [S.l.], 2014. p. 210–221. Citado nas páginas 37, 39, 41, 57 e 58.

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. 573, 2017. Citado na página 26.

MCCALLUM, A.; NIGAM, K. *et al.* A comparison of event models for naive bayes text classification. In: CITESEER. **AAAI-98 workshop on learning for text categorization**. [S.l.], 1998. v. 752, n. 1, p. 41–48. Citado na página 40.

ÖZSU, M. T. A survey of rdf data management systems. **Frontiers of Computer Science**, Springer, v. 10, n. 3, p. 418–432, 2016. Citado na página 43.

PEMBERTON, S. **Fellini—Satyricom (1969)—Synopsis. IMDb-The Internet Movie Database**. [S.l.]: IMDb, 2008. Citado nas páginas 42 e 99.

PIO, G.; SERAFINO, F.; MALERBA, D.; CECI, M. Multi-type clustering and classification from heterogeneous networks. **Information Sciences**, Elsevier, v. 425, p. 107–126, 2018. Citado nas páginas 25, 37 e 40.

ROSSI, R. G.; LOPES, A. de A.; FALEIROS, T. de P.; REZENDE, S. O. Inductive model generation for text classification using a bipartite heterogeneous network. **Journal of Computer Science and Technology**, Springer, v. 29, n. 3, p. 361–375, 2014. Citado na página 40.

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 26 e 40.

____. Using bipartite heterogeneous networks to speed up inductive semi-supervised learning and improve automatic text categorization. **Knowledge-Based Systems**, Elsevier, v. 132, p. 94–118, 2017. Citado na página 34.

SERAFINO, F.; PIO, G.; CECI, M. Ensemble learning for multi-type classification in heterogeneous networks. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 30, n. 12, p. 2326–2339, 2018. Citado na página 40.

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 25, 37 e 41.

SHI, C.; KONG, X.; HUANG, Y.; PHILIP, S. Y.; WU, B. Hetesim: A general framework for relevance measure in heterogeneous networks. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 26, n. 10, p. 2479–2492, 2014. Citado na página 37.

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 32 e 99.

SHI, C.; ZHOU, C.; KONG, X.; YU, P. S.; LIU, G.; WANG, B. Heterecom: a semantic-based recommendation system in heterogeneous networks. In: ACM. **Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2012. p. 1552–1555. Citado nas páginas 25 e 37.

SINGHAL, A. Introducing the knowledge graph: things, not strings. **Official google blog**, v. 5, 2012. Citado nas páginas 42, 43 e 99.

SUN, Y.; HAN, J. Mining heterogeneous information networks: principles and methodologies. **Synthesis Lectures on Data Mining and Knowledge Discovery**, Morgan & Claypool Publishers, v. 3, n. 2, p. 1–159, 2012. Citado nas páginas 25, 31, 35 e 37.

____. Mining heterogeneous information networks: a structural analysis approach. **Acm Sigkdd Explorations Newsletter**, ACM, v. 14, n. 2, p. 20–28, 2013. Citado na página 34.

SUN, Y.; HAN, J.; YAN, X.; YU, P. S.; WU, T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. **Proceedings of the VLDB Endowment**, Citeseer, v. 4, n. 11, p. 992–1003, 2011. Citado nas páginas 36, 37 e 58.

SUN, Y.; HAN, J.; ZHAO, P.; YIN, Z.; CHENG, H.; WU, T. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: ACM. **Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology**. [S.l.], 2009. p. 565–576. Citado nas páginas 25 e 63.

SUN, Y.; NORICK, B.; HAN, J.; YAN, X.; YU, P. S.; YU, X. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM, v. 7, n. 3, p. 11, 2013. Citado na página 36.

SZUMMER, M.; JAAKKOLA, T. Partially labeled classification with markov random walks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2002. p. 945–952. Citado nas páginas 40 e 49.

USLU, T.; MEHLER, A. Polyviz: a visualization system for a special kind of multipartite graphs. **Proceedings of the IEEE VIS 2018**, 2018. Citado nas páginas 17 e 84.

VALEJO, A.; GOES, F.; ROMANETTO, L. M.; OLIVEIRA, M. C. F.; LOPES, A. A. A benchmarking tool for the generation of bipartite network models with overlapping communities. **Knowledge and information systems, accepted paper**, 2019. Citado nas páginas 26, 79 e 83.

WU, M.; SCHÖLKOPF, B. Transductive classification via local learning regularization. In: **Artificial Intelligence and Statistics**. [S.l.: s.n.], 2007. p. 628–635. Citado na página 45.

YANG, Y.; CHAWLA, N.; SUN, Y.; HANI, J. Predicting links in multi-relational and heterogeneous networks. In: IEEE. **2012 IEEE 12th international conference on data mining**. [S.l.], 2012. p. 755–764. Citado na página 81.

YIN, X.; HAN, J.; YU, P. S. Linkclus: efficient clustering via heterogeneous semantic links. In: VLDB ENDOWMENT. **Proceedings of the 32nd international conference on Very large data bases**. [S.l.], 2006. p. 427–438. Citado na página 37.

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 26, 38, 39, 46 e 51.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado na página 51.


ICMC
USP
SÃO CARLOS