Pós-Graduação em Ciência da Computação

PREDIÇÃO TEMPORAL DE LINKS BASEADA NA EVOLUÇÃO DE TRÍADES

Por

HUGO NEIVA DE MELO

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE

2016

HUGO NEIVA DE MELO

"PREDIÇÃO TEMPORAL DE LINKS BASEADA NA EVOLUÇÃO DE TRÍADES"

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: *Ricardo Bastos Cavalcante Prudêncio*

RECIFE

2016

Catalogação na fonte
Bibliotecário Jefferson Luiz Alves Nazareno CRB 4-1758

M528p Melo, Hugo Neiva de.
Predição temporal de links baseada na evolução de tríades / Hugo Neiva de Melo – 2016.
115.: fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, Recife, 2016.
Inclui referências.

1. Inteligência artificial. 2. Redes sociais on-line. 3. Predição de links. I. Prudêncio, Ricardo Bastos Cavalcante. (Orientador). II. Titulo.

006.3 CDD (22. ed.) UFPE-MEI 2017-009

Hugo Neiva de Melo

Predição Temporal de Links Baseada na Evolução de Tríades

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 25/08/2016

BANCA EXAMINADORA

_____

Prof. Dr. Leandro Maciel Almeida
Centro de Informática / UFPE

_____

Prof. Dr. André Câmara Alves do Nascimento
Departamento de Estatística e Informática/UFRPE

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática /UFPE
(**Orientador**)

Dedico esta dissertação a toda a minha família, em especial meus pais e meu avô Ubirajara, aos amigos e meu orientador, que me ajudaram durante todo esse tempo.

Agradecimentos

Gostaria de agradecer, primeiramente, aos meus pais, Ubirajara e Betânia, por todo o apoio dado e pela compreensão durante esses dois anos de mestrado, quando tive que deixar algumas coisas um pouco de lado para poder me dedicar. Agradeço também aos meus avós, em especial ao meus avôs Ubirajara e Severino, que queriam presenciar a conclusão do meu mestrado, porém a vida não lhes permitiu.

Um grande obrigado também ao meu orientador, Ricardo Prudêncio, que confiou na minha capacidade mesmo passando por vários problemas durante esses dois anos de mestrado e durante a confecção deste trabalho.

Agradeço também aos meus amigos, Arthur Ramos, Davi Duarte, David Hulak e Tullio Lucena por sempre estarem por perto e por não me deixar desanimar, ajudando inclusive com algumas dicas para o mestrado. Também quero citar meus colegas de trabalho, Marcelo, Fernando, Leo, Evertons e Miotto por me ajudarem sempre que o tempo apertava, enquanto trabalhei com eles no projeto CIn-Motorola. Gostaria de agradecer também a Arley, por me ajudar em várias cadeiras e com o mestrado.

Por fim, termino agradecendo o apoio de todos que de alguma forma colaboraram comigo para que eu pudesse concluir o presente trabalho. Sem todos vocês, não estaria onde estou hoje.

Resumo

Atualmente, com o crescimento da área de inteligência artificial e devido à necessidade do estudo das redes sociais no mundo virtual, ficou em evidência a importância da análise dessas redes. Existem vários tipos de problemas que podem ser levantados nesse sentido, entre eles, o problema de Predição de Links dentro de uma rede social, tarefa associada à Análise de Redes Sociais. Atualmente as abordagens buscam observar algum tipo de padrão na rede, sendo esses padrões estruturais, de similaridades entre os indivíduos, estatísticos, até modelos mais complexos, como padrões temporais. Este trabalho tem como objetivo propor uma nova metodologia temporal, chamada de Predição Temporal de Links baseada na Evolução de Tríades, de modo a prover uma solução mais satisfatória e computacionalmente viável para o problema de Predição de Links. Para isto, foi criado um novo modelo temporal de dados, chamado de Tensor de Transições de Tríades, que serve de base para o cálculo de modelos de predição temporal estatística de séries temporais. Este modelo foi concebido a partir da análise das principais abordagens vistas na literatura e identificação das suas vantagens e limitações. Os resultados obtidos mostraram que, em relação às abordagens de trabalhos relacionados, houve uma considerável melhora na qualidade da predição ao utilizar o modelo criado.

Palavras-chave: Inteligência Computacional. Análise de Redes Sociais. Predição Temporal de Links. Tensor de Transições de Tríades. Predição Temporal.

Abstract

Nowadays, with the development of artificial intelligence and the need to study virtual social networks, the importance of the analysis of such networks has grown. There are many problems that arise when studying these networks, including the Link Prediction problem in a social network, a task associated with Social Network Analysis. The current state-of-the-art on Link Prediction seeks to find a hidden pattern in the network, including structural patterns, similarities and statistical characteristics and evolving to more complex models, like temporal patterns. This work aims to create a new temporal method, called Temporal Link Prediction based on Triads Evolution, which provides a more satisfactory and efficient solution for the Link Prediction problem. To achieve this goal, a new temporal data model, the Triad Transition Tensor, was created and used as a source to compute temporal forecasting statistic models based on time series. This method was conceived from a wide analysis of the state-of-the-art of the Link Prediction methods and identifying it's advantages and limitations. The results in this work show that, compared to other methods found in related works, there was a considerable improvement in the quality of the predictions when using the proposed method.

Keywords: Computational Intelligence. Social Network Analysis. Temporal Link Prediction. Triad Transition Tensor. Temporal Forecasting.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Acrônimos

Sumário

1

Introdução

Este capítulo tem por objetivo apresentar algumas informações iniciais sobre o presente trabalho, como o contexto e a área de estudo em que o trabalho está inserido, as motivações para se resolver o problema abordado, ou seja, a Predição de Links, os objetivos que serão alcançados com o desenvolvimento deste trabalho, bem como suas contribuições. Por fim, será mostrado como os capítulos seguintes desta dissertação estão organizados.

1.1 Contexto

O estudo sobre redes sempre foi de interesse de várias áreas do conhecimento, entre elas a matemática e sociologia. Na matemática, o problema das pontes de Leonard Euler é um exemplo conhecido. O problema relata uma cidade cortada por rios e conectada por pontes. Desejava-se saber se existia um caminho que passasse por todas as pontes da cidade, passando por cada ponte apenas uma vez (NEWMAN; BARABASI; WATTS, 2011).

Além da rede descrita pelo problema de Euler, que é formada pelas seções da cidade (indivíduos da rede) e as pontes que as conectam (conexões entre os indivíduos), existem outros tipos de redes reais. Essas redes podem ser constituidas por diversos tipos de membros, como seres humanos, animais ou até objetos inanimados (é o caso do problema de Euler) e tipos de conexões diferentes. Em cada rede, as conexões podem ter conotações diferentes, por exemplo, em uma comunidade, os indivíduos da rede são os moradores e as conexões podem ser de vizinhança ou amizade (NEWMAN; BARABASI; WATTS, 2011).

A sociologia é uma área que estuda extensivamente as redes sociais do mundo real. Além das redes sociais, as redes biológicas também são objetos de estudo, como redes de trasmissão de doenças e reações entre proteínas. A principal característica comum entre essas redes é que elas ocorrem naturalmente e evoluem, tipicamente, de modo não planejado e descentralizado (NEWMAN; BARABASI; WATTS, 2011). Redes que apresentam essas características são chamadas de redes complexas. Além disso, redes complexas passam por processos que, ao longo do tempo, adicionam ou removem conexões de tais redes.

Com a criação e a popularização dos computadores e, posteriormente, das redes de *internet*, nas últimas décadas, novos tipos de redes começaram a surgir. Essas redes são chamadas

de virtuais, pois as interações entre seus membros ocorre por meios eletrônicos, como um computador. Isso se deu porque o usuário, à medida que a *World Wide Web* (WWW) evoluiu, deixou de compor o polo passivo (apenas visitar páginas da *internet*) passando a ser ativo, ou seja, começou a criar seu próprio conteúdo. Sites, mensagens e outros tipos de dados fizeram com que a *web* se tornasse um ambiente bem mais dinâmico onde qualquer usuário pode participar criando ou mudando conteúdo (CORMODE; KRISHNAMURTHY, 2008).

A partir disso, as redes sociais virtuais foram criadas e consolidadas, como, por exemplo, o Facebook[1] ou o Twitter[2]. Apesar de populares, as redes virtuais são apenas uma pequena parcela dos vários tipos de redes existentes no mundo real. Entre outros exemplos, pode-se citar as redes onde os indivíduos ou componentes não são seres humanos, a exemplo das colônias de insetos ou até mesmo das redes que representam a porosidade do solo. As redes virtuais são uma consequência do avanço da tecnologia e do comportamento humano de coletividade.

A Análise de Redes Sociais (ARS) (SA; PRUDENCIO, 2011) está focada em investigar padrões de comportamento que ditam como indivíduos se relacionam, além de tentar esclarecer algumas questões sobre a formação ou extinção de relacionamentos. Estudar o comportamento dos usuários dentro dessas redes é importante em várias aplicações, como *marketing* (análise de compras), segurança (redes terroristas), entre outros, pois seus resultados permitem tirar conclusões estratégicas baseadas em informações extraídas dessas redes (WASSERMAN; FAUST, 1994).

A popularização das redes sociais virtuais facilitou o desenvolvimento da ARS, já que o custo para adquirir os dados de redes virtuais é bem menor, visto que a coleta dos dados é feita de forma automática, e os dados são mais completos (as redes virtuais possuem milhares e até milhões de usuários). Essa popularização também permitiu análises mais complexas e a obtenção de resultados mais sólidos se comparado a levantamentos manuais feitos em outras redes. Com isso, redes sociais têm atraído a atenção da comunidade acadêmica e de empresas privadas, que passaram a lucrar com a obtenção de dados importantes de tais redes (COOKE, 2006; Lü; ZHOU, 2011).

A partir dos estudos realizados (BARABASI; ALBERT, 1999), pôde-se constatar que as redes sociais, assim como boa parte das redes governadas por regras não determinísticas (redes complexas), apresentam comportamento naturalmente evolutivo, isto é, ao longo do tempo, membros e conexões tendem a aparecer ou desaparecer, conforme um processo que não pode ser determinado com exatidão, ante a própria essência da rede complexa. Por tal motivo, as pesquisas nessa área tornam-se de extrema importância, já que são necessárias para o entendimento dos processos pelos quais passam essas redes durante sua evolução, de modo a determinar com mais exatidão como elas estarão configuradas no futuro.

[1] www.facebook.com
[2] www.twitter.com

1.2 Motivação

A determinação de todas as conexões ou as conexões faltantes dentro de uma rede, no futuro, é denominada como problema de Predição de Links, uma subárea da Mineração de Links (ML). A predição de novos relacionamentos entre indivíduos de uma sociedade é, em suma, feita pela análise das relações já existentes, a partir do cálculo de métricas, descoberta de padrões e extração de informações (GETOOR; DIEHL, 2005). A Predição de Links é um problema dentre vários na área de ARS e tem sido tradicionalmente definido como: "Dado um estado da rede no tempo t, busca-se predizer que conexões serão adicionadas a rede em um tempo futuro $t + 1$"(LIBEN-NOWELL; KLEINBERG, 2003). Ela consiste em basicamente duas perspectivas. A primeira é sob o aspecto de conexões faltantes: o problema é detectar conexões implícitas dentro da rede e utilizá-las para predizer relacionamentos. A segunda perspectiva é sobre temporalidade da predição, ou seja, como os *links* da rede evoluem ao longo do tempo é levado em consideração para a predição de conexões em um futuro próximo. A segunda vertente do problema considera a rede como uma estrutura evolutiva, procurando identificar como a rede estará estruturada no futuro.

Majoritariamente, os trabalhos, por exemplo, os que exploram padrões estruturais da rede, baseiam-se na primeira perspectiva (Lü; ZHOU, 2011), isto é, utilizam dados das redes em um determinado período de tempo, e, em sua maioria reunidos e representados por uma única estrutura de dados. Em posse desse conjunto, é possível calcular diferentes métricas, como as de similaridade, as estruturais ou as baseadas em métodos estatísticos, que servem de fonte exploratória para a predição de conexões faltantes. Essas abordagens são chamadas de estáticas, pois se entende, pela modelagem, que uma rede social mantém a mesma estrutura durante todo o período considerado (SA; PRUDENCIO, 2011).

Contudo, as redes mudam ao longo do tempo, e a temporalidade dessas mudanças é pouco explorada em estratégias para o problema de Predição de Links (Lü; ZHOU, 2011; SANTORO et al., 2011). Uma maneira simples de inserir um contexto de tempo é dividir a rede em períodos de mesmo tamanho, calcular *scores* (a relevância de um *link*, presente ou não na rede) a partir de alguma métrica, como, por exemplo, padrões estruturais da rede, e atribuir um peso a cada *score* obtido. *Scores*, sendo os relacionados a períodos mais antigos com um peso menor, enquanto os mais recentes com um peso maior. Essa abordagem dá mais importância aos *scores* calculados a partir de dados mais recentes. De modo geral, estratégias com essas características são chamadas de temporais e tentam incorporar, de alguma maneira, a temporalidade das mudanças que ocorrem ao longo do tempo dentro da rede. Apesar de aumentar a complexidade dos algoritmos, estratégias temporais podem prover melhores resultados se comparadas às estáticas (HUANG; LIN, 2009; POTGIETER et al., 2009; TYLENDA; ANGELOVA; BEDATHUR, 2009; BRINGMANN et al., 2010). A predição temporal de *links* é uma linha de pesquisa em crescimento e ainda não consolidada, com muitas vertentes a serem exploradas.

1.3 Objetivos

Tendo por base a problemática apresentada, o trabalho objetiva a proposição de uma nova estratégia capaz de coletar informações temporais e fazer uso destas para solucionar o problema de Predição de Links para uma determinada rede social. Durante o desenvolvimento deste trabalho, o problema principal foi dividido em subproblemas menores, de modo que ao resolver todos, tem-se uma solução mais satisfatória para a Predição de Links. São objetivos:

- Realizar um levantamento dos estudos mais relevantes e recentes sobre a Predição de Links, de modo a entender como funcionam os principais métodos utilizados para resolver o problema;
- Realizar um estudo sobre as predições temporais para entender como funcionam as representações de tempo mais conhecidas, atentando para as suas vantagens e desvantagens;
- Identificar padrões de evoluções dentro das redes sociais analisadas, a fim de escolher o melhor método estatístico preditivo a ser incorporado ao algoritmo;
- Identificar vantagens, desvantagens e avaliar a estratégia criada, comparando-a com as metodologias mais tradicionais e trabalhos relacionados;
- Otimizar o algoritmo criado a fim de conhecer as limitações e torná-lo utilizável em redes sociais de grande escala (com muitos indivíduos).

1.4 Contribuições

Para alcançar os objetivos enumerados, foi, inicialmente, feito um levantamento das principais abordagens encontradas na literatura. A partir do conjunto dessas abordagens, foram analisadas as vantagens e limitações de cada uma. Em especial, foram verificadas duas limitações principais onde: a primeira diz respeito à ausência de uma modelagem temporal dos dados e a segunda, que várias métricas, temporais ou não, são derivadas de cálculos de características topográficas (ou outras métricas estáticas, como similaridade) da rede.

Com essas limitações em mãos, foi proposto um método temporal e elaborado um algoritmo que descreve, estatísticamente, como ocorre a evolução de estruturas simples e não triviais que podem ser observadas em um grafo: as tríades. Uma tríade, como será visto posteriormente, é um subgrafo composto de três vértices quaisquer do grafo. A tríade foi escolhida como objeto de observação por já ser abordada em diversos trabalhos (LESKOVEC; HUTTENLOCHER; KLEINBERG, 2010; DAVIS; LICHTENWALTER; CHAWLA, 2011; JUSZCZYSZYN; MUSIAL; BUDKA, 2011; DONG et al., 2012).

O modelo temporal proposto é baseado no trabalho desenvolvido por JUSZCZYSZYN; MUSIAL; BUDKA (2011). Esse trabalho consiste em identificar todas as tríades em uma rede e

como elas mudam ao longo de um certo período de tempo observado, para predizer os estados dessas tríades (em termos das conexões entre os vértices dessas tríades) no futuro. Para tal, criam uma estrutura chamada *Matriz de Transição de Tríades* (TTM) que armazena as probabilidades dessas mudanças de estado ocorrerem e calculam valores que indicam o quão provável é de que uma tríade qualquer mudar, posteriormente, para um estado diferente.

Porém, algumas limitações foram identificadas: a primeira diz respeito à TTM, pois pode apenas armazenar informações sobre um período de tempo. Essa primeira limitação está relacionada à primeira perspectiva do problema de Predição de Links, citada na Seção 1.2, Página 15. A segunda limitação refere-se à predição temporal sugerida por JUSZCZYSZYN; MUSIAL; BUDKA (2011), onde são calculadas as TTMs referentes às diversas observações no tempo, mas o valor final é a média aritmética dos valores intermediários encontrados, que não leva em consideração qualquer padrão evolutivo visto durante as observações. As contribuições deste trabalho consistem em implementar melhorias sobre o algoritmo de JUSZCZYSZYN; MUSIAL; BUDKA (2011), explorando as limitações encontradas, para criar um modelo mais completo e capaz de obter melhores resultados de predições. Algumas das contribuições do trabalho são enumeradas a seguir:

1. Proposição de uma análise temporal para a predição de relacionamentos;
2. Elaboração de um modelo flexível, que permite uma fácil adaptação para o uso de outras técnicas para a Predição de Links;
3. Comparação de resultados com o trabalho desenvolvido por JUSZCZYSZYN; MUSIAL; BUDKA (2011), um modelo temporal, pois foi utilizado como base para o modelo proposto. Além desse, foram escolhidas outras duas métricas estáticas que alcançam um bom desempenho em trabalhos relacionados, para fins de comparação;
4. Obtenção de ótimos resultados, avaliados através de duas métricas robustas, as curvas *Receiver-Operating Characteristic* (ROC) e *Precision-Recall* (PR) (YANG; LICHTENWALTER; CHAWLA, 2015).

Estas contribuições serão melhor detalhadas no Capítulo 6.

1.5 Organização da Dissertação

O trabalho está disposto em seis capítulos, divididos de acordo com os objetivos levantados, da seguinte maneira:

- **Capítulo 2**: Elucida alguns conceitos fundamentais necessários a compreensão da área de ARS. Estes conceitos servem de base e/ou justificativa para a adoção das estratégias para resolver o problema;

- **Capítulo 3**: Define o escopo do problema de Predição de Links abordado e investiga mais detalhadamente as estratégias utilizadas na bibliografia, citando algumas vantagens e limitações de cada uma;
- **Capítulo 4**: Mostra, passo a passo, como foi elaborada a estratégia proposta por este trabalho e o algoritmo final resultante, justificando as escolhas feitas, além de alguns detalhes técnicos, e um quadro comparativo com as outras abordagens apresentadas;
- **Capítulo 5**: Este capítulo contém os resultados obtidos com os experimentos realizados utilizando a estratégia proposta, comparando-a com as metodologias tradicionais e alguns trabalhos relacionados;
- **Capítulo 6**: Por fim, serão apresentadas as considerações finais sobre este trabalho, identificando pontos positivos e negativos da estratégia proposta, além de sugestões de algumas melhorias e otimizações possíveis.

2

Terminologia, Conceitos e Fundamentos

Este capítulo tem por objetivo apresentar os conceitos de ARS, abordando fundamentos e definindo algumas terminologias utilizadas em pesquisas desenvolvidas na literatura.

2.1 Teoria dos Grafos - Conceitos gerais

Um grafo é uma estrutura abstrata cujas primeiras noções originaram-se de um dos trabalhos de Leonhard Euler, em 1736, e que representa um conjunto de objetos onde alguns deles estão ligados entre si. Cada objeto é representado no grafo como um nó ou vértice, enquanto uma ligação é representada por uma aresta. Formalmente, um grafo é definido como um conjunto V de vértices e um conjunto A de pares de vértices, que representam as arestas. Esses dois conjuntos formam a estrutura G(V,A). Quando o conjunto V é vazio, o grafo pode ser chamado de grafo nulo e quando A é vazio, de grafo vazio (WASSERMAN; FAUST, 1994).

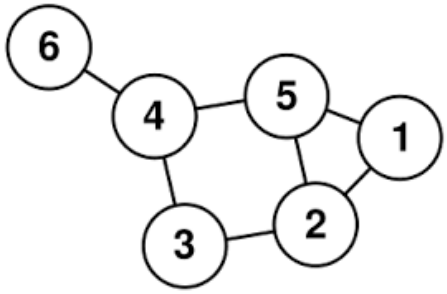

Figura 2.1: Exemplo de um grafo não-nulo. As circunferências numeradas representam vértices, enquanto as linhas que as conectam representam as arestas.

No caso das redes que interessam ao presente trabalho, os grafos analisados sempre serão não-nulos e não-vazios (ver Figura 2.1).

Uma rede social pode ser representada por meio de um grafo, onde cada indivíduo que pertence à rede é representado como um vértice, e os relacionamentos entre indivíduos representados como arestas entre os respectivos vértices. As arestas representam relações sociais, como amizade, hierarquias, relacionamentos familiares, entre outros. A aresta pode ou não ser direcionada, se os relacionamentos forem recíprocos (por exemplo, a relação de amizade) ou não (relação de citação em artigos científicos). Em outros tipos de redes, como por exemplo, as de

proteínas, os vértices representam proteínas e as arestas, reações químicas entre elas (JEONG et al., 2000; ZENG et al., 2013).

A representação de um grafo pode ser ampliada de modo a armazenar informações contextuais sobre a rede que representa. Por exemplo, em uma rede social, cada vértice poderia conter informações sobre o usuário, como preferências pessoais ou localização, e as arestas, o peso atribuído ao relacionamento de amizade (por exemplo, colegas de trabalho, amigos, melhores amigos e familiares), momento em que o relacionamento surgiu, quem fez a solicitação de amizade, etc. Um grafo costuma ser chamado de multimodal quando as informações contextuais estão armazenadas nos vértices, diferenciando-os entre si, ou multirrelacional se essas informações estiverem presentes nas arestas, o que permite a criação de diferentes tipos de arestas dentro do mesmo grafo (WASSERMAN; FAUST, 1994).

Outro ponto importante a ser observado é sobre as características das arestas. As duas mais observadas são o direcionamento e o valor da aresta, explicadas a seguir:

- **Direcionamento**: arestas de um grafo podem ser de dois tipos, direcionadas ou não-direcionadas. É importante diferenciar esses dois tipos de arestas, pois elas são usadas para representar relacionamentos unilaterais ou mútuos, respectivamente. Do ponto de vista das abordagens, a lógica dos algoritmos também é um pouco alterada, pois grafos direcionados permitem uma maior quantidade de arestas e requerem tratamento diferenciado. Grafos direcionados são aqueles em que se define o vértice de origem e de destino: a aresta sai de um vértice e chega em outro. X se relacionar com Y não implica que Y se relaciona com X. A representação gráfica de arestas desse tipo é feita comumente através de setas. Já em grafos não-direcionados ocorre o oposto: todos os vértices ligados por arestas transmitem e recebem a relação, ao mesmo tempo, ou seja, se X se relaciona com Y, Y se relaciona com X. A representação gráfica é feita através de uma linha.

- **Valores ou pesos**: Arestas podem ou não possuir pesos, que indicam, principalmente, a importância da relação dentro da rede. Quando uma aresta não é valorada, ela só possui dois estados: 0 ou 1. Ou a aresta existe ou é ausente. Quando as arestas são valoradas, elas podem existir com diferentes pesos (que podem ser valores discretos ou contínuos). Como no exemplo já citado, uma rede social pode ter seus relacionamentos valorados como colegas de trabalho, amigos, melhores amigos e familiares, implicando em diferentes valores atribuídos a cada aresta.

2.1.1 Definições

Todos os grafos possuem propriedades comuns e alguns tipos de grafos possuem propriedades específicas, dando possibilidade a criação de uma classificação baseada nessas características. Algumas propriedades e tipos de grafos têm importância na estratégia proposta por este trabalho, sendo eles:

- **Grau de um vértice**: o grau é uma propriedade dos vértices de um grafo e conta a quantidade de arestas que incidem em um determinado vértice. Quando o grafo é direcionado, tem-se dois tipos de grau: o de entrada, ou seja, quantas arestas "entram" no vértice, e o de saída, que conta quantas arestas "saem" do vértice;
- **Grafo nulo**: é um grafo $G(V,A)$ onde $V = \emptyset$;
- **Grafo vazio**: é um grafo onde todos os vértices possuem grau 0 (seja de incidência ou de entrada e saída), isto é, $A = \emptyset$;
- **Subgrafo**: de um grafo $G(V,A)$ é um grafo $G'(V',A')$, onde $V' \subseteq V$ e $A' \subseteq A$. A' não possui elementos que conectam vértices que não pertencem a V';
- **Clique**: é um subgrafo onde todos os vértices estão conectados entre si. Um clique de n vértices costuma ser chamado de n-clique ou K_n;
- **Caminho**: dados dois vértices i e j quaisquer de um grafo, diz-se que existe um caminho entre eles se existir uma sequência de arestas que levem de i a j, passando por outros vértices do grafo. Por exemplo, num grafo onde $V = \{i, j, k, l, m\}$ e $A =$ {(i,k),(k,l),(l,m),(m,j),(j,k)}, a sequência {(i,k),(k,l),(l,m),(m,j)} é um caminho entre i e j;
- **Menor caminho**: é o caminho entre dois vértices i e j com menor custo de travessia (soma dos pesos das arestas). Quando as arestas do grafo não são valoradas, essa definição pode ser reduzida para "o caminho com menor número de arestas".
- **Componente conectado**: é um subgrafo onde, para quaisquer pares de vértices pertencentes a ele, existe um caminho entre o par;
- **Vizinhança de um vértice**: é o conjunto de outros vértices do grafo que estão diretamente conectados a um vértice por uma aresta;
- **Circuito**: é um caminho de tamanho maior que 0 que começa e termina no mesmo vértice;
- **Grafo bipartido**: é um grafo cujo conjunto de vértices pode ser dividido em dois subconjuntos disjuntos onde não há aresta entre dois vértices de um mesmo subconjunto. Ainda, para ser bipartido, o grafo não pode ter circuitos de tamanho ímpar.

Além dessas definições básicas, ao analisar redes sociais, pode-se observar algumas características peculiares, que ocorrem devido a natureza não determinística da formação das mesmas. A seguir serão citadas algumas dessas propriedades, tanto locais quanto globais, e métricas utilizadas para quantificar essas propriedades.

- **Centralidade**: medidas de centralidade tentam quantificar a importância de um vértice dentro do grafo, ou seja, vértices relevantes para o fluxo de informação da rede. A centralidade pode estar relacionada a uma de três características: o vértice central é o de maior grau (mais bem conectado), o vértice central é aquele que está mais próximo dos outros (proximidade) ou é aquele que está entre dois vértices quaisquer (intermediação), na maioria das vezes (FREEMAN, 1979);

- **Diâmetro**: é o maior caminho possível entre dois vértices quaisquer de um grafo *G*. Enumerando-se todos os menores caminhos entre os possíveis pares de vértices, o maior entre eles é chamado de diâmetro da rede;

- **Densidade**: mede o grau de coesão entre os vértices da rede, ou seja, o quão próximos estão os indivíduos da rede, de um modo geral. Uma rede mais densa torna mais fácil o trânsito de informação entre os indivíduos, porque, em média, eles estarão mais próximos uns dos outros (grau de proximidade maior). A equação de densidade pode ser escrita como:

$$densidade(G) = \frac{2|A|}{n * (n-1)}, \tag{2.1}$$

  onde $|A|$ é o tamanho do conjunto de arestas do grafo e n é o tamanho do conjunto de vértices. Caso o grafo seja direcionado, retira-se o fator 2 do numerador, pois existem arestas em duas direções diferentes;

- **Máximo Componente Conectado**: é o maior componente conectado em termos de número de vértices;

- ***Cluster***: ou comunidade é um subgrafo mais denso, considerando os vértices que pertencem a ele, porém pouco denso considerando as ligações a vértices fora do subgrafo. A Figura 2.2, Página 23, exemplifica a presença de três *clusters* dentro de um grafo. *Clusters* estão presentes em diversas redes complexas, como redes sociais, redes de computadores, entre outros tipos (BARABASI, 2002). Exemplos atuais de *clusters* são: redes internas de universidades ou empresas privadas ou os grupos do Facebook[1]. Apesar de *clusters* se formarem em diversos tipos de redes diferentes, a conotação varia de acordo com o tipo de rede, por exemplo, grupos do Facebook reúnem pessoas com interesses parecidos enquanto redes internas de computadores reúnem computadores pertencentes à mesma instituição ou empresa.

- **Coeficiente de Clusterização Global**: quantifica o número de 3-cliques que existem no grafo normalizado pelo número total de 3-cliques possíveis. A diferença entre o coeficiente global e o local é que o local é calculado para um vértice, apenas,

[1] www.facebook.com

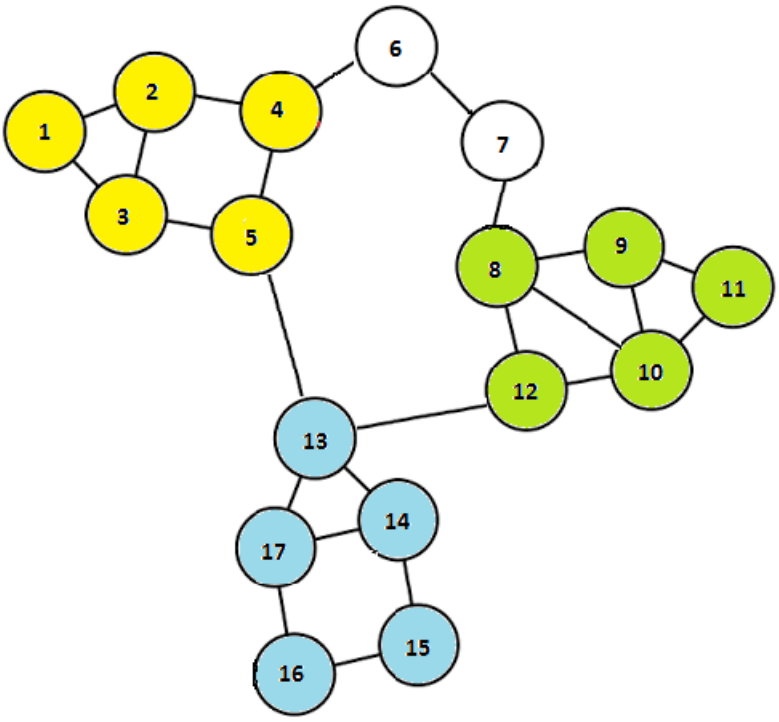

Figura 2.2: Exemplos de clusters dentro de um grafo. Existem três clusters, cujos vértices estão próximos uns dos outros, formados pelos vértices (1, 2, 3, 4 e 5), (8, 9, 10, 11 e 12) e (13, 14, 15, 16 e 17).

enquanto o global com relação ao grafo inteiro (WASSERMAN; FAUST, 1994). O coeficiente global é calculado da seguinte maneira:

$$CC_g(G) = \frac{3 * \#\Delta}{\#\Delta_T}, \tag{2.2}$$

onde $\#\Delta$ é o número de 3-cliques presentes e $\#\Delta_T$ é o total de 3-cliques possíveis para o grafo G;

- **Coeficiente de Clusterização Local**: de um vértice mede quanto falta para seus vizinhos formarem um clique. Esta medida é feita contabilizando o número de 3-cliques que o vértice faz parte, normalizado pelo número máximo de 3-cliques possíveis no grafo. Matematicamente, o coeficiente é calculado como se segue:

$$CC_l(i) = \frac{2 * \Delta(i)}{|\Gamma(i)| * (|\Gamma(i)| - 1)}, \tag{2.3}$$

onde $\Delta(i)$ é o número de 3-cliques aos quais o vértice i pertence e $|\Gamma(i)|$ é o tamanho do conjunto de vizinhos ao vértice i;

- **Resiliência**: é uma propriedade da rede que consiste em manter suas outras características quando há uma remoção de vértices. A remoção de um vértice pode ocasionar a diminuição das outras métricas vistas anteriormente, prejudicando, em

geral, o fluxo de informações dentro da rede. Remover um vértice pode ocasionar no desaparecimento de um *cluster*, pode fazer com que um componente do grafo deixe de ser conectado, além de poder afetar a densidade e os coeficientes de clusterização local e global.

Apesar da representação gráfica do grafo ser de fácil entendimento e simples de se esquematizar, existem algumas representações mais adequadas para o processamento por um programa de computador, que tornam mais fáceis a extração de dados e a análise automática. Os formatos mais comuns são detalhados a seguir.

2.1.2 Outras representações de grafos

- **Matriz de adjacência**: representação matricial de um grafo e simples de ser lida, essa estrutura é uma matriz M quadrada e de tamanho n x n (ou de tamanho m x n, caso o grafo seja bipartido), onde n é o número de vértices do grafo e onde as linhas e colunas representam esses vértices. As arestas podem ser representadas por um *bit* (indicando existência ou ausência) ou por um peso (indicando a importância). Os *bits* ou pesos, referentes ao par de vértices i e j, ficam armazenados posição $M(i,j)$ da matriz. Caso o grafo seja não-direcionado, a matriz M se torna simétrica, já que para todo i, j, se $(i,j) \in A, (j,i) \in A$. Dado um grafo G(V,A), a matriz de adjacência M pode ser definida como:

$$M(i,j) = \begin{cases} 1, \ se \ (i,j) \in A; \\ 0, \ se \ (i,j) \notin A. \end{cases} \tag{2.4}$$

ou

$$M(i,j) = \begin{cases} w_{i,j}, \ se \ (i,j) \in A; \\ 0, \ se \ (i,j) \notin A. \end{cases} \tag{2.5}$$

O primeiro caso aplica-se a grafos com arestas não valoradas, enquanto o segundo aos grafos com arestas valoradas.

- **Lista de adjacência**: é uma lista de listas cujos itens são identificados pelos vértices do grafo. Quando uma aresta existe entre os vértices i e j, a lista indexada por i contém j como um vértice adjacente e assim por diante. A vantagem de se usar essa representação é quando o que mais interessa no grafo são as arestas presentes e pode ser bastante útil para poupar memória e processamento durante sua leitura, quando o grafo é esparso;

- **Matriz de incidência**: Na matriz de incidência, os vértices identificam as linhas da matriz enquanto que as colunas são indexadas pelas conexões do grafo. Um cruzamento entre uma linha e uma coluna da matriz indica se um determinado vértice

é incidente a uma aresta específica. Assim como na matriz de adjacência, também é possível a manipulação de grafos direcionados e com arestas valoradas neste tipo de representação (WASSERMAN; FAUST, 1994).

Adiante são mostrados alguns outros conceitos simples e que se relacionam com a teoria dos grafos. Alguns desses conceitos são bastante importantes, já que servem de base para o algoritmo implementado neste trabalho.

2.2 Teoria dos Grafos - Conceitos específicos

2.2.1 Tríade

É o menor subgrafo não trivial, que possui apenas três vértices e que pode ter de zero até três arestas, quando as arestas não são direcionadas, e zero até seis arestas, quando as arestas são direcionadas. É chamado de não trivial pois os vértices pertencentes a uma tríade podem ter mais de um vizinho, enquanto uma díade, um subgrafo de dois vértices, pode conter no máximo apenas uma (ou duas, para grafos direcionados) aresta (cada vértice só pode ter no máximo um vizinho). Uma tríade não pode ser nula e pode ser vazia, o que não é interessante para este trabalho. Os motivos que levam a descartar tríades vazias serão mostrados posteriormente. Uma tríade também pode ser equivalente a um 3-clique, se ela possuir 3 arestas (ou 6 arestas, para grafos direcionados), o máximo possível. Nas tríades, são consideradas apenas as conexões que os vértices têm ou podem ter entre si, desconsiderando arestas que ligam a vértices fora da tríade, assim como qualquer subgrafo (MOODY, 1998; BATAGELJ; MRVAR, 2001; JUSZCZYSZYN; MUSIAL; BUDKA, 2011). O número de possíveis tipos de tríades depende do tipo de grafo: se um grafo for não direcionado, existem 8 possibilidades de configurações de arestas (ver Figura 2.3), pois elas sempre são bilaterais.

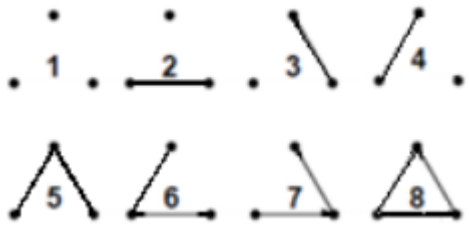

Figura 2.3: Os 8 tipos possíveis de tríades presentes em grafos não direcionados. Fonte: (JUSZCZYSZYN; MUSIAL; BUDKA, 2011)

Caso o grafo seja direcionado, a análise é mais extensa, pois existem 64 configurações possíveis de arestas (ver Figura 2.4, Página 26), já que as arestas são unilaterais.

É importante notar todos os tipos possíveis de tríades, pois as mudanças de configurações das arestas das tríades dentro de um grafo, ao longo do tempo, serão um dos elementos-chave para a estratégia proposta por este trabalho.

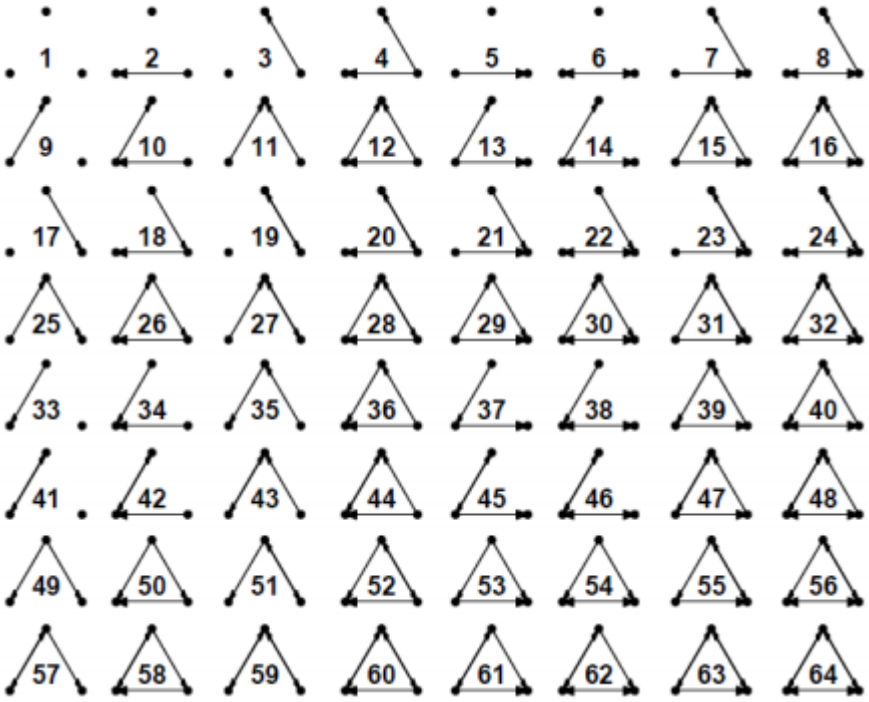

Figura 2.4: Os 64 tipos possíveis de tríades presentes em grafos direcionados. Fonte: (JUSZCZYSZYN; MUSIAL; BUDKA, 2011)

2.2.2 *Triad Census*

Triad Census (DAVIS; LEINHARDT, 1967; MOODY, 1998; BATAGELJ; MRVAR, 2001) é um algoritmo estatístico que visa contabilizar quantas tríades de cada tipo (variando se o grafo é direcionado ou não) estão presentes no grafo em questão. Basicamente, considera todas as combinações de vértices tomados três a três possíveis. Sua primeira versão é custosa, com $O(n^3)$ operações para o pior caso. A primeira versão do algoritmo é descrita pelo Algoritmo 1 a seguir.

Algorithm 1 Triad_Census()

Entrada: $grafo\ G(V,A)$
Saida: $array\_tipos\_triades$ ▷ Totais de todos os tipos de tríades presentes.
$array\_tipos\_triades \leftarrow \{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\}$ ▷ 64 índices se o grafo for direcionado.
for $i \in V$ **do**
    for $j \in V$ **do**
        for $k \in V$ **do**
            if $i < j$ AND $j < k$ **then** ▷ Não contar a mesma tríade 3 vezes.
                $tipo\_triade \leftarrow get\_tipo\_triade(i,j,k)$
                $array\_tipos\_triades[tipo\_triade] \leftarrow array\_tipos\_triades[tipo\_triade] + 1$
return $array\_tipos\_triades$

Algumas variações do *Triad Census*, entre elas as implementadas por MOODY (1998) e BATAGELJ; MRVAR (2001) conseguem diminuir a complexidade de tempo para $O(n^2)$, e em alguns casos (redes esparsas e com grau máximo pequeno), $O(n)$.

2.2.3 Matriz de Transição de Tríades

Uma rede complexa, em especial, uma rede social, evolui ao longo do tempo, ou seja, sua estrutura está constantemente mudando. Isto significa que novos indivíduos passam a fazer parte ou deixam a rede e que relacionamentos são desfeitos ou criados. Ao analisar uma tríade entre um determinado tempo t e um tempo $t + \Delta$ posterior, pode-se verificar que a tríade permaneceu do mesmo tipo (a configuração das arestas não mudou) ou que *links* possam ter desaparecido ou surgido, fazendo assim com que a tríade passe a ser de um tipo diferente. Esta mudança, verificada ao comparar duas observações em tempos diferentes, da configuração de arestas de um trio específico de vértices é chamada de transição de tríade.

A partir de uma análise onde se identificam todas as transições de tríades que ocorrem entre um determinado tempo t e um tempo $t + \Delta$ posterior, pode-se construir uma Matriz de Transição de Tríades (*Triad Transition Matrix - TTM*, em inglês) (JUSZCZYSZYN; MUSIAL; BUDKA, 2011), estrutura que armazena a contagem de todas as transições de tríades observadas, de modo a permitir o cálculo de probabilidades de uma tríade de um tipo qualquer encontrada no grafo mudar a disposição das arestas ou de se manter. Essa matriz tem tamanho $64 * 64$, já que existem 64 possibilidades de tríades, para grafos direcionados, e tamanho $8 * 8$ para grafos não direcionados. Os valores encontrados na matriz são definidos como:

$$TTM_t(i,j) = P(g_i[t] \rightarrow g_j[t+1]). \tag{2.6}$$

Ou seja, a TTM para um tempo t pode ser definida como a probabilidade de transição de uma tríade de um tipo i presente em t para uma tríade do tipo j no tempo $t + 1$, por exemplo, a probabilidade de uma tríade do tipo 6, observada no tempo t, passar a ser do tipo 7 no tempo $t + 1$. Vale ressaltar que nessa definição, o somatório de todas as probabilidades da tríade do tipo i se transformar em qualquer uma das outras ou de se manter somam 1, isto é, todos os valores de uma linha qualquer da matriz TTM somam 1. Esta representação é mais robusta, pois não só contabiliza a probabilidade de arestas aparecerem como também a de arestas desaparecerem, ou seja, relacionamentos dentro de uma rede deixarem de existir.

O cálculo de $P(g_i[t] \rightarrow g_j[t+1])$ é feito através da contagem das tríades do grafo e da contagem de todas as transições ocorridas a partir dessas tríades. A Figura 2.5, Página 28, exemplifica essa contagem: analisando as 13 tríades do tipo 5 observadas no tempo t, verifica-se, no tempo $t + 1$, que 4 delas passaram a ser do tipo 4, 7 não apresentaram qualquer alteração, continuando a ser do tipo 5 e duas ganharam mais arestas, sendo classificadas como do tipo 7. Para este caso, $P(g_5[t] \rightarrow g_4[t+1]) = 4/13$, $P(g_5[t] \rightarrow g_5[t+1]) = 7/13$ e $P(g_5[t] \rightarrow g_7[t+1]) = 2/13$. Portanto, $P(g_i[t] \rightarrow g_j[t+1])$ é um valor bem particular a cada grafo diferente em que uma TTM é calculada.

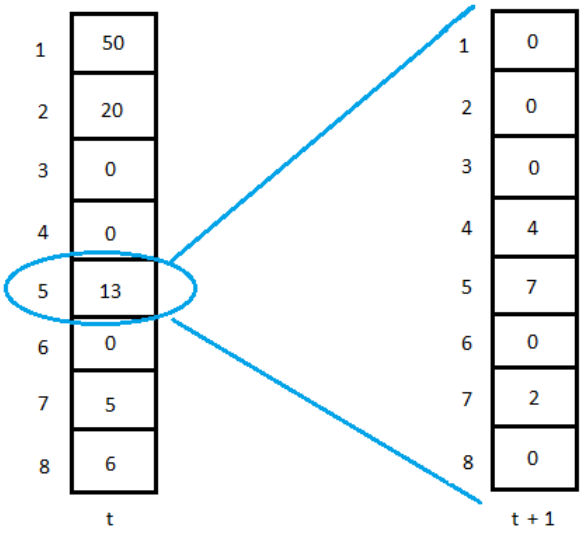

Figura 2.5: Exemplo de contagem de tríades. Nesse caso, ao analisar todas as 13 tríades do tipo 5 observadas no tempo t, verifica-se que, no tempo $t + 1$, 4 passaram a ser do tipo 4, 7 continuaram do tipo 5 e 2 passaram a ser do tipo 7.

2.2.4 Tensores

Um Tensor (DUNLAVY; KOLDA; ACAR, 2011; SPIEGEL et al., 2012) Z é uma estrutura de dados que generaliza o conceito de matriz, podendo considerar mais de duas dimensões, capaz de armazenar sequências de informações, geralmente sobre o mesmo objeto de interesse. Um tensor bidimensional (semelhante a uma matriz) pode armazenar sequências de dados unidimensionais, como vetores, e um tensor tridimensional armazena múltiplos dados bidimensionais, e assim por diante. A Figura 2.6, Página 29, mostra um tensor Z tridimensional que armazena múltiplas matrizes de adjacência de um grafo, observadas do tempo t_1 até o tempo t.

É interessante notar a utilidade do tensor tridimensional: os dados bidimensionais que se quer armazenar são exatamente matrizes de adjacência de um grafo, enquanto a terceira dimensão pode ser interpretada como intervalos de tempo (dias, meses, anos, etc). Portanto, pode servir para armazenar sequências dessas matrizes, ou seja, *snapshots* do grafo ao longo do tempo. Podemos definir um tensor Z, de tamanho $n * n * t$, tridimensional da seguinte maneira:

$$Z(i,j,t) = \begin{cases} 1, \ se \ (i,j) \in A_t; \\ 0, \ se \ (i,j) \notin A_t. \end{cases} \tag{2.7}$$

A_t é o conjunto de arestas do grafo G no tempo t (G_t). Vale ressaltar que é simples adaptar essa representação para arestas com pesos:

$$Z(i,j,t) = \begin{cases} w_{i,j,t}, \ se \ (i,j) \in A_t; \\ 0, \ se \ (i,j) \notin A_t. \end{cases} \tag{2.8}$$

Assim como está esquematizado na Figura 2.6, Página 29, um Tensor Z pode ser usado para armazenar representações temporais de grafos, ou seja, os estados do grafo do tempo t_1 até

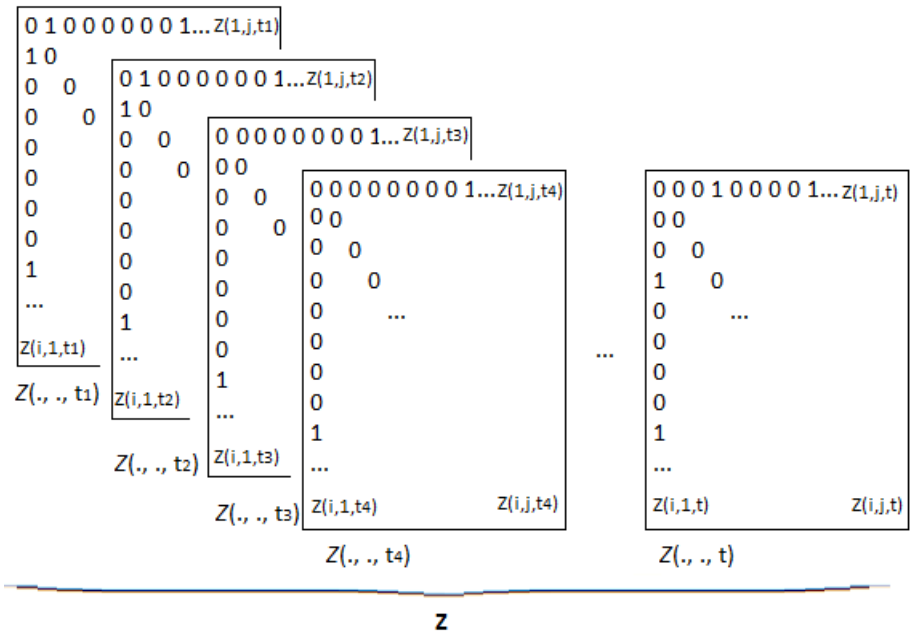

Figura 2.6: Exemplo gráfico de um tensor tridimensional.

o tempo t, onde cada índice $Z(., ., t)$ armazena a matriz de adjacência do grafo correspondente ao que foi observado em t. Existem técnicas de processamento de tensores, como a decomposição canônica (DUNLAVY; KOLDA; ACAR, 2011), que visam extrair informações relevantes e que são de interesse da ARS, em especial para o problema de predição temporal de *links*.

2.3 Análise de Redes Sociais

Uma rede social é uma estrutura formada por um conjunto de indivíduos ou organizações. Elas existem desde quando o ser humano começou a viver em sociedade e sempre se caracterizou pela constante mudança em sua estrutura interna. Com o aparecimento e popularização da *internet*, um tipo novo e diferente (pois as interações entre indivíduos ocorrem por meios eletrônicos) de rede surgiu: a rede social virtual. Existem diversas redes desse tipo, cada uma reunindo um grupo de pessoas com interesses em comum, como compartilhar conhecimento, expor a vida social e fazer amizades, desenvolver *softwares* livres, redes internas de empresas privadas, entre outros. Redes *online* se popularizaram muito nos últimos anos, dentre as quais pode-se citar: Facebook[2], Twitter[3] e diversas outras. Outro exemplo bastante interessante que se pode observar são redes colaborativas, isto é, redes formadas por acadêmicos e as colaborações científicas entre eles.

Hoje em dia, essas redes são alvos de empresas especializadas em estudos sobre o comportamento das pessoas (por exemplo, preferências de compras), por envolverem uma grande quantidade de indivíduos. Também, pelo mesmo motivo, são alvos de estudos por diversas áreas do conhecimento, entre elas a computação. A ARS é o ramo da Computação que

[2] www.facebook.com
[3] www.twitter.com

estuda o surgimento, características e evolução de redes sociais em geral e atualmente abrange diversos temas sobre redes. Em resumo, ARS faz uso de várias métricas para extrair informações sociológicas sobre uma rede social (WASSERMAN; FAUST, 1994).

As redes sociais vêm sendo objeto de estudo há muitos anos na área da Sociologia, Antropologia, Matemática, Física, Estatística, entre outras (ALBERT; BARABASI, 2002; NEWMAN; BARABASI; WATTS, 2011). Um dos problemas encontrados anteriormente é o fato das redes serem extensas e complexas, e a coleta de dados sobre tais redes ser feita manualmente, o que prejudicava a completude dos dados e podendo introduzir erros de coleta. O avanço tecnológico e o posterior surgimento de redes sociais virtuais permitiu, contudo, a automatização do processo de coleta, além de poder-se avaliar redes com milhares e até milhões de indivíduos, minimizando a possibilidade de erros durante a análise. Dessa forma, as redes (não apenas as sociais) inseridas no contexto digital podem ser consideradas abstrações que podem representar, até certo ponto, o mundo real. Através delas, sistemas podem ser representados e problemas podem ser investigados (ALBERT; BARABASI, 2002).

2.3.1 Tipos de redes

Até agora foi citado apenas que redes, de um modo geral, e inclusive as redes sociais, se formam através de processos complexos e não-determinísticos, o que dificulta o entendimento dos comportamentos que levam ao surgimento de relacionamentos, principalmente. Apesar disso, existem modelos de rede que tentam explicar tais comportamentos. Alguns dos principais modelos são explicados adiante.

2.3.1.1 Rede aleatória

É um modelo de rede proposto por SOLOMONOFF; RAPOPORT (1951) e por ERDöS; RéNYI (1959), que assume que conexões têm a mesma probabilidade de aparecer entre quaisquer vértices da rede, independente de qualquer fator. Em um grafo $G_{n,p}$ que segue esse modelo, o aparecimento de conexões ocorrem seguindo uma distribuição binomial, isto é, a probabilidade de existir n arestas é dado por $p^n * (1-p)^{N-n}$, onde N é o número máximo possível de arestas de $G_{n,p}$, ou seja, $\frac{n*(n-1)}{2}$. Concluindo, o grau esperado de um vértice é dado por $p*(n-1)$ (NEWMAN, 2003).

Ainda de acordo com NEWMAN (2003), quando o número de vértices aumenta consideravelmente, essa distribuição se aproxima de uma distribuição de Poisson (ver Figura 2.7, Página 31), já que os eventos que representam o surgimento e remoção de arestas são independentes. Desse modo, a probabilidade de um vértice qualquer ter grau k é dada como:

$$p_k = \begin{pmatrix} n \\ k \end{pmatrix} p^k * (1-p)^{n-k}. \tag{2.9}$$

$$p_k \simeq \frac{\lambda^k * e^{-\lambda}}{k!}. \tag{2.10}$$

Quando, em média, uma rede aleatória possui a média dos graus de todos os vértices maior que 1, o número de vértices de grau 0 diminui exponencialmente. Isso implica em que os componentes conectados da rede passam a crescer em tamanho, até que se tenha um grande componente conectado ou componente gigante (GIRVAN; NEWMAN, 2002; BARABASI, 2002; NEWMAN, 2003). A principal importância de um componente gigante dentro da rede é que qualquer informação possa chegar a maioria dos vértices, ou seja, um componente gigante garante a propagação da informação a todos os indivíduos. Em redes sociais, a existência de um componente gigante é comumente observada em um Modelo de Mundo Pequeno (STROGATZ, 2001).

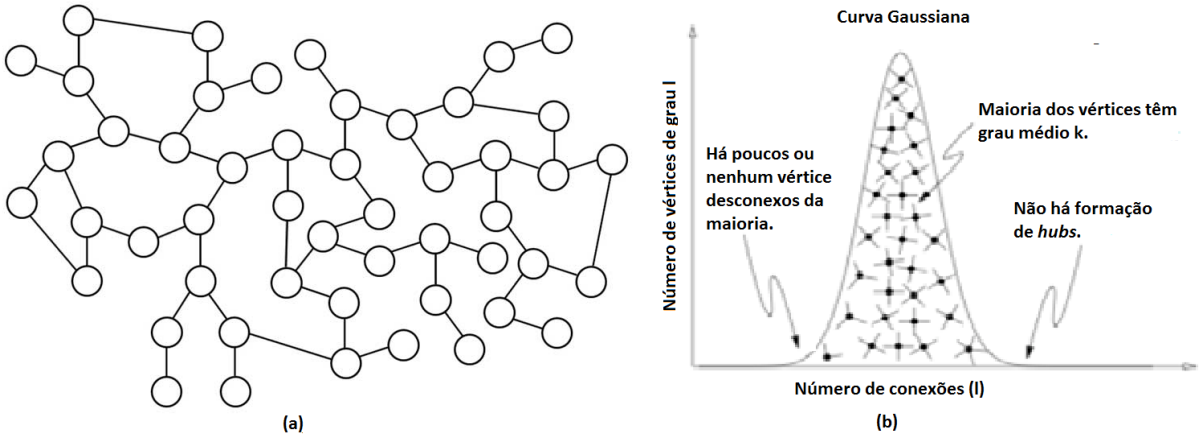

Figura 2.7: Exemplo de uma rede aleatória (a) e da distribuição dos graus (b). Nota-se, em (a), que o Maior Componente Conectado é o próprio grafo e que não há vértices com grau alto (*hubs*) e, em (b), que a curva formada é semelhante a Gaussiana, já que a distribuição é binomial ou de Poisson.

2.3.1.2 Modelo de Mundo Pequeno

Quantos contatos intermediários separam uma pessoa de qualquer outra ao redor do mundo? Esta pergunta, sem resposta inicialmente, despertou o interesse de acadêmicos, que passaram a experimentar sobre o quão próximos estão dois indivíduos quaisquer em uma rede. Um experimento feito por Stanley Milgram e Jeffrey Travers, na década de 1960, consistia em escolher remententes nos Estados Unidos e fazer com que eles enviassem uma carta para um determinado destinatário final. Caso o remetente o conhecesse pessoalmente, deveria enviar a carta diretamente ao destinatário final e caso contrário, deveria enviá-la para um contato que fosse conhecido pessoalmente (pelo primeiro nome) e que provavelmente conhecesse o destinatário final. O resultado da experiência deu origem a expressão "Seis Graus de Separação", pois Stanley descobriu que as cartas passavam, em média, por seis contatos intermediários até chegar no destinatário final. Mais tarde, foi verificado que não somente sociedades, mas também vários

tipos de redes complexas são densas, isto é, a distância média entre dois vértices quaisquer numa rede é pequena, fenômeno que foi chamado de "Mundo Pequeno" (BARABASI, 2002).

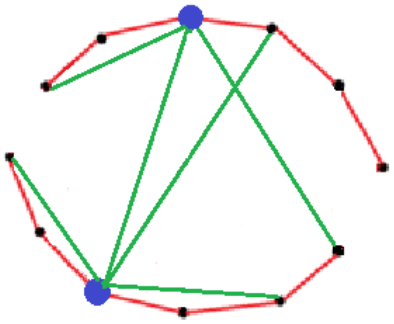

Figura 2.8: Exemplo de uma rede de mundo pequeno. É possível ver os *hubs* em azul e as arestas verdes, elos mais fracos que servem como atalhos para a propagação da informação para vértices mais distantes.

Na década de 1970, (GRANOVETTER, 1973) publicou a obra chamada de "The Strength of Weak Ties", propondo que as informações que transitam mais eficientemente por uma rede se deslocam por conexões de pouca relevância da rede ao invés de *links* sólidos. Desse modo, a sociedade seria composta de *clusters*, ou seja, comunidades altamente conectadas, e existiriam algumas conexões externas e pouco frequentes que interligariam aqueles *clusters* (ver Figura 2.8), evitando o isolamento do resto da rede, mesma observação feita por GIRVAN; NEWMAN (2002). O fato de informações semelhantes transitarem dentro de comunidades e de informações mais diversas se deslocarem entre aquelas comunidades, ou seja, através de conexões menos relevantes, justifica a afirmação de Granovetter (BARABASI, 2002).

Baseado nas observações de Granovetter, por que razão uma rede de mundo pequeno não pode ser totalmente modelada de modo aleatório, como nas redes aleatórias? Se isso fosse possível, a probabilidade de dois indivíduos próximos, ou seja, dentro da mesma comunidade, se conhecerem deveria ser a mesma de dois indivíduos bem distantes um do outro (indivíduos de *clusters* diferentes), o que não ocorre no mundo real. Em 1998, Duncan J. Watts e Steven Strogatz criaram o Modelo de Mundo Pequeno (WATTS; STROGATZ, 1998), uma alternativa ao modelo de rede aleatória, de modo a explicar o aparecimento de comunidades no mundo real.

Este modelo cria inicialmente aglomerados densos, que simulam a formação de comunidades, e redefinem algumas conexões para vértices aleatoriamente escolhidos, possivelmente criando atalhos entre vértices distantes, evitando o já citado isolamento de um *cluster* e diminuindo o grau de separação médio da rede (WATTS; STROGATZ, 1998; WATTS, 1999; NEWMAN, 2003). Uma das grandes vantagens de redes de mundo pequeno é que elas são bem mais robustas, ou seja, são mais tolerantes a perturbações na rede (como a remoção de vértices ou conexões), sendo considerado uma vantagem evolutiva, já que se um vértice for removido, por exemplo, ainda existirão alguns outros caminhos entre comunidades por onde a informação pode transitar, não isolando-os do resto da rede.

Por fim, o modelo consegue criar um mundo pequeno compatível com o experimento

dos "Seis Graus de Separação" de Stanley e Jeffrey e com o experimento de GRANOVETTER (1973). Uma das evidências disso é que as cartas do experimento dos seis graus de separação só conseguem chegar ao destinatário devido a presença de um componente gigante na rede, gerado pelas redefinições aleatórias de conexões.

2.3.1.3 Redes livres de escala

Este modelo foi derivado a partir do estudo das redes de *internet* como uma rede complexa (BARABASI; ALBERT, 1999). Dentro da *web*, foi descoberto que a distribuição de vértices não é igualitária e também a presença de vértices com grau muito alto, isto é, muito bem conectados dentro da rede. Esses vértices passaram então a ser chamados de *hubs* e concluiu-se a importância deles para o fluxo de informações dentro das mais diversas redes, por apresentarem vários possíveis canais por onde essas informações podem transitar para chegar ao destino (BARABASI, 2002; NEWMAN, 2003). Para explicar o surgimento de *hubs*, as redes deveriam possuir alguma característica particular, que foi explicada pela Lei da Potência, basicamente ditando como se dá a distribuição de relações dentro da rede. A Lei da Potência pode ser expressada como:

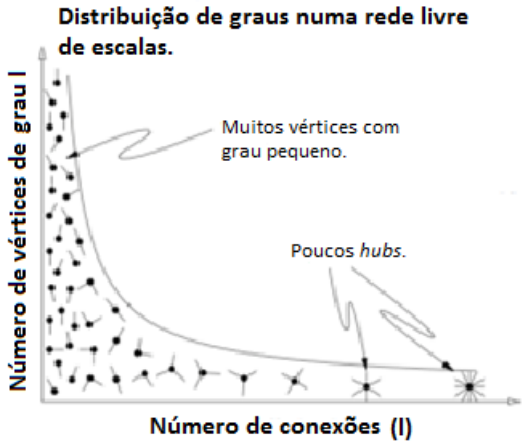

Figura 2.9: Representação gráfica de como estão distribuídos conexões que seguem a Lei da Potência (BARABASI, 2002).

$$p_k \sim k^{-\gamma}, \tag{2.11}$$

onde p_k é a probabilidade de existir um vértice com grau k, e γ é o expoente do grau, que na maioria dos casos varia entre 2 e 3 (BARABASI, 2002). A Figura 2.9 mostra, graficamente, como ocorre a distribuição entre o número de conexões e o grau dos vértices, seguindo a Lei da Potência. Essa lei consegue explicar o fato da grande maioria dos vértices, dentro de redes, possuírem poucas conexões, ou seja, um grau pequeno, enquanto os *hubs* são tão bem conectados. Mas, como explicar a ocorrência dessa lei? BARABASI; ALBERT (1999) verificaram que se dá devido a ocorrência de um fenômeno chamado de conexão preferencial. À primeira vista,

presumiu-se que essa preferência era por vértices mais antigos, ou seja, os primeiros indivíduos que surgiram na rede deveriam ser os de maior grau. Entretanto, o que realmente ocorre é que vértices de grau mais alto tendem a atrair conexões de outros vértices da rede, fenômeno conhecido como "as pessoas tendem a querer se relacionar com quem é mais popular" ou "o rico fica mais rico". Dessa forma, vértices com maior grau tendem a se conectar com novos vértices em uma velocidade superior se comparados a outros vértices da rede. A Figura 2.10 exemplifica uma rede livre de escalas. Percebe-se que existem poucos *hubs* (vértices marcardos), que tem grau alto, e muitos vértices com grau baixo, exatamente como descrito pela Lei da Potência (ver Figura 2.9, Página 33).

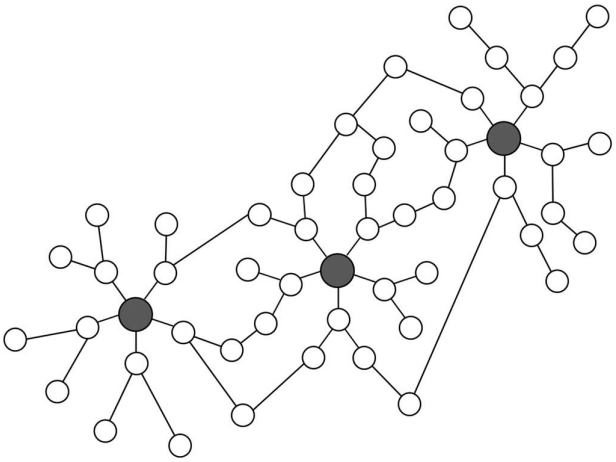

Figura 2.10: Exemplo de uma rede livre de escalas. Os *hubs* (grau 5) são representados pelos vértices marcados. Visualmente percebe-se que os *hubs* estão em menor número se comparados ao resto da rede e eles possuem bem mais conexões.

2.4 Outros exemplos de redes

Esta seção mostrará alguns exemplos de redes que seguem, em parte, os comportamentos apresentados pelos modelos mostrados.

Rede biológica

- **Memória de curto prazo**: pode-se fazer uma analogia entre o sistema nervoso humano e redes complexas. Um exemplo de mundo pequeno (ROXIN; RIECKE; SOLLA, 2004) pode ser visto no conjunto de neurônios que são responsáveis pelas memórias recentes (encontrados no córtex pré-frontal). Esses neurônios, assim como os outros, podem estar em dois estados diferentes, ativo (quando guarda uma memória) ou inativo. A Figura 2.11, Página 35, mostra um esquema de neurônios onde uma falha pode facilmente impedir a transmissão de impulsos nervosos para neurônios mais distantes.

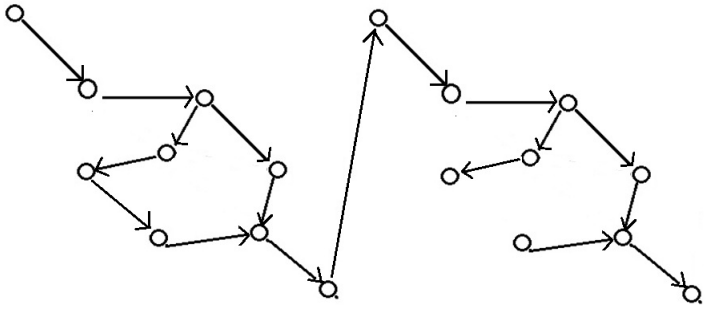

Figura 2.11: Exemplo de esquematização de neurônios do córtex.

Foi analisado que a adição de atalhos, da mesma maneira que ocorre em um modelo de mundo pequeno, provê meios de transmitir impulsos nervosos de maneira mais eficiente, além de aumentar a robustez da rede em caso de uma perturbação. Um exemplo é a ocorrência da morte de um neurônio (ver Figura 2.12). Como existem caminhos alternativos, os impulsos ainda conseguem chegar no neurônio final, além de que aqueles caminhos permitem a retro-alimentação da rede, ou seja, que o impulso nervoso se mantenha.

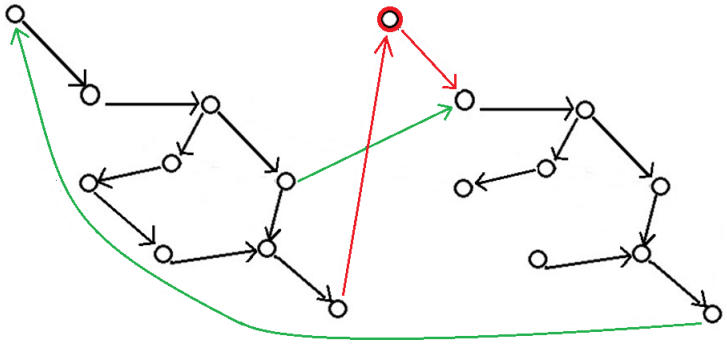

Figura 2.12: Adicionar "atalhos" (arestas verdes) aumenta a robustez da rede em caso de falhas (por exemplo, a morte de um neurônio, representado pelo vértices e arestas vermelhas) e permite retro-alimentação do sinal nervoso.

Redes de computadores

- ***World Wide Web***: rede formada por páginas de *internet* e por *hyperlinks* entre elas, começou como páginas em HTTP (*Hyper Text Transfer Protocol*) e contendo texto. Essas páginas referenciavam outras páginas, geralmente dentro de uma mesma rede interna de uma universidade. Foi verificado que a *web* se comporta, em parte, como um mundo pequeno (ALBERT; JEONG; BARABASI, 1999) e que a Lei da Potência pode ser aplicada a ela (KUMAR et al., 2000). Como a *web* é o maior exemplo de rede disponível publicamente, tem sido objeto de muitos estudos, inclusive na ARS (ALBERT; BARABASI, 2002; NEWMAN, 2003). Atualmente, possui mais de 1.1 bilhão de páginas[4] e é considerada uma rede direcionada, pois *hyperlinks* são

[4]Retirado de: http://www.internetlivestats.com/watch/websites/

unidirecionais.

Geofísica

- **Permeabilidade do solo**: YANG (2001) descobriu que a permeabilidade do solo também pode ser entendida como uma rede complexa, e em parte como de mundo pequeno. Isso se dá pelo fato das rochas serem porosas, permitindo a passagem de fluídos (interpretado como a informação que transita pela rede), e que esses poros formam conexões que facilitam a difusão do fluído dentro daquela rocha (a permeabilidade está diretamente relacionada a essa facilidade de difusão). As fissuras, ou "atalhos" presentes nas rochas podem ser interpretadas como os elos fracos citados por Granovetter, servindo de caminho por entre partes sólidas da rocha e aumentando ainda mais a permeabilidade desta.

Redes acadêmicas

- **Redes de coautoria**: rede formada por autores de artigos e trabalhos científicos e colaborações entre eles nos mesmos trabalhos acadêmicos, também conhecida como rede de coautoria. É uma rede social onde os vértices são representados pelos autores e as arestas ou relacionamentos existem se determinado par de autores já publicou um artigo juntos. Como a relação "publicar juntos" é mútua, o grafo que representa essa rede é não direcionado. Redes desse tipo são bastante exploradas pois os dados referentes a elas são públicos, ao contrário de outras redes com dados privados. Estudos apontam que redes de coautoria se comportam como um mundo pequeno, com um alto grau de proximidade, formação de vários *clusters*, separação média pequena entre vértices distintos e distribuição de graus seguindo a Lei da Potência (NEWMAN, 2001a; BARABáSI et al., 2002).

2.5 Considerações finais

Este capítulo explicou vários conceitos e exemplos sobre duas temáticas importantes neste trabalho: Análise de Redes Sociais e Teoria dos Grafos. Foram vistas propriedades e definições básicas sobre grafos, sua definição formal e outras representações mais adequadas para processamento automático. Além disso, foram discutidos também algumas características um pouco mais complexas, que servem para analisar grafos de uma perspectiva topológica.

Outros fundamentos relacionados ao tema de grafos também foram mais detalhados. O conceito de tríade foi explicado e seus tipos possíveis enumerados, pois serão abordados posteriormente neste trabalho. Um algoritmo estatístico para contagem de tríades, originalmente chamado de *Triad Census*, foi estudado e descrito pelo Algoritmo 1, Página 26. Por fim, duas

estruturas de dados mais complexas, os tensores e as matrízes de transição de tríades também foram apresentadas, pois têm importância na estratégia proposta por este trabalho.

Finalmente, foi dada uma breve explicação sobre a área de ARS, campo maior em que está inserido o problema de Predição de Links, problema-alvo da estratégia que será proposta adiante. Também foram mostrados, ainda relacionados à ARS, modelos de redes que tentam explicar comportamentos do mundo real, as motivações, alguns modelos matemáticos que descrevem tais comportamentos e exemplos práticos sobre esses modelos de redes. A Tabela 2.1 compara os aspectos principais dos modelos de rede apresentados.

Tipo de rede	**Modelo para o surgimento de conexões**	**Distribuição de graus dos vértices**	**Propagação da informação pela rede**
Aleatória	Totalmente aleatório.	Segue uma distribuição de Poisson.	É propagada para a grande maioria dos vértices. Quando a média dos graus dos vértices é maior que 1, o grafo se torna conexo e a informação alcança toda a rede.
Mundo pequeno	Aglomerados, ou *clusters*, densos e bem conectados. Algumas conexões são redefinidas aleatoriamente, possivelmente conectando vértices distantes.	Distribuição de graus similar a uma rede aleatória, mas com redefinição aleatória de poucas arestas.	Transita facilmente entre vértices de um mesmo *cluster*. Entre *clusters* diferentes, a informação é propagada através das conexões redefinidas, isto é, "elos fracos".
Livre de escalas	A maioria dos vértices da rede possuem grau baixo, enquanto poucos vértices, os *hubs*, possuem grau bem acima da média da rede. *Hubs* também tendem a ganhar novas conexões em um ritmo acima da média da rede (conexões preferenciais).	Segue a Lei da Potência	*Hubs* provêem vários canais diferentes para que a informação transite para outros vértices da rede.

Tabela 2.1: Quadro comparativo entre os modelos de rede apresentados.

Dentre os assuntos vistos ao decorrer deste capítulo, é importante destacar que as definições básicas de grafos, como propriedades e as outras representações para processamento, os conceitos relacionados à tríades (incluíndo os tipos de tríades, o *Triad Census* e a matriz de transição de tríades) e os conceitos sobre tensores são usados extensivamente neste trabalho. De modo geral, o estudo sobre os tipos de redes apresentados também propiciou um entendimento inicial sobre como é possível identificar os padrões de evolução que ocorrem em cada rede.

Apesar do esforço da comunidade acadêmica em entender como funcionam redes não-determinísticas (não só as formadas por seres humanos) no mundo real, ainda está longe de se ter um modelo satisfatório, pois existem vários tipos de redes diferentes (WATTS; STROGATZ, 1998; BARABASI; ALBERT, 1999), cada um com princípios próprios que determinam a sua evolução, e que cuja complexidade aumenta à medida em que o tamanho da rede também

aumenta (em termos de indivíduos ou relacionamentos), tornando-os ainda mais difíceis de explicar e simular.

No capítulo seguinte será discutido mais a fundo o problema-alvo deste trabalho: a Predição de Links em redes sociais. Será mostrado como o problema está relacionado às redes, como estruturas complexas e evolutivas e as principais abordagens, estáticas e temporais, utilizadas para tentar resolver satisfatoriamente este problema.

3

Predição de Links em redes sociais

Este capítulo irá explorar mais detalhadamente o problema de Predição de Links em redes sociais e como é tratado pelas abordagens, tanto tradicionais quanto as mais específicas que tratam de temporalidade encontradas na literatura. No fim serão apresentadas aplicações práticas e desafios da área.

3.1 Definição

A Predição de Links, como já foi falado na Seção 1.2, Página 15, é um problema dentre vários na área de ARS, que consiste em duas perspectivas, sendo elas a predição de conexões faltantes e analisar com se dá a evolução de uma rede ao longo do tempo, tratando a rede como uma estrutura dinâmica. GETOOR; DIEHL (2005) e HAN; PEI; KAMBER (2011) consideram a Predição de Links como um subproblema da ML e que emprega técnicas de Mineração de Dados (HAN; PEI; KAMBER, 2011), considerando explicitamente as conexões no desenvolvimento de predições ou descrições de dados. A ML está relacionada com a ARS pois não foca apenas na extração pura de dados, mas tenta entender a rede como um todo, destacando as condições que levam ao surgimento de conexões entre os vértices.

Na Predição de Links, a descoberta de conexões faltantes pode ser útil quando conexões dentro da rede são omissas, por exemplo, em redes criminosas (CLAUSET; MOORE; NEWMAN, 2008). Analisar como a rede evolui ao longo do tempo, porém, é uma tarefa mais complicada, já que descobrir os fatores que influenciam as mudanças ao longo dessa evolução (poucos desses fatores são triviais e evidentes na rede) é necessário para desenvolver uma solução precisa e que faça uma predição satisfatória das conexões que têm grandes chances de se formar no futuro (HUANG, 2006; MURATA; MORIYASU, 2008). Entretanto, o trabalho de predizer conexões no futuro pode envolver mais que "descobrir as possíveis novas conexões em $t + 1$, dado o histórico da rede até o tempo t", isto é, pode englobar a predição para um outro intervalo de tempo entre $t + 1$ e $t + \Delta$. A definição mais comum para predições do ponto de vista temporal pode ser resumida como: "Dado o histórico condensado da rede, do tempo 1 ao tempo t, predizer as possíveis conexões entre os tempos $t + 1$ e $t + \Delta$".

A definição do problema de Predição de Links formulada por LIBEN-NOWELL; KLEIN-

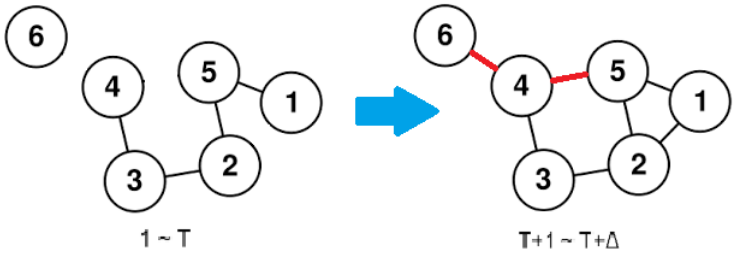

Figura 3.1: Exemplo gráfico da definição tradicional de Predição de Links dada por (LIBEN-NOWELL; KLEINBERG, 2003). O primeiro grafo representa a configuração final da rede, até o tempo t, e o objetivo é predizer a configuração da rede de t + 1 até t + Δ.

BERG (2003) e representada graficamente pela Figura 3.1 tem servido como referência para diversos trabalhos (WANG et al., 2015). Os autores ainda demonstraram a viabilidade da predição de conexões a partir da utilização de métodos baseados exclusivamente na estrutura da rede.

Existem basicamente dois tipos de entradas que alimentam algoritmos para o problema de Predição de Links: i) características dos membros da rede (informações ou contextos) e ii) informações estruturais e topológicas sobre a rede. Essas informações servem para o cálculo de métricas, baseadas em preceitos sociológicos, que de alguma forma conseguem extrair e representar padrões, de forma mensurável e de modo a servir de base para valorar o grau de similaridade ou proximidade entre dois indivíduos, ou seja, pares de vértices dentro da rede. De posse dessas métricas, métodos de aprendizagem podem ser aplicados na tarefa final de predizer futuros relacionamentos. Existem vários métodos que produzem resultados satisfatórios para a predição de relacionamentos. Esses métodos costumam seguir três abordagens (XIANG, 2008; VARTAK, 2008; Lü; ZHOU, 2011): a baseada em similaridade entre vértices, baseada em padrões estruturais ou baseada em modelos probabilísticos. Normalmente, um valor é atribuído a cada par de vértices e quanto maior esse valor, maior a probabilidade de que os vértices em questão venham a se conectar no futuro.

Como mencionado, as informações e modelos gerados (por exemplo, informações de distância e vizinhança de vértices e modelos probabilísticos) a partir dessas análises são fundadas em preceitos sociológicos, e podem ser utilizadas para alimentar sistemas de aprendizagem. YIN et al. (2010) descrevem alguns desses preceitos, resumidos a seguir:

- **Relacionamentos em comum**: o número de relacionamentos em comum que dois membros de uma sociedade possuem torna maior a probabilidade de que tais pessoas venham a se conhecer no futuro. Os indivíduos que pertencem aos relacionamentos em comum intermediam informações que fazem com que os dois membros em questão possam vir a se conhecer;

- **Homofilia**: dois indivíduos que possuem mais características em comum um com o outro, ao invés de com outros indivíduos da sociedade, tem mais chances de estabelecer uma relação no futuro;

- **Raridade**: características comuns entre os membros de uma sociedade tendem a ser menos importantes e a ter menos influência no surgimento de um relacionamento entre dois indivíduos. Características mais peculiares, pelo contrário, tendem a interferir mais no surgimento de tais relacionamentos, por exemplo, dentro de uma comunidade de programadores, simpatizantes de uma mesma linguagem de programação tendem a criar mais relacionamentos entre si;

- **Exclusividade**: indica que os poucos relacionamentos de certos indivíduos podem ser bem mais relevantes (sólidos) e duradouros do que os relacionamentos de alguém popular, que costumam ser menos importantes e intermitentes.

- **Influência social**: características de membros de uma sociedade que se relacionam com um certo indivíduo tendem a influenciar outro indivíduo qualquer, que possui as mesmas características, a se relacionar com o indivíduo específico. Por exemplo: se muitos amigos de um indivíduo A gostam de programar, é mais provável que outro indivíduo B, que também goste de programar, venha a se relacionar com A;

- **Conexões preferenciais**: fenômeno oposto ao de exclusividade, sugere que membros de uma sociedade tendem a criar relações com outros mais populares (BARABASI; ALBERT, 1999), ou seja, vértices mais centrais dentro do grafo;

- **Proximidade social**: a proximidade entre dois vértices (indivíduos) dentro de um grafo (distância entre os vértices) tende a ser um fator relevante para o surgimento de uma aresta (relacionamento) entre eles. Uma característica interessante é a expressa na teoria do mundo pequeno, que afirma que a distância entre dois usuários em uma rede social, através da relação “amigo do amigo”, é tipicamente pequena.

3.2 Estratégias para Predição de Links

Existe uma variedade de abordagens presentes na literatura que visam lidar com a predição de novos relacionamentos em redes sociais (Lü; ZHOU, 2011). Essas abordagens variam de resultados puros da análise de similaridade, topográfica ou utilizando modelos probabilísticos até a utilização de técnicas de aprendizagem de máquina (XIANG, 2008). Comumente, o resultado dessas análises é transformado em uma pontuação, ou *score*, retornando-se os maiores obtidos ou alimentando-os como entrada para algum algoritmo de aprendizagem. Além disso, os métodos que serão apresentados podem ser classificados quanto a considerarem ou não o tempo na predição. Eles podem ser chamados de abordagens estáticas (consideram a rede como uma estrutura que não muda de estado ao longo do tempo) ou temporais (tentam entender como a rede evolui). É importante lembrar que uma rede é uma estrutura evolutiva e que considerar o tempo pode melhorar a qualidade das predições.

3.2.1 Padrões estruturais

Tanto os padrões estruturais quanto outras características topológicas de um grafo costumam não ser visíveis à primeira vista, porém são relativamente simples de extrair automaticamente, quando se tem uma representação mais adequada, como, por exemplo, uma matriz de adjacência do grafo (visto no capítulo anterior). A partir disso, é possível descobrir características globais ou locais, como distâncias ou vizinhança (dois vértices que têm vizinhos em comum tem mais chances de se relacionarem (JIN; GIRVAN; NEWMAN, 2001; DAVIDSEN; EBEL; BORNHOLDT, 2002)), que são base para as métricas que serão apresentadas a seguir:

3.2.1.1 Vizinhos em comum (*Common neighbors*)

Métrica mais simples que serve como base para a maioria das outras métricas calculadas e baseia-se na idéia de que dois vértices tem maior probabilidade de se conectarem se o número de vizinhos em comum também for maior (NEWMAN, 2001b). Seja $\Gamma(i)$ o conjunto de vértices vizinhos a um vértice *i* qualquer, e *i* e *j* vértices de um grafo, então:

$$CN(i,j) = |\Gamma(i) \cap \Gamma(j)|. \tag{3.1}$$

3.2.1.2 Conexões preferenciais (*Preferential attachment*)

Uma conexão preferencial representa o fato de que membros populares de uma sociedade do mundo real tendem a criar mais relações quando comparados a outros membros. A partir deste pensamento, isto é, vértices que detém um grande número de conexões tendem a criar novas conexões mais rapidamente (NEWMAN, 2001b; BARABáSI et al., 2002), verificou-se que a probabilidade de uma conexão entre dois vértices pode ser calculada e é proporcional ao produto do número de vizinhos $|\Gamma|$ que cada vértice possui.

$$PA(i,j) = |\Gamma(i) * \Gamma(j)|. \tag{3.2}$$

3.2.1.3 Coeficiente de Jaccard (*Jaccard's coefficient*)

Métrica que calcula a razão entre o número de vizinhos em comum e o número total de vizinhos distintos de dois vértices *i* e *j*, ou seja, é a porcentagem de vizinhos comuns entre *i* e *j* em relação ao total de vizinhos distintos de *i* e de *j*. Esta porcentagem representa a probabilidade de se escolher um vizinho comum a *i* e *j*, e é dada pelo coeficiente.

$$JC(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}. \tag{3.3}$$

3.2.1.4 Proximidade de Adamic-Adar

Baseado no índice de Adamic e Adar e no número de vizinhos em comum, essa medida representa o grau de proximidade entre dois vértices i e j, fazendo uma simples contagem dos vizinhos em comum e atribuindo pesos maiores aos vizinhos com menor número de conexões (ADAMIC; ADAR, 2003; LU; ZHOU, 2010).

$$AA(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{log(|\Gamma(k)|)}. \tag{3.4}$$

Esse primeiro conjunto de métricas analisa aspectos mais locais (vizinhança dos vértices). As métricas que serão mostradas adiante também são estruturais, mas consistem na análise de caminhos para uma exploração mais global dos padrões estruturais da rede ao invés de uma análise local entre dois vértices. As próximas métricas consideram a análise conjunta de tipos de caminhos entre dois vértices dentro de um grafo e refinam o conceito de menor distância, partindo do princípio que se existem caminhos que ligam indiretamente dois vértices, é provável que eles venham a se conectar em algum momento no futuro.

3.2.1.5 Menor caminho (*Shortest path*)

Considera que quanto menor a distância entre dois vértices, maior a probabilidade de eles se conectarem. O menor caminho entre dois vértices em um grafo é o caminho com menor custo de travessia, ou seja, a soma dos pesos das arestas é a menor possível (SA; PRUDENCIO, 2011). Como esses dois valores são inversamente proporcionais, o valor é o negativo da menor distância.

$$SP(i,j) = -(min\_distance(i,j)). \tag{3.5}$$

3.2.1.6 Coeficiente de Katz

Valor resultante da soma de todos os caminhos entre dois vértices, ponderada pelo tamanho dos caminhos, com maiores pesos dados aos caminhos mais curtos (KATZ, 1953; WANG et al., 2015). Existem duas versões dessa métrica, sendo elas a sem peso, onde cada caminho tem peso um, e com peso, onde cada caminho tem peso igual ao número de vezes que o relacionamento foi observado. A versão sem pesos da métrica é dada a seguir (WANG et al., 2015):

$$Katz(i,j) = \sum_{l=1}^{\infty} \beta^l * |paths_{i,j}^{<l>}| = \beta^1(A_{ij}^1) + \beta^2(A_{ij}^2) + \beta^3(A_{ij}^3) + \ldots, \tag{3.6}$$

onde $paths_{i,j}^{<l>}$ é o conjunto de todos os caminhos de tamanho l que tem i como vértice inicial e j como vértice final e $0 \leq \beta \leq 1$ pondera a relevância dos caminhos de acordo com seus tamanhos.

A forma matricial dessa métrica pode ser obtida através da seguinte fórmula:

$$Katz(M) = (I_n - \beta * M)^{-1} - I_n, \tag{3.7}$$

onde M é a matriz de adjacência da rede e I_n é a matriz identidade de ordem n (mesma ordem de M). De acordo com DUNLAVY; KOLDA; ACAR (2011), a métrica de Katz é uma das melhores para a Predição de Links, por ter um desempenho superior em relação a várias outras abordagens.

3.2.1.7 *Hitting time (HT) e Commute Time (CT)*

Hitting time é o número de passos esperado que será gasto, através de uma caminhada aleatória, ao sair de um vértice i e chegar a outro vértice j. Uma caminhada aleatória, ou *random walk*, é feita a partir de um vértice i, onde se escolhe aleatóriamente um vizinho k e a partir de k, novamente se escolhe outro vértice, iterando sobre esses passos até que se chegue no vértice j desejado. Essa propriedade é característica de cadeias de Markov (SPITZER, 2013). Como $HT_{i,j}$ e $HT_{j,i}$ geralmente não são simétricos, costuma-se calcular o *Commute time* normalizado (CT), ou seja, *hitting times* de ida e volta (LIBEN-NOWELL; KLEINBERG, 2003):

$$CT(i,j) = -((HT_{i,j} * \pi_j) + (HT_{j,i} * \pi_i)). \tag{3.8}$$

A probabilidade estacionária π de um vértice é um valor proporcional a quantidade de vezes que o vértice é selecionado dentro de uma caminhada aleatória, sendo utilizado como fator de normalização.

3.2.1.8 *Rooted PageRank*

O algoritmo de PageRank (PAGE et al., 1998) é muito utilizado para a recuperação de informação na *internet*. Sua proposta inicial foi indicar o quanto uma página *web* é importante na rede. O *Rooted PageRank* é uma variação para predição de *links*. A partir de um vértice i, deseja-se, iterativamente, chegar a outro vértice j. Em cada iteração, é escolhido aleatoriamente um vizinho de i, mas com uma probabilidade α de se retornar ao vértice inicial (reiniciar o processo) e probabilidade 1 - α de se mover para o vizinho escolhido. Após a caminhada parar em j (ficar estável em j, ou seja, não existe mais movimentação para vértices vizinhos), o *score* da métrica é dado pela probabilidade de se manter estável. Para um dado par de vértices i e j, o *score* é calculado da seguinte maneira:

$$RPR(i,j) = (1-\alpha) * (I - (\alpha.T))^{-1}, \tag{3.9}$$

onde $T = D^{-1} * M$, D é uma matriz diagonal tal que $D(i,i) = \sum_j M(i,j)$ e M é a matriz de adjacência.

3.2.1.9 *SimRank* (JEH; WIDOM, 2002)

É uma definição recursiva resultante da medida de similaridade dos vizinhos de dois vértices i e j: i e j têm maior probabilidade de se relacionarem se seus vizinhos possuirem características em comum. Este conceito de vizinhança de um vértice compreende todos os outros que podem ser alcançados a partir do vértice inicial (isto é, exista um caminho entre eles).

$$SimRank(i,j) = \frac{\gamma * (\sum_{a \in \Gamma(i)} \sum_{b \in \Gamma(j)} SimRank(a,b))}{|\Gamma(i)| * |\Gamma(j)|}, \tag{3.10}$$

onde $0 \leq \gamma \leq 1$ e $SimRank(i,i) = 1$ (caso base).

Todas as métricas explicadas até agora, apesar de terem sido apresentadas em seu formato original, possuem versões matriciais. Devido à ordem de grandeza dos dados e do tamanho das redes sociais e bases de dados atuais, é preferível utilizar métodos de fatoração de matrizes, por dois motivos principais: i) a redução do tamanho dos dados e ii) a redução do ruído (casos anormais) (LIBEN-NOWELL; KLEINBERG, 2007). Fatorar as matrizes de entrada antes de aplicá-las à alguma métrica, técnica conhecida como *low-rank approximation*, ajuda a melhorar os resultados e diminuir o seu custo computacional. Alguns tipos de *low-rank approximations* (fatorações) serão explicados a seguir:

3.2.1.10 *Singular Value Decomposition (SVD)*

Método que se baseia na matriz de adjacência M de um grafo para gerar uma segunda matriz M_k, diminuindo a esparsidade dentro da matriz original, uma espécie de redução de ruído. Essa redução se dá através do uso de decomposição em valores singulares, uma fatoração de matriz computada por uma rotação V^* e aplicação de escala $\sum$. Esta fatoração é conhecida como *SVD*. Para obter M_k, a matriz M deve ser decomposta nos fatores a seguir:

$$M_k = U \sum V^*, \tag{3.11}$$

onde U é uma matriz unitária, $\sum$ é uma matriz diagonal com números reais não negativos na diagonal principal e V^* (conjugada transposta de V) é uma matriz unitária. As três matrizes reduzidas (estas matrizes são de ordem k, onde k é menor que a ordem de M) resultantes da decomposição de M são geradas de modo a preservar informações da estrutura global da rede. Elas também permitem recompor M (a matriz recomposta é chamada de M_k), de modo mais eficiente e sem a presença de ruídos, mantendo a maioria da estrutura de M em uma estrutura mais simplificada (XIANG, 2008). Após obter M_k, deve-se aplicar os métodos descritos anteriormente sobre essa nova matriz, o que se chama de aproximação com *rank* reduzido (por um fator k).

3.2.1.11 *Truncated Singular Value Decomposition (TSVD)*

É uma *low-rank approximation* bastante conhecida e utilizada para decompor matrizes ou tensores. Apesar de uma SVD reduzir a ordem da matriz para k, qual seria o melhor k a ser utilizado na decomposição? DUNLAVY; KOLDA; ACAR (2011) citam uma versão truncada (*Truncated Singular Value Decomposition* (TSVD)) e que utiliza $k = 1$, dada pela seguinte aproximação:

$$M_k \approx U_k \sum_k V_k, \tag{3.12}$$

onde U_k e V_k compreendem as k primeiras colunas de U e V e $\sum_k$ é uma submatriz principal quadrada de ordem k de $\sum$. Essa aproximação pode também ser reescrita como uma soma de matrizes de *rank* 1:

$$M_k \approx \sum_{k=1}^{K} \sigma_k u_k v_k, \tag{3.13}$$

onde u_k e v_k são as k-ésimas colunas de U e V. Pode-se utilizar, então, a matriz S resultante do TSVD como base para a aplicação de métricas para predizer *links*:

$$S = U_k \sum_k V_k. \tag{3.14}$$

A Figura 3.2 mostra graficamente como ocorre a aproximação da matriz original através dos valores truncados.

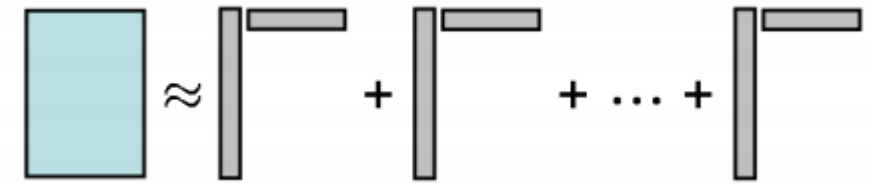

Figura 3.2: Representação gráfica da soma dos valores truncados de U e V (u_k e v_k). Fonte: (DUNLAVY; KOLDA; ACAR, 2011).

Low-rank approximations baseadas em SVD tiveram desempenho positivo em várias aplicações, como, por exemplo, aprendizado para reconhecimento facial (ZHANG; LI, 2010). Esta aproximação é chamada de *"low-rank approximation: matrix entry"* em (LIBEN-NOWELL; KLEINBERG, 2007).

Por fim, existem alguns outros tipos de abordagens de alto nível (LIBEN-NOWELL; KLEINBERG, 2007) que utilizam informações sobre os dados do problema (metadados) ou alteram os dados iniciais (como o grafo que representa a rede) em conjunto com os métodos explicados anteriormente, resultando em métricas mais complexas, porém mais precisas. Como exemplos, pode-se citar duas técnicas conhecidas, a Clusterização (LIBEN-NOWELL; KLEINBERG, 2007) e a identificação de Bigramas Escondidos (WANG; MANNING, 2012):

3.2.1.12 *Clustering factorization*

Consiste em melhorar a predição através da remoção de arestas menos importantes, baseado na própria métrica que está sendo utilizada. Esta remoção ocasiona a formação de aglomerados dentro do grafo final, por isso o nome de *Clustering*. Primeiramente, define-se um limiar l para a remoção de arestas: se a pontuação de uma aresta for menor que l, a aresta será removida, até que a fração de (1 - l) de arestas removidas seja atingida.

Para qualquer métrica utilizada, computa-se o *score*(i, j) para todas as arestas presentes no grafo, removendo, após isso, as 1 - l arestas com piores *scores* computados. O fator escolhido é arbitrário, mas de modo que não distorça a formação inicial do grafo, fazendo-o perder dados relevantes. Ainda, os *scores* são normalizados para valores entre 0 e 1 e $0 \leq l < 1$. Após a remoção das 1 - l arestas com menores *scores*, a métrica utilizada inicialmente deve ser recalculada para todas as arestas que sobraram, aumentando assim a confiança dos resultados, já que foram excluídas arestas menos relevantes. Por fim, a estrutura gerada é mais simples e as conexões presentes, mais relevantes a análise em curso, podendo os pares de vértices serem avaliados dentro dessa versão mais simples da rede (LIBEN-NOWELL; KLEINBERG, 2003).

3.2.1.13 Bigramas escondidos (WANG; MANNING, 2012)

Baseado no problema original de modelagem de linguagens e processamento de linguagens naturais, esta abordagem consiste em melhorar a predição tentando mapear casos reais com outros semelhantes que já foram classificados. Como a maioria dos exemplos de teste (no caso, pares de vértices (i, j) e se um *link* existirá entre eles) não são encontrados nos conjuntos de treinamento, tenta-se encontrar outros vértices k e l com características semelhantes a i e j, imaginando-se que terão comportamentos parecidos e criando um *score* (UB) que reflita essa semelhança.

Dado um conjunto S de vértices semelhantes a i, utilizando como métrica qualquer um dos métodos já apresentados, pode-se computar UB como o tamanho do conjunto interseção entre os vizinhos de j e S (não ponderado):

$$UB(i,j) = |\{z : z \in \Gamma(j) \cap S\}|. \tag{3.15}$$

Pode-se também ponderar o valor final somando-se os *scores* de todos os elementos do conjunto interseção entre os vizinhos de j e S:

$$UB(i,j) = \sum_{z \in \Gamma(j) \cap S} score(i,z). \tag{3.16}$$

Métricas estruturais são independentes do domínio que está sendo estudado, podendo ser aplicadas a qualquer tipo de rede, isto porque as operações são feitas sempre sobre representações de grafos. A grande desvantagem é o custo computacional de tempo e memória, pois as redes

estudadas têm um grande número de individuos e os dados costumam ser bidimensionais.

3.2.2 Similaridade entre vértices

Medidas de similaridade entre vértices tentam encontrar semelhanças presentes em dados sobre os indivíduos da rede, para quantificar o quão parecidos eles são. Essas semelhanças são informações contextuais sobre esses indivíduos, ou seja, atributos relevantes dentro do cenário e que caracterizam o indivíduo dentro da rede. Em uma rede virtual, como o Facebook[1], atributos relevantes podem ser a localização, interesses pessoais, entre outros. Porém, esses atributos não são tão relevantes quando se está analisando uma rede de coautoria, por exemplo (atributos mais importantes seriam: áreas de pesquisa, número de artigos publicados, entre outros). Daí a importância de se analisar o contexto no qual um indivíduo está inserido antes de iniciar a quantificação dessas semelhanças.

Os atributos a_i de um indivíduo podem ser reunidos em um vetor $\vec{v_c} = \{a_1, a_2, ..., a_n\}$ de características, que são utilizadas como entrada para métricas de similaridade, medindo o quão parecido aquele indivíduo é com outros da rede analisada, com relação aos atributos que possui. Esta técnica é aplicada em Predição de Links analisando as maiores similaridades entre vértices (similaridades maiores que um limiar, definido manualmente ou com a ajuda de alguma técnica de aprendizagem de máquina) (XIANG, 2008) e indicando essas maiores similaridades como possíveis relacionamentos futuros, se aplicável.

Os dados utilizados como atributos dos indivíduos de uma rede variam com o contexto da rede em questão. Assim sendo, a similaridade também pode variar de acordo com aquele contexto. Por isso, as métricas costumam mudar conforme o problema abordado, como, por exemplo, informação mútua (HINDLE, 1990), inverso da distância euclidiana, similaridade de cosseno e o coeficiente de Dice (SALTON; MCGILL, 1983).

De maneira geral, a similaridade entre vértices i e j pode ser calculada pela razão entre a quantidade de informação que descreve as semelhanças entre i e j e a quantidade de informação necessária para descrever i e j, individualmente:

$$sim(i,j) = \frac{logP(common(i,j))}{logP(description(i,j))}, \tag{3.17}$$

onde *common*(i, j) e *description*(i, j) estão atreladas a um domínio de aplicação específico (XIANG, 2008). Alternativamente, pode-se utilizar métricas baseadas na união de características dos vértices. Desse modo, as informações numéricas dos atributos devem ser agregadas matematicamente através de alguma função, como soma, média, máximo ou mínimo, para combinar todos os atributos em uma única medida (HASAN et al., 2006).

A principal vantagem de métricas baseadas em similaridades se dá no fato de conseguirem identificar propriedades intrínsecas aos indivíduos da rede, dificilmente vistas por uma análise

[1] www.facebook.com

estrutural. Entretanto, as informações sobre indivíduos são pessoais e privadas e nem sempre estão disponíveis publicamente.

3.2.3 Modelos probabilísticos

Para tentar resolver satisfatoriamente o problema de Predição de Links, as abordagens apresentadas até agora se utilizam de informações estruturais ou contextuais da rede. Modelos probabilísticos, por outro lado, procuram extrair uma série de parâmetros que servirão de base para criar um modelo probabilístico. Esse modelo é capaz de abstrair as informações relevantes de membros da rede e de suas conexões e é utilizado na predição de novos *links* (GETOOR; DIEHL, 2005; YU; HAN; FALOUTSOS, 2010; Lü; ZHOU, 2011). O destaque vai para as redes Bayesianas (POTGIETER et al., 2009) e as redes de Markov (WANG; SATULURI; PARTHASARATHY, 2007), que têm um bom desempenho, mas ainda não consideram várias informações relacionais implícitas nas redes, como, por exemplo, hierarquias dentro de uma comunidade (Lü; ZHOU, 2011). A principal idéia é extrair de um grafo $G(V,A)$ um conjunto θ de parâmetros que servirão de base para que o modelo possa abstrair outras características. A existência de um *link*, então, é dada pela probabilidade condicional $P((i,j) \in A|\theta)$, isto é, a probabilidade de i e j se relacionarem dado que os parãmetros de θ foram observados.

Para melhor aproveitar as informações presentes em dados com estruturas relacionais, existem dois *frameworks* conhecidos: Modelo Probabilístico Relacional e Relacionamento de Entidades Probabilísticas Acíclicas Direcionadas:

3.2.3.1 Modelo Probabilístico Relacional

Apresentado inicialmente por POOLE (1993) e KOLLER; PFEFFER (1998), Modelo Probabilístico Relacional (MPR) representa uma distribuição de probabilidade conjunta sobre a rede. Assim, as propriedades de um objeto dependem probabilisticamente das propriedades dos objetos relacionados. MPR também tenta trabalhar com modelos compactos de grafos (FRIEDMAN et al., 1999), utilizando uma série de grafos para modelar a distribuição de probabilidade, ao contrário das modelagens tradicionais de grafos. A idéia principal é trabalhar com três grafos, o grafo de dados G_D, o de modelo, G_M e o de inferência, G_I (NEVILLE, 2006).

G_D é utilizado para representar a rede em questão, onde cada vértice e aresta está associado a um tipo de elemento, e cada tipo representa um conjunto de atributos. O grafo G_M é usado para representar dependências entre atributos diferentes. Atributos de um tipo podem depender probabilisticamente de outros do mesmo tipo ou de atributos relacionados. O grafo G_I, por fim, representa as dependências probabilísticas entre todas as variáveis de um conjunto de teste, usado para modelar um novo $G_{D'}$ e no grafo G_M. Para cada par elemento-atributo em $G_{D'}$, recupera-se sua distribuição de probabilidade condicional a partir do grafo G_M. Para a Predição de Links, cada par é representado por uma aresta e um atributo binário que indica a sua existência ou não (TASKAR et al., 2004).

MPR pode ser classificado como redes Bayesianas relacionais (GETOOR, 2001) ou redes de Markov relacionais (TASKAR; ABBEEL; KOLLER, 2002), se utilizam as redes Bayesianas ou de Markov para a geração do modelo que abstrai a rede analisada.

3.2.3.2 Relacionamento de Entidades Probabilísticas Acíclicas Direcionadas (DAPER, em inglês)

Framework baseado no modelo entidade-relacionamento que é capaz de manipular a representação de parâmetros fundamentada em uma abordagem Bayesiana, fornecendo uma personalização das entidades e dos relacionamentos do grafo (HECKERMAN; MEEK; KOLLER, 2004).

O modelo consiste em seis tipos de classes, sendo elas (Lü; ZHOU, 2011; XU et al., 2012):

1. Entidades, que especificam os indivíduos;
2. Relacionamentos, que representam as conexões entre indivíduos;
3. Atributos, que guardam os atributos dos indivíduos e conexões;
4. Classes de arco, que representam as dependências probabilísticas entre atributos;
5. Distribuição local, que geram as distribuições locais de um atributo e
6. Restrição, que determinam como gerar o grafo de inferência.

O DAPER é normalmente usado em situações nas quais a estrutura relacional encontra-se em um estado de incerteza, sendo mais expressivo que o MPR (HECKERMAN; MEEK; KOLLER, 2007; Lü; ZHOU, 2011).

3.2.4 Estratégias temporais

As modelagens apresentadas se baseiam na definição de LIBEN-NOWELL; KLEINBERG (2003), utilizando uma estrutura única como fonte de dados e para extrair características. Nesse caso, como mencionado na Seção 3.1, Página 39, as abordagens são consideradas estáticas. A seguir, serão vistas abordagens temporais, ou seja, que consideram o tempo como fator na predição de novos relacionamentos e que redes funcionam como estruturas evolutivas. Uma definição resumida para esta vertente poderia ser: "Dado o histórico detalhado da rede, do tempo 1 ao t, identificar as mudanças e transições entre períodos para predizer o que irá ocorrer em $t + \Delta$". A Figura 3.3, Página 51, representa graficamente a definição desta vertente.

A forma como as várias observações ao longo do tempo são tratadas varia. Pode-se manter o tempo como um fator explícito e levar em consideração cada observação individualmente para o cálculo do *score* final ou agregar os dados de tais observações utilizando, por exemplo, alguma das técnicas de agregação explicadas a seguir.

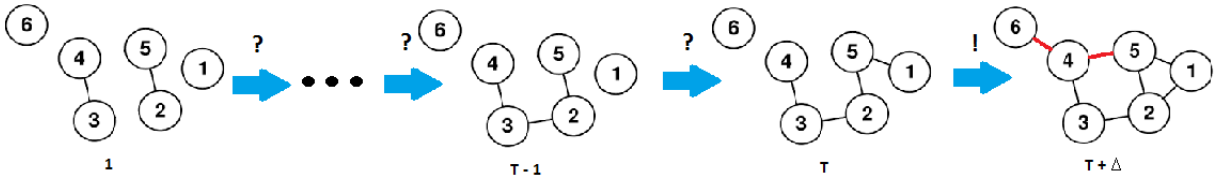

Figura 3.3: Representação gráfica do problema de predição temporal de *links*. Importante frisar a necessidade da análise das transições (marcadas com "?") entre os períodos de observação da rede para predizer relacionamentos futuros.

3.2.4.1 Agregação de tensores

DUNLAVY; KOLDA; ACAR (2011) detalharam esta abordagem, que consiste em agregar as matrizes de entrada em uma única matriz que contém os dados consolidados de todos os períodos de tempo. A entrada é constituída de T matrizes M x N, onde T é o número de fatias de tempo, e M e N as dimensões das matrizes de adjacência. Uma estrutura Z desse tipo pode ser chamada de tensor e é representada formalmente como se segue:

$$Z(i,j,t) = \begin{cases} 1, \textit{ se existe aresta que liga } i \textit{ a } j \textit{ no tempo } t; \\ 0, \textit{ caso contrário.} \end{cases} \tag{3.18}$$

Vale ressaltar que é simples adaptar essa representação para arestas com pesos, como a seguir:

$$Z(i,j,t) = \begin{cases} w_{i,j,t}, \textit{ se existe aresta que liga } i \textit{ a } j \textit{ no tempo } t; \\ 0, \textit{ caso contrário.} \end{cases} \tag{3.19}$$

Existem diversas formas de se agregar matrizes desse tipo, que podem ser divididas basicamente em dois tipos: as que consideram todas as fatias de tempo como tendo a mesma importância e as que dão importância à recenticidade dos dados.

O primeiro e mais simples modo de condensar os dados das matrizes é utilizar como *score* a soma de todos os valores encontrados ao longo do tempo, isto é:

$$S(i,j) = \sum_{t=1}^{T} Z(i,j,t). \tag{3.20}$$

A matriz Z' resultante é chamada de tensor colapsado (TC), já que a dimensão do tempo foi condensada. Esse tensor colapsado é obtido de forma similar por LIBEN-NOWELL; KLEINBERG (2007). DUNLAVY; KOLDA; ACAR (2011) citam ainda outra maneira de combinar os dados de um tensor, onde os dados mais antigos tem um peso menor no resultado final. Essa forma colapsada é chamada de tensor colapsado valorado (TCV) e é obtida pela

seguinte função:

$$S(i,j) = \sum_{t=1}^{T} Z(i,j,t) * (1-\theta)^{T-t}, \tag{3.21}$$

onde (1 - θ)$^{T-t}$ é o peso atribuído a cada Z(i, j, t) e $0 \leq \theta \leq 1$ pode ser escolhido arbitrariamente ou empiricamente (θ que obteve melhores resultados durante a validação, por exemplo). Percebe-se que quando $t = T$, o peso atribuído é igual a 1 (o maior peso possível). Para todo $0 < t \leq T$, o peso cresce exponencialmente até se igualar a 1. A Figura 3.4 mostra o gráfico da função f(t).

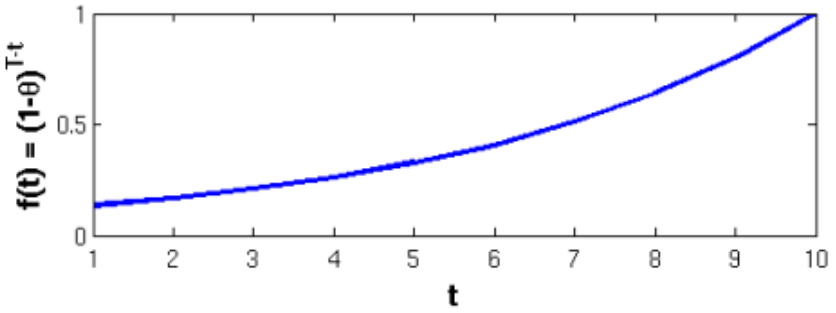

Figura 3.4: Visualização gráfica do crescimento da função mostrada, para T = 10 e θ = 0.2. Quando t = T, o peso será 1 (Fonte: (DUNLAVY; KOLDA; ACAR, 2011)).

O resultado da agregação dos valores de um tensor pode ser utilizado como *score* final para um determinado par de vértices. Seja um tensor Z(i, j, t) cujos índices (i, j, t) armazenam o *score* relativo ao par (i, j) no tempo t, a função descrita na Figura 3.4 pode ser utilizada para agregar esses valores e calcular o *score* final para o par (i, j), considerando o tempo.

3.2.4.2 Decomposição canônica/de fatores paralelos (CP)

Diferentemente das outras técnicas de *low-rank approximation*, essa decomposição (CARROLL; CHANG, 1970) não necessita que a matriz tridimensional (ou tensor) original seja condensada. Isso permite que o fator tempo permaneça explícito durante todo o processo. Isso ocorre porque essa decomposição é uma generalização do SVD para T dimensões. Dado um tensor Z de ordem M x N x T como descrito anteriormente, a matriz tridimensional pode ser decomposta da seguinte maneira:

$$Z \approx \sum_{k=1}^{K} \lambda_k \, a_k \, o \, b_k \, o \, c_k. \tag{3.22}$$

Ou seja, Z pode ser aproximado como um somatório, onde cada parcela é chamada de componente, a_k o b_k o c_k é o produto externo entre os vetores (fatores) a_k, b_k e c_k, $\lambda_k > 0$ é o peso atribuído à k-ésima decomposição e os três fatores de cada decomposição estão normalizados. Essa abordagem é análoga à decomposição de valores singulares truncada porque o tensor é fatorado em uma soma de tensores de *rank* 1, assim como o TSVD fatora uma matriz numa soma de matrizes de *rank* 1. Para efeito de interpretação, os fatores a_k e b_k representam os

relacionamentos entre os dois tipos de entidades presentes no grafo, enquanto c_k captura os comportamentos temporais presentes no tensor. A Figura 3.5 mostra graficamente como um tensor é fatorado utilizando a decomposição canônica.

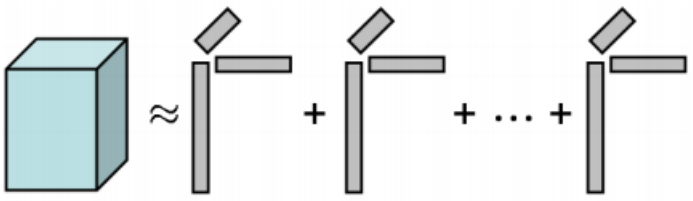

Figura 3.5: Representação gráfica da fatoração de um tensor utilizando decomposição canônica (Fonte: (DUNLAVY; KOLDA; ACAR, 2011)).

Para computar a matriz final para essa decomposição, DUNLAVY; KOLDA; ACAR (2011) enumeram duas maneiras: usando uma heurística ou predição temporal. A heurística consiste em quantificar os relacionamentos entre objetos utilizando o produto externo entre a_k e b_k, e definir o peso λ_k baseado nos T_Δ (o trabalho escolhe $T_\Delta = 3$) últimos anos. Eles então definem a medida de similaridade entre dois vértices i e j como a entrada $S(i, j)$ da seguinte matriz:

$$S = \sum_{k=1}^{K} \gamma_k \lambda_k \, a_k \, b_k^T, \tag{3.23}$$

onde $\gamma_k = \frac{1}{T_0} \sum_{t=T-T_0+1}^{T} c_k$.

Já a predição temporal consiste em utilizar a predição de Holt-Winters aditiva (ASSIS et al., 2013), pois é mais aplicável a dados com padrões distribuídos periodicamente, isto é, dados sazonais. Para uma predição de Δ fatias de tempo no futuro será gerado um tensor S de tamanho M x N x Δ:

$$S = \sum_{k=1}^{K} \lambda_k \, a_k \, o \, b_k \, o \, \gamma_k, \tag{3.24}$$

onde γ_k é um vetor de tamanho Δ que contém a predição para as próximas Δ fatias de tempo do método de Holt-Winters, com c_k como entrada. Novamente, a similaridade entre dois vértices i e j é obtida pelo valor de S(i, j).

3.2.4.3 Grafos variantes no tempo e Distância temporal

Redes sociais são estruturas dinâmicas, ou seja, invidíduos participam, competem, atraem outros indivíduos, desaparecem e afetam as conexões presentes dentro da rede. Ainda assim, definições, modelos ou métricas para a Predição de Links ainda captam, majoritariamente, informações estáticas de tais redes (SCELLATO et al., 2010; SANTORO et al., 2011).

O aumento da disponibilidade de conjuntos de dados de redes reais (por exemplo, *e-mails* ou metadados sobre publicações científicas), bem como a popularização de dispositivos

eletrônicos, como *smartphones*, têm formentado a pesquisa sobre Análise de Redes Dinâmicas (CARLEY, 2003; SCELLATO et al., 2010; SANTORO et al., 2011). Os primeiros trabalhos sobre transporte e redes tolerantes a atrasos (redes que não apresentam conexões fim-à-fim instantâneas, como, por exemplo, redes de comunicação) foram responsáveis pela criação de conceitos mais específicos, dentre os principais (SANTORO et al., 2011):

- **Jornada ou caminho temporal**: é um tipo de caminho específico cujas arestas não são seguidas uma após a outra, instantaneamente. Ao invés disso, induzem um determinado tempo de espera nos vértices intermediários;
- **Partida e Chegada**: de uma jornada são os tempos t_0 em que a jornada começa e t_k em que a jornada termina. Além disso, o comprimento temporal de uma jornada é definido como *Chegada - Partida*;
- **Conectividade temporal**: entre dois vértices i e j é denominada como a existência de uma jornada entre i e j.

Existem, dentro da ARS, algumas definições de redes temporais, ou seja, redes que tentam incorporar o conceito de tempo dentro da sua própria estrutura. Nas primeiras definições, assim como em (KEMPE; KLEINBERG; KUMAR, 2002), uma rede temporal pode ser descrita como um grafo G onde todas as arestas pertencentes à G são etiquetadas com um *timestamp*, isto é, o momento exato (baseado na granularidade de tempo definida para cada problema) quando essas arestas apareceram em G. A partir daí, os caminhos computados deveriam seguir a restrição de cronologia do aparecimento das arestas, ou seja, duas arestas de tempos distintos não podem constituir um caminho dentro de G (SCELLATO et al., 2010; SANTORO et al., 2011).

Um dos problemas dessa abordagem, apontado em (TANG et al., 2009), é que esse modelo não permite a análise da frequência de contatos entre vértices ou grupos de vértices, além de não dar suporte a vértices que estão temporalmente desconectados do grafo. Por fim, o foco de trabalhos que utilizam esse modelo é capturar mudanças locais dentro de redes sociais dinâmicas, enquanto TANG et al. (2009) tentam identificar aspectos mais globais dos processos de alteração da estrutura da rede. Baseadas nessa premissa, são apresentadas métricas que tentam computar a distância temporal entre dois vértices. A métrica principal é o tamanho médio do caminho temporal, que diz o quão rápido a informação se espalha na rede em questão. O tamanho médio do caminho temporal proposto retém a cronologia de aparecimento das arestas, ocorrências de arestas repetidas, tempo de contato entre vértices e a deleção de arestas.

Grafos variantes no tempo

Um grafo variante no tempo ou grafo temporal (SCELLATO et al., 2010) G_t^w pode ser representado como uma sequência de janelas de tempo de tamanho w, onde cada janela consiste em um *snapshot* do grafo naquele momento. Dado que a rede em questão começa em um tempo

t_0 e termina em T, o contato (qualquer tipo de relacionamento R) entre dois vértices i e j no tempo S é dado por $R^S_{i,j}$ e o grafo temporal $G^w_t(t_0,\mathrm{T})$ com N vértices é a sequência de grafos G_{t_0}, G_{t_0+w}, ..., G_T.

Cada G_t presente em G^w_t é definido como um conjunto V de vértices e um conjunto A de arestas. Dois vértices i e j pertencem à G_t se satisfazem a seguinte condição:

$$i, j \in V \leftrightarrow \exists R^s_{i,j},\ tal\ que\ t \leq S \leq t+w. \tag{3.25}$$

Dadas as definições de grafo temporal e contato entre dois vértices, é possível quantificar a medida de distância temporal:

- **Menor distância temporal**: Dados dois vértices i e j, a distância temporal (ver Figura 3.6) pode ser definida como o tamanho do conjunto de caminhos $p^h_{i,j}(t_{min},t_{max})$ que começam em i e terminam em j, passando pelos vértices n^t_1, n^t_2, ..., n^t_i, onde $t_{min} \leq \mathrm{t} \leq t_{max}$ é a fatia de tempo em que o vértice n é visitado e h é o número de fatias de tempo percorridas dentro do tempo t (SCELLATO et al., 2010; SANTORO et al., 2011). A menor distância temporal $d^h_{i,j}(t_{min},t_{max})$ é igual ao menor $p^h_{i,j}(t_{min},t_{max})$. Em outras palavras, a menor distância temporal é o menor número de fatias de tempo necessárias para uma informação sair de i e chegar em j.

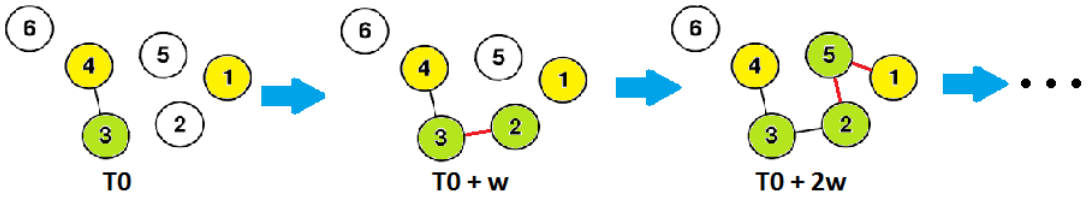

Figura 3.6: Exemplo do caminho formado entre os vértices 4 e 1 ao longo do tempo. As arestas marcadas em vermelho surgiram no intervalo de tempo indicado, até que existisse um caminho. Este caminho demorou $2w$ para se formar, então pode-se dizer que a distância temporal entre os vértices 4 e 1 é 2.

Para computar a menor distância, foi implementada uma busca em profundidade que calcula a distância temporal de i a todos os outros vértices do grafo. O algoritmo ainda mantém duas listas, para cada vértice: D, que guarda o número de fatias de tempo até chegar a um vértice, e R, contendo todos os vértices que podem ser alcançados. Partindo de t_{min}, verifica-se se o vértice i, a origem, está presente. Se sim, então, para todo vértice k que ainda não foi alcançado (R[k] = falso), é feita outra busca em profundidade para saber se existe um caminho entre ele e um vértice k' que já foi alcançado (R[k'] = verdadeiro) na fatia de tempo anterior. A profundidade máxima da busca é ditada por h, e se o vértice k pode ser alcançado, então R[k] passa a ser verdadeiro. Caso contrário, incrementa-se D[k] e repete-se os passos para todas as fatias de tempo subsequentes.

- **Métricas globais**: A partir do conceito de menor distância temporal, algumas métricas temporais são descritas por TANG et al. (2009) e SCELLATO et al. (2010). São elas:

  1. **Eficiência temporal**: $E^h_{T_{i,j}}(t_{min}, t_{max}) = \frac{1}{d^h_{i,j}(t_{min},t_{max})}$
  2. **Menor caminho temporal**: $L^h(t_{min},t_{max}) = \frac{1}{N(N-1)} \sum_{ij} d^h_{i,j}(t_{min}, t_{max})$
  3. **Eficiência temporal global**: $E^h_{glob}(t_{min},t_{max}) = \frac{1}{N(N-1)} \sum_{ij} E^h_{T_{i,j}}(t_{min}, t_{max})$

  Vale ressaltar que vértices desconectados (distância infinita) tem eficiência igual a 0, pois $1/\infty = 0$.

3.2.4.4 Séries temporais

Uma série temporal é uma sequência de dados discretos indexados pelos respectivos tempos de observação, em ordem cronológica (ESLING; AGON, 2012). Seja S um conjunto de dados discretos, indexados do tempo t_0 ao tempo T, a série temporal relativa à S pode ser definida como:

$$S = \{S_{t_0}, S_{t_1}, S_{t_2}, ..., S_t, ..., S_T\}, \tag{3.26}$$

onde S_t é o dado discreto observado no tempo t e $t_0 \leq t \leq T$. O *período* de uma série temporal é definido como a diferença entre dois tempos de observação consecutivos, isto é, $|t_1 - t_0|$ (o valor absoluto de $t_1 - t_0$). Séries temporais são aplicadas em Estatística, Predição de Links, Matemática financeira, Hidrologia, entre outras áreas (ESLING; AGON, 2012).

A análise de séries temporais compreende a aplicação de técnicas estatísticas, tendo como base uma série temporal, para extrair características importantes sobre os dados observados e criar modelos estatísticos (ESLING; AGON, 2012). Alguns modelos utilizados incluem modelos de médias móveis ponderadas exponencialmente (Alisamento Exponencial) (HOLT, 2004; ASSIS et al., 2013), modelos de Regressão linear (NASEEM; TOGNERI; BENNAMOUN, 2010) e ARMA e ARIMA (FARUK, 2010; VALIPOUR; BANIHABIB; BEHBAHANI, 2013). Já a predição de séries temporais, em suma, isola os componentes da série temporal coletados pelo modelo criado para calcular os valores que S assumirá no futuro, ou seja, S_{T+1}, S_{T+2} e assim por diante.

HUANG; LIN (2009) basearam-se em redes de vigilância e monitoramento de comunicações para utilizar uma abordagem temporal, construindo uma série temporal $S_t(i,\ j)$, onde $t_0 \leq t \leq T$, associada a cada par de vértices (i, j) da rede, que foi dividida em T *snapshots*, cada um representado pelas matrizes de adjacência (M_{t_0}, M_{t_1}, ..., M_T). As séries foram estruturadas de maneira que para um par (i, j), o t-ésimo termo da série fosse dado por $M_t(i, j)$. Outros trabalhos utilizam modelos mais complexos baseados em séries temporais, como o modelo ARIMA (FARUK, 2010; VALIPOUR; BANIHABIB; BEHBAHANI, 2013).

O cálculo das pontuações associadas aos pares de vértices da rede, para um par (i, j) qualquer, é dado pela probabilidade de que o valor predito do próximo elemento da série $S_t(i, j)$

seja maior que um, ou seja:

$$score(i,j) = P(\widehat{S}_{T+1}(i,\ j)\ >\ 1). \tag{3.27}$$

HUANG; LIN (2009) também desenvolveram uma abordagem híbrida, através da combinação dos *scores* gerados pelo método baseado em séries temporais e os gerados pelas abordagens estáticas (através do uso de métricas estruturais). Verificou-se que a abordagem híbrida, baseada em séries temporais, obteve um desempenho superior quando comparada às abordagens estáticas.

Além das estratégias citadas, existem outras que utilizam diversas características ou conceitos do mundo real para predizer novas conexões, como a proposta por POTGIETER et al. (2009): a geração de *scores* a partir de métricas financeiras, como retorno (fração de crescimento ou decaimento de um valor no tempo), média móvel (cálculo da média sobre os n elementos mais recentes de um conjunto de valores ordenados de acordo com o tempo) e recenticidade (indica quanto tempo se passou desde o último estabelecimento de conexão entre dois vértices).

Várias abordagens apresentaram melhora dos resultados em relação às abordagens estáticas apresentadas anteriormente (HUANG; LIN, 2009; POTGIETER et al., 2009; DUNLAVY; KOLDA; ACAR, 2011), o que mostra o grande ponto positivo de incluir o tempo dentro de uma modelagem para o problema de Predição de Links. O principal ponto negativo é que, se antes a ordem de grandeza dos dados era bidimensional, a adição do conceito de tempo faz com que os dados passem a ser tridimensionais, aumentando ainda mais o custo de tempo e memória, além da alta complexidade de implementação das soluções. Outra desvantagem é que a grande maioria das estratégias baseiam-se nas métricas mais tradicionais, que podem não ser tão representativas quando o problema é abordado do ponto de vista temporal.

3.2.4.5 Transições de tríades

Uma dificuldade de incorporar o conceito de tempo ao problema de predição temporal de *links* é que as abordagens temporais deixam um tipo de comportamento mais explícito: existem curtos períodos de tempo onde se tem intensa atividade de formação de novos relacionamentos sucedidos por longos períodos de inatividade. Esse padrão de comportamento acaba sendo introduzido na aprendizagem e altera bastante os resultados de abordagens clássicas utilizadas na análise de redes sociais. Uma solução para isso é aplicar os métodos vistos anteriormente às fatias de tempo e tentar descobrir o padrão de mudança dos resultados. Mesmo assim, não há um balanceamento quanto ao tamanho de uma fatia de tempo: pequenos intervalos detalham muito mais mudanças (o que pode fazer parecer com que a rede mude aleatoriamente) do que períodos maiores, o que dificulta o estudo da evolução da rede.

Motivados por essa problemática, JUSZCZYSZYN; MUSIAL; BUDKA (2011) propõem uma análise das mudanças estruturais que podem ser encontradas nos menores grupos não triviais que compõem a rede: as tríades. Enquanto a maioria dos métodos de ARS consistem em computar propriedades comuns, como distribuições, aglomerações, tamanhos de caminhos, entre

outros, as tríades são, por sua vez, uma análise mais sensível à estrutura do grafo. Uma tríade é considerada como o menor subgrafo não trivial que pode ser encontrado dentro de um grafo: um conjunto de três vértices que se podem se conectar de qualquer maneira.

Como já mencionado, o número de possíveis configurações (como as arestas estão dispostas dentro da tríade) de tríades varia dependendo se o grafo é direcionado ou não. Para grafos direcionados, existem 64 configurações de arestas possíveis, enquanto em um grafo não direcionado, existem apenas oito. JUSZCZYSZYN; MUSIAL; BUDKA (2011) enumeram cada configuração diferente de tríade em um índice, que será utilizado em uma Matriz de Transição de Tríades, já que esta não é indexada por vértices do grafo, mas as linhas e colunas representam cada tipo de tríade, e o índice (i, j) representa a probabilidade de uma tríade do tipo i passar a ser do tipo j, durante a transição entre t e $t + 1$. Uma TTM, para um grafo direcionado, seria uma matriz TTM_{64x64} enquanto para um grafo não direcionado seria uma matriz TTM_{8x8}.

	1	2	3	4	5	6	7	8
1	1	0	0	0	0	0	0	0
2	0	3/4	0	0	0	0	1/4	0
3	0	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0	0
5	0	0	0	0	1/2	0	0	1/2
6	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	1/3	2/3

Figura 3.7: Exemplo de uma TTM_{8x8}. A matriz é indexada pelos índices atribuídos aos tipos de tríades. As linhas representam o estado inicial das tríades enquanto as colunas representam o estado final. Os índices TTM(i,j) representam as probabilidades de transições. Note que a soma de todas as probabilidades de uma linha é sempre 1.

Observe, por exemplo, a TTM apresentada na Figura 3.7. Nela, pode-se identificar que ocorreram algumas transições. A primeira diz respeito a triades do tipo 2: há uma probabilidade de 75% de tríades desse tipo se manterem (TTM(2,2)) enquanto há 25% de chance de se tornarem do tipo 7 (TTM(2,7)), isto é, uma aresta entre dois vértices foi criada durante essa transição. A segunda é com relação a tríades do tipo 4, todas perderam uma aresta (TTM(4,1)), tornando-se tríades vazias (100% de probabilidade). Metade das tríades do tipo 5 se manteve, enquanto a outra metade ganhou uma aresta, passando a ser um 3-clique e, por fim, há 66.67% de chance de tríades do tipo 8 se manterem, enquanto há 33.33% de chance de perderem uma aresta, passando a ser do tipo 7.

Uma TTM computa a quantidade de tríades de cada tipo, encontradas em um grafo G. Essa contagem utiliza o algoritmo de *Triad Census* proposto por DAVIS; LEINHARDT (1967).

Algumas abordagens conseguem otimizar a complexidade de tempo, como em (MOODY, 1998), para O(n^2) ou em (BATAGELJ; MRVAR, 2001), para O(n), dadas certas condições, aproveitando o fato de que redes sociais mais próximas da realidade são bastante esparsas (maioria das tríades não possuem arestas). Para computar os *scores* do preditor, os autores propõem a seguinte abordagem: dado uma TTM, um grafo G e um par de vértices (i, j) o qual deseja-se avaliar e $i \neq j$, chama-se de $\Delta_{(i,\ j)}$ o conjunto de todas as tríades não-vazias (ou seja, com pelo menos uma aresta) que contém i e j. Olhar apenas tríades com pelo menos uma aresta é justificado pelo fato do grafo ser esparso, e de que a grande maioria das tríades são vazias e permanecerão vazias no futuro (JUSZCZYSZYN; MUSIAL; BUDKA, 2011).

Partindo desse ponto, deseja-se analisar a evolução de todas as tríades a quem i e j pertencem, a fim de determinar possíveis conexões que posssam ser formadas no futuro. Seja $e \in \Delta_{(i,\ j)}$ uma tríade, d o tipo de e e TTM(d) a linha da TTM contendo as probabilidades de transição para o tipo d. Os valores dessa linha podem ser divididos em dois conjuntos:

1. $TTM^0(d)$ como o conjunto das entradas que não possuem aresta entre i e j e
2. $TTM^1(d)$ como o conjunto das entradas que possuem aresta entre i e j.

Para computar o *score* de (i, j), determina-se primeiramente $\Delta_{(i,\ j)}$. Em seguida, para todo $e \in \Delta_{(i,\ j)}$, computa-se $TTM^0(d)$ e $TTM^1(d)$. Por fim, a pontuação final do par de vértices (i, j) é dada como:

$$scoreTTM(i, j) = \sum_{\Delta_{(i,\ j)}} \sum TTM^1(d). \tag{3.28}$$

Ou seja, o *score* do par (i, j) pode ser computado como o somatório de todos os valores dos conjuntos TTM^1 de todas as tríades onde (i, j) está presente e há uma aresta adjacente a i e j. As sugestões de novos relacionamentos são dadas ordenando-se as pontuações obtidas em ordem decrescente e retornando os n primeiros pares de vértices.

3.3 Avaliação das predições

Independentemente do modelo utilizado para a Predição de Links, as predições devem ser avaliadas para computar o desempenho do modelo. Na área de Aprendizagem de Máquina, existem problemas chamados de problemas de classificação. Soluções para problemas de classificação consistem em um modelo de aprendizagem que consegue distinguir os exemplos de entrada em uma ou mais classes diferentes, isto é, o modelo de aprendizagem classifica cada exemplo de entrada (Lü; ZHOU, 2011; SA; PRUDENCIO, 2011). As classes dos exemplos podem ser conhecidas ou não de antemão.

O problema de Predição de Links consiste na classificação de pares de vértices da rede em pares que apresentam um *link* entre os vértices e pares que não apresentam um *link* entre os vértices. A Predição de Links pode, portanto, ser considerada um problema de classificação

binária (WANG et al., 2015), pois existem duas classes que deseja-se diferenciar: i) a classe de exemplos que apresentam um *link* entre os vértices e ii) a classe de exemplos que não apresentam *links*. No caso da Predição de Links, todas as classes possíveis são conhecidas previamente. Também, na Predição de Links, há um interesse maior em predizer exemplos da classe que apresentam *links* entre os vértices, e, portanto, essa classe é chamada de classe positiva (ou classe 1). A classe dos exemplos que não apresentam *links* entre os vértices é chamada de classe negativa, ou classe 0. Quando existem muito mais exemplos de uma classe do que de outra, diz-se que há um desbalanceamento entre as classes.

Como visto ao longo deste capítulo, as abordagens apresentadas geram um *score* para cada exemplo de entrada, isto é, cada par de vértices (i, j) da rede em questão. A fim de decidir quais relacionamentos serão preditos, ou seja, pares de vértices que apresentaram *scores* mais representativos, existem dois tipos de avaliação utilizados (WANG et al., 2015): o método de avaliação supervisionada e o método de avaliação não supervisionada.

3.3.1 Método de avaliação supervisionada

Em uma abordagem supervisionada, o problema de Predição de Links é tratado como um problema de classificação binária. A rede é inicialmente dividida em dois períodos distintos e que não se intersectam. O primeiro período (observações do tempo t_0 ao tempo T), composto de observações mais antigas da rede em relação ao período subsequente, é utilizado como base para construir um conjunto chamado de conjunto de treinamento. O segundo período (observações do tempo $T+1$ ao tempo $T+\Delta$) é chamado de conjunto de testes.

Os exemplos que constituem o conjunto de treinamento são rotulados de modo que o modelo de aprendizagem identifique a que classe pertence cada exemplo. A seguir, o modelo computa todos os atributos relevantes de cada exemplo de entrada e associa essas características à classe rotulada no exemplo (SA; PRUDENCIO, 2011; WANG et al., 2015). Com base nesse conjunto de treinamento, o modelo aprende a diferenciar exemplos de cada classe. Ainda, pode existir um conjunto de validação, dentro do período de treinamento, que serve para ajudar a otimizar hiperparâmetros do modelo com relação a alguma métrica ou decidir quando o treinamento deve ser encerrado. O conjunto de testes é usado para avaliar o resultado das predições feitas pelo modelo treinado. No conjunto de testes, ao contrário do que ocorre com o conjunto de treinamento, os exemplos não são rotulados, de modo que o modelo treinado deve dizer a qual classe pertence cada exemplo.

Tendo em mãos os conjuntos de treinamento e testes, algum método de aprendizagem de máquina, por exemplo, *Support Vector Machines*, Árvores de decisão, Redes Neurais Artificiais, entre outros métodos (LICHTENWALTER; LUSSIER; CHAWLA, 2010; BENCHETTARA; KANAWATI; ROUVEIROL, 2010; SA; PRUDENCIO, 2011), utiliza o conjunto de treinamento para treinar um modelo para a tarefa de classificar exemplos de entrada. De posse do modelo treinado, os exemplos do conjunto de testes são utilizados para avaliar (em termos de acertos e

erros) as predições realizadas. Várias métricas podem ser utilizadas para essa avaliação, como, por exemplo, a Precisão, o *Recall*, *F-measure* ou curvas de limiares, como a curva ROC e a curva PR, e suas respectivas áreas abaixo da curva (*Area Under the Curve* (AUC)) (YANG; LICHTENWALTER; CHAWLA, 2015).

3.3.2 Método de avaliação não supervisionada

A principal diferença entre a avaliação não supervisionada e a supervisionada é que a Predição de Links não é considerada um problema de classificação binária. Entretanto, assim como ocorre para a avaliação supervisionada, na avaliação não supervisionada, a rede é dividida em dois conjuntos distintos, o de treinamento e o de testes, porém apenas o de testes é usado. Na avaliação não supervisionada, não há uma fase de aprendizagem ou aquisição de conhecimento (LU; ZHOU, 2010). A seguir, *scores* são calculados, baseados em uma métrica (como qualquer uma das apresentadas neste capítulo), para todos os exemplos de entrada (geralmente exemplos que não possuem *links* entre os vértices do par) do conjunto de testes. Esses *scores* são então ordenados decrescentemente (SA; PRUDENCIO, 2011). Como não existe qualquer tipo de aprendizado ou classificação dos exemplos, na avaliação não supervisionada, a Predição de Links não é considerada um problema de classificação.

Existem dois métodos utilizados para definir quais relacionamentos serão sugeridos pelo preditor (WANG et al., 2015): o primeiro é chamado de *Top K*, em que os K pares de vértices associados aos K maiores *scores* são sugeridos como futuros relacionamentos, e no segundo método é definido um limiar θ, onde todo par de vértices (i, j) com um *score* superior a θ é sugerido como um futuro relacionamento.

Por fim, assim como também é feito para a avaliação supervisionada, as predições podem ser avaliadas por diversas métricas, como a Precisão, o *Recall* e as outras já citadas.

3.4 Outros problemas relacionados

A Predição de Links é apenas um dos problemas associados à ML e relacionado à ARS. As estratégias apresentadas até agora também são utilizadas em outros problemas, como por exemplo, o *PageRank* é utilizado no *ranking* de vértices de um grafo (mais usado para páginas de *internet*) e a abordagem de Bigramas escondidos, utilizada em processamento de linguagens naturais. Outros subproblemas da ML estão resumidos a seguir:

Ranking de vértices

Consiste em ordenar os vértices de uma rede baseado nas conexões que possuem ou na análise de centralidade dos vértices, levando em conta características como grau, proximidade e intermediação, que podem indicar a relevância de um vértice em um contexto (FREEMAN, 1979). Este problema costuma ser abordado em tarefas de Recuperação de Informação, onde

uma base de documentos deve ser analisada para decidir os mais relevantes de acordo com uma determinada consulta. Algoritmos como o PageRank (PAGE et al., 1998) e HITS (KLEINBERG, 1999) são bastante conhecidos, utilizados e referenciados na literatura, possuindo variações, inclusive para a Predição de Links: o *Rooted PageRank* mostrado neste capítulo.

Classificação de vértices

Como o próprio nome sugere, o problema trata de identificar e rotular indivíduos de uma rede de acordo com um conjunto finito de classes, em que os dados são tipicamente independentes e identicamente distribuídos (TANG; AGGARWAL; LIU, 2016). Por outro lado, os vértices da rede formam uma estrutura heterogênea, onde os dados relacionados são interdependentes. Rotular vértices depende então da exploração das relações entre os objetos da rede e como um vértices é afetado por outros ao redor.

Classificação de grafos

Visa classificar grafos inteiros com relação à determinadas propriedades observadas na rede. É um problema de aprendizagem de máquina e não requer uma análise coletiva como ocorre na classificação de vértices, já que a formação de um grafo geralmente é independente. BENCHERIF et al. (2015) tentam classificar imagens em alta resolução utilizando uma estratégia híbrida, onde a aprendizagem é feita por uma *Extreme Learning Machine* e os *pixels* das imagens são representados em uma estrutura de grafo.

Análise de subgrafos

Problema que compreende encontrar padrões (subgrafos) relevantes ou recorrentes dentro de um conjunto de grafos. Esses padrões são então aplicados em atributos para predição e na classificação de grafos. BEUTEL; AKOGLU; FALOUTSOS (2015) aplicam a detecção de subgrafos em redes para o problema de detecção de fraudes, ou seja, identificar indivíduos com comportamento anômalo e modelar essas anomalias encontradas dentro da rede.

Detecção de grupos

Compreende a descoberta de comunidades, ou *clusters*, numa rede. É bastante importante na análise de redes pois identificar *clusters* permite isolar tipos de dados e descobrir como eles estão interligados, além de como o tráfego de informações ocorre dentro da rede, diminuindo a complexidade e dimensão do problema. Esta representação resumida pode facilitar a visualização da rede e elucidar características importantes, o que normalmente não é possível numa análise da rede como um todo (BACKSTROM et al., 2006).

3.5 Aplicações práticas

As aplicações do problema de Predição de Links no mundo real variam de acordo com o domínio do problema e descobrir relacionamentos antes que eles aconteçam pode ser muito importante para determinadas instituições ou empresas, atraindo a atenção de vários pesquisadores (COOKE, 2006). Alguns domínios de aplicação estão descritos a seguir.

- **Sistemas de recomendação**: bastante conhecidos na prática, sistemas de recomendação indicam sugestões baseadas nas características e preferências de um usuário e na análise de preferências de usuários parecidos. Sistemas conhecidos, como o Netflix[2], a Amazon[3] e Foursquare[4] sugerem uma variedade de objetos, como filmes e seriados, produtos e locais para visitar (restaurantes, pontos turísticos, etc), respectivamente. Além destes, existem muitos outros sistemas. A Predição de Links pode ser usada para resolver este problema modelando o grafo da rede como um grafo bipartido, existindo dois tipos de vértices: o primeiro tipo representa o usuário e o segundo, preferências em geral. Os relacionamentos preditos entre usuários e preferências são, então, retornados como sugestões da aplicação ao usuário (CHEN; LI; HUANG, 2005; WANG et al., 2015). Além da recomendação de itens, outro emprego conhecido de sistemas de recomendação é o de sugestão de amizades em redes sociais, como o Facebook. A Predição de Links é feita normalmente com base em um usuário, e as sugestões de amizade são os pares com maiores *scores* em que o usuário analisado aparece.

- **Redes de contágio de doenças**: o sucesso do controle de epidemias está fortemente relacionado a compreensão de como as doenças se propagam. Em uma rede desse tipo, os vértices são indivíduos, contaminados ou não, e as arestas indicam a transmissão da doença de um indivíduo a outro da rede. A análise do histórico da rede junto com a Predição de Links pode ajudar a revelar padrões de transmissão e os caminhos mais prováveis por onde a doença pode transitar no futuro. Dessa forma, consegue-se antecipar ações de controle e impedir a infecção de novos indivíduos, sejam humanos ou outros seres vivos (ZANETTE, 2002; KEELING; EAMES, 2005; CRAFT, 2015).

- **Redes criminosas**: redes criminosas, como por exemplo, quadrilhas ou redes terroristas, são caracterizados pela confidencialidade da estrutura, ou seja, os relacionamentos entre os integrantes dessas redes estão ocultos propositadamente. A quebra de sigilos permite às forças de segurança a descoberta de apenas uma parte da estrutura. A Predição de Links pode ser utilizada neste contexto para descobrir as conexões omissas entre criminosos, contribuindo assim para o desmantelamento dessas redes e a resolução de crimes (KREBS, 2002; ZHENG; SKILLICORN, 2015).

[2] www.netflix.com
[3] www.amazon.com
[4] www.foursquare.com

- **Redes de coautoria**: talvez o tipo de rede mais abordado em trabalhos sobre Predição de Links, a grande vantagem dessas redes é que seus dados estão disponíveis publicamente, como já citado. Consiste em identificar pares de autores que ainda não publicaram um trabalho juntos, mas que tem grande probabilidade de fazê-lo em um futuro próximo. Redes deste tipo têm sido objeto de experimentos em muitos trabalhos acadêmicos, como base para testar os desempenhos das abordagens propostas.

3.6 Tendências futuras

Os principais problemas relacionados às estratégias de Predição de Links citadas ao longo do capítulo são a análise do tempo como um fator da predição e a alta complexidade de tempo e memória. Alguns outros pontos, como digrafos, arestas com pesos e estruturas multidimensionais também não foram suficientemente explorados (Lü; ZHOU, 2011).

Analisar grafos direcionados pode ser mais complexo, pois existem mais casos a tratar (JUSZCZYSZYN; MUSIAL; BUDKA, 2011), referentes às direções das arestas, tornando até a aplicação de métricas tradicionais, como distância entre vértices, mais problemática.

Com relação as arestas valoradas, determinar a importância de relacionamentos dentro da rede pode melhorar o resultado da predição (MURATA; MORIYASU, 2008; LU; ZHOU, 2010). Apesar de pesos agregarem mais informação à representação da rede, alguns autores, como MURATA; MORIYASU (2008), chegaram a conclusão de que arestas com pesos maiores são mais relevantes enquanto GRANOVETTER (1973) e LU; ZHOU (2010) sugerem que conexões fracas são mais significativas. Além disso, não é simples definir valores para os relacionamentos de uma rede, como, por exemplo, a força de um relacionamento entre duas pessoas.

Muitas redes do mundo real são constituídas por *links* com diferentes conotações, por exemplo, uma relação entre duas pessoas pode ser de amizade ou inimizade. Isso faz com que essas redes sejam estruturas multidimensionais, ou seja, diferentes atributos podem dar origem a diferentes tipos de arestas. Algumas soluções incluem modelos de *random walks* supervisionados para predição de *links* considerando diferentes atributos (BACKSTROM; LESKOVEC, 2011), predição de sinais (relacionamentos positivos ou negativos) de um relacionamento (KUNEGIS; LOMMATZSCH; BAUCKHAGE, 2009; LESKOVEC; HUTTENLOCHER; KLEINBERG, 2010), entre outros.

Além desses pontos, existem problemas bem comuns de aprendizagem de máquina presentes na Predição de Links. O custo computacional de algoritmos desse tipo aliado ao desbalanceamento de classes (existem muito mais exemplos de determinadas classes do que de outras) foi responsável pela aplicação de *low-rank approximations*, por exemplo, no processamento da rede, antes da computação das pontuações, visto que redes do mundo real são esparsas (muito mais exemplos de relacionamentos ausentes do que presentes). Outras técnicas para tratar o desbalanceamento de classes incluem *undersampling* (reduzir o número de instâncias das classes

predominantes) (LICHTENWALTER; LUSSIER; CHAWLA, 2010), *cost-sensitive learning* (PAN et al., 2008) e *chance-constrained* (DOPPA et al., 2010), aplicadas para aumentar as taxas de acerto dos algoritmos.

Por fim, a tendência é que trabalhos relacionados tentem utilizar algoritmos híbridos, de modo a aproveitar as vantagens de cada um, combinando diversos tipos de predição. Alguns trabalhos vistos neste capítulo, como, por exemplo, a TTM e os desenvolvidos por ZHENG; SKILLICORN (2015) e BENCHERIF et al. (2015), explicam que utilizar algoritmos híbridos ajuda a criar um modelo de predição mais complexo e satisfatório. A inclusão de atributos em vértices e outros tipos de informações externas pode enriquecer os resultados das predições. Lü; ZHOU (2011) acreditam que o sucesso comercial de sistemas de recomendação aumente ainda mais o interesse de acadêmicos e empresas por melhores soluções para tarefa de Predição de Links.

3.7 Considerações finais

Foram apresentados, no decorrer deste capítulo, vários aspectos sobre o problema de Predição de Links em redes sociais. À primeira vista, parece ser um problema estático, de predição de relacionamentos faltantes, mas foi visto que redes no mundo real são estruturas evolutivas, e que inserir o conceito de tempo dentro dos modelos de predição ajuda a melhorar a qualidade das predições. A segunda vertente do problema de Predição dos Links consiste em analisar a rede como uma estrutura evolutiva, detalhando as transições que representam as mudanças nas arestas presentes na rede, ao longo do tempo.

Também foram apresentadas diversas abordagens para o problema de Predição de Links, tanto do ponto de vista de conexões faltantes como o de análise da evolução da rede e temporalidade. As estratégias foram classificadas, baseado nas características principais de cada modelo utilizado, em estruturais, de similaridade, probabilísticas e temporais. Entretanto, não utilizam um modelo puramente, mas sim versões híbridas que unem aspectos de mais de um modelo. Algumas vantagens e desvantagens de cada uma foram sumarizadas e foi visto que, mesmo diante de algumas falhas, esses métodos funcionam e podem gerar modelos de predição aceitáveis.

É importante enfatizar que a estratégia proposta por este trabalho é baseada inicialmente nos princípios de transições de tríades detalhados por JUSZCZYSZYN; MUSIAL; BUDKA (2011). Entretanto, aplica outras técnicas um pouco mais complexas, entre elas a análise estatística de séries temporais (HUANG; LIN, 2009) e os conceitos de tensores (DUNLAVY; KOLDA; ACAR, 2011), vistas ao longo deste capítulo, para criar um modelo temporal capaz de servir como solução satisfatória para o problema de Predição de Links. A escolha de tais técnicas foi feita para tentar solucionar as principais deficiências encontradas nos trabalhos apresentados nesse capítulo, como, por exemplo, o uso de métricas estáticas como base para uma predição temporal ou o uso de funções mais simples de agregação de *scores*, o que ocasiona a perda de

informações temporais relevantes.

Por fim, este capítulo citou exemplos de aplicações de Predição de Links no mundo real, sendo os sistemas de recomendação um exemplo conhecido e de sucesso comercial. Além disso, foram apresentados alguns desafios enfrentados ou pouco explorados e um pouco sobre as tendências das abordagens que podem surgir no futuro.

No próximo capítulo será detalhada a estratégia proposta por este trabalho, mostrando o passo a passo da elaboração e implementação e o algoritmo final resultante, justificando as escolhas feitas, além de alguns detalhes técnicos e um comparativo com as abordagens já apresentadas.

4

Predição Temporal de Links baseada na evolução de subgrafos

A definição de LIBEN-NOWELL; KLEINBERG (2003) serve de base para a maioria das estratégias de predição. Contudo, existem aspectos que ainda são pouco explorados e que podem servir para oferecer uma solução mais precisa. Descobrir padrões de evolução de *links* em uma rede é importante para criar um bom modelo de predição. Um exemplo simples é sobre redes de coautoria: novos relacionamentos vão surgir em épocas de publicação de artigos em conferências, entre outras reuniões de trabalhos científicos. Fora desses períodos, não há muita atividade, o que torna uma parte dos dados bem menos relevante para o estudo. Como citado anteriormente, atividades de surgimento de relacionamentos podem estar atreladas a eventos do mundo real, onde pode se considerar que a rede está ativa, seguidas por longos períodos de inatividade. Descobrir eventos que funcionam como gatilhos para a criação de novas conexões pode ser a chave para criar um modelo preciso e satisfatório.

A estratégia desenvolvida neste trabalho é híbrida, do ponto de vista das abordagens vistas no capítulo anterior. Essa estratégia analisa, estatisticamente, como se dá a evolução das tríades, já mencionadas, ao longo do tempo. Pode-se dizer, então, que a estratégia proposta é uma abordagem temporal que faz uma análise estatística sobre a evolução de padrões estruturais relacionados às tríades. Uma modelagem temporal de dados foi criada para identificar e guardar todas as mudanças relevantes que ocorrem na rede, do ponto de vista estrutural, e um método de predição estatística para séries temporais foi escolhido para ser aplicado sobre o conjunto de informações armazenados no modelo.

No Capítulo 3 foram vistos diversos tipos de abordagens para a Predição de Links, onde cada uma explora um determinado aspecto (como padrões estruturais) da rede. Em especial, foi visto o trabalho de JUSZCZYSZYN; MUSIAL; BUDKA (2011), que discute aspectos sobre as tríades, a estrutura base que será observada por este trabalho. JUSZCZYSZYN; MUSIAL; BUDKA (2011) identificam as mudanças por que passam as tríades (transições entre tipos distintos) e criam uma TTM para armazenar as probabilidades dessas mudanças ocorrerem. Vale lembrar que uma TTM não é indexada pelos vértices do grafo que representa a rede, mas sim pelos tipos possíveis de tríades, que, como visto no Capítulo 3, dependem do tipo das arestas do grafo (se as arestas são direcionadas ou não). Cada índice (a, b) da TTM indica a probabilidade de que uma tríade do tipo a, em um determinado tempo t, passe a ser do tipo b no tempo $t + 1$.

Essa mudança de tipo é chamada pelos autores de transição e a TTM armazena as probabilidades de que essas transições ocorram.

Porém, algumas limitações foram identificadas: a primeira diz respeito à TTM, pois pode apenas armazenar informações sobre as transições entre dois tempos t e $t + 1$ apenas, o que acaba não sendo tão diferente da primeira perspectiva do problema de Predição de Links, já que a TTM ou contém informações sobre uma observação apenas (a observação, no caso, diz respeito às transições que ocorreram entre dois intervalos de tempo) ou contém informações dos dados condensados de observações do tempo t_0 ao tempo T. De toda forma, as informações da rede sempre estão reunidas e representadas por uma única estrutura de dados (no caso, a TTM). A segunda limitação refere-se à predição temporal sugerida por JUSZCZYSZYN; MUSIAL; BUDKA (2011), onde os autores sugerem o cálculo das TTMs referentes às diversas observações (entre t_0 e t_1, t_1 e t_2, e assim por diante), mas o valor final é a média aritmética dos valores intermediários encontrados, o que não descreve a evolução dos valores intermediários encontrados nem identifica nenhum padrão de evolução que possa existir dentro do período de tempo observado. Estas limitações serão exploradas como pontos de melhoria neste trabalho.

4.1 Algoritmo

O algoritmo desenvolvido está dividido em diversas tarefas, cada uma responsável por uma parte específica da predição. Um pseudocódigo completo do algoritmo é apresentado pelo Algoritmo 2, Página 69, e cada tarefa encontra-se comentada e numerada.

Seja T o conjunto de períodos de observação da rede e t uma variável que indexa cada período de T. Seja $G(V,A)$ um grafo G com um conjunto V de vértices e um conjunto A de arestas, chamaremos $G_t(V, A_t)$ o grafo G observado no tempo t. O conjunto V de vértices não muda, pois será assumido que o grafo não perde ou ganha vértices, de modo a focar apenas nas mudanças das arestas. Uma aresta (ou par de vértices) também será identificada pela tripla (i, j, t), ou seja, o par (i, j) observado no tempo t, já que pode existir mais de um estado para (i, j), além de situar temporalmente o par. Seja também *lista_arestas* uma lista que enumera todas as arestas (i, j, t) observadas, e Δ o número de períodos à frente para a predição.

A seguir, um resumo sobre cada tarefa numerada no Algoritmo 2, Página 69, é apresentado. Os detalhes sobre cada tarefa (bem como as variáveis presentes no pseudocódigo) serão esclarecidos posteriormente. As tarefas 1, 2, 3 e 4 foram desenvolvidas utilizando a linguagem de programação Python[1], enquanto a tarefa 5 foi implementada utilizando a linguagem estatística R[2].

1. **Modelagem temporal dos dados**: primeiramente, será criado um tensor Z de três dimensões para armazenar as matrizes de adjacência para cada período de observação da rede. Este tensor será indexado pelo tempo (de t_0 a T) e pelos vértices do grafo

[1] www.python.org
[2] cran.r-project.org

Algorithm 2 Computar_Scores_TTT()

Entrada: $G(V,A)$; (i,j); Δ; *tipo_modelo* ▷ $i,j \in V$. (i,j) é um parâmetro opcional.
Saída: *score_intermediario(i,j,T + Δ - 1)*

Inicializar Z (dimensões $|V| * |V| * T$) com 0's ▷ 1. Modelagem temporal dos dados
for all $(i,j,t) \in lista\_arestas$ **do**
    $Z(i,j,t) \leftarrow 1$

$total\_tipos\_triades \leftarrow 0$ ▷ 2. Indexação
if $G(V,A)$ é direcionado **then**
    $total\_tipos\_triades \leftarrow 64$
else
    $total\_tipos\_triades \leftarrow 8$
$v \leftarrow \{\}$ ▷ Inicializar vetor de índices
for $i \leftarrow 0$ **to** $total\_tipos\_triades - 1$ **do**
    Adicionar i ao vetor v ▷ Adicionar tipos de tríades ao vetor de índices

for $t \leftarrow t_0$ **to** T **do** ▷ 3. Cálculo das probabilidades de transições
    $TTT(.,.,t) \leftarrow calcular\_probabilidades\_transicoes(Z,v,t)$

$pares\_vertices \leftarrow \{\}$ ▷ Pares de vértices cujos *scores* serão calculados
if (i,j) for fornecido como entrada **then**
    $pares\_vertices \leftarrow \{(i,j)\}$
else ▷ Caso contrário, o cálculo é feito para toda a rede
    for all $i \in V$ **do**
        for all $j \in V$ **do**
            if $i \neq j$ **then**
                Adicionar (i,j) ao vetor $pares\_vertices$

for all $(i,j) \in pares\_vertices$ **do** ▷ 4. Cálculo de scores intermediários
    for $t \leftarrow t_0$ **to** T **do**
        Construir o conjunto *triades_nao_vazias(i, j, t)*
        $score\_intermediario(i,j,t) \leftarrow 0$
        for all $e \in$ *triades_nao_vazias(i, j, t)* **do**
            Construir o conjunto *transicoes(i, j, t, e)*
            $TD(e) \leftarrow 0$
            for all $tr(a,b,t+1) \in transicoes(i,j,t,e)$ **do**
                $TD(e) \leftarrow TD(e) + TTT(a,b,t+1)$
            $score\_intermediario(i,j,t) \leftarrow score\_intermediario(i,j,t) + TD(e)$
        Adicionar *score_intermediario(i, j, t)* ao vetor $scoresTTT(i,j)$
    Persistir $scoresTTT(i,j)$

for all $scoresTTT(i,j)$ **do** ▷ 5. Predição temporal estatística
    $modelo\_serie\_temporal \leftarrow criar\_modelo(scoresTTT(i,j), tipo\_modelo)$
    $score\_intermediario(i,j,T+\Delta-1) \leftarrow predicao\_modelo(modelo\_serie\_temporal, \Delta)$
    Persistir $score\_intermediario(i,j,T+\Delta-1)$

que representa a rede observada. Basicamente, deseja-se armazenar uma sequência de matrizes de adjacência relativas à rede em análise;

2. **Indexação**: tarefa que identifica o tipo das arestas do grafo (direcionadas ou não direcionadas) para definir as dimensões do Tensor de Transições de Tríades (TTT) ($8 * 8 * (T$ - $1)$ ou $64 * 64 * (T$ - $1)$) que será criado durante a tarefa seguinte e atribuir um índice para cada tipo de tríade diferente;

3. **Cálculo das probabilidades de transições**: nesta etapa, para resolver a primeira limitação citada anteriormente, o conceito de TTM é extendido utilizando-se outro tensor de três dimensões, chamado aqui de TTT. Como os dados da rede estão divididos em períodos, que iniciam na observação da rede no tempo t_0 e terminam na observação feita no tempo T, uma dimensão do TTT será indexada pelo tempo. As outras duas dimensões serão indexadas pelos índices que foram atribuídos a cada tipo de tríade e cada índice TTT$(a,b,t+1)$ guardará a probabilidade de uma tríade do tipo a no tempo t passar a ser do tipo b no tempo t + 1, assim como ocorre em uma TTM. Em resumo, a estrutura de dados utilizada para armazenar as probabilidades de transições a serem calculadas será um tensor TTT(a,b,t), em que a e b são relativos aos índices atribuídos pela tarefa anterior e t ($t_0 < t \leq T$) é indexado por cada período de tempo. Depois de instanciar o TTT, para cada t, onde $t_0 \leq t \leq T$, será aplicado o algoritmo *Triad Census* (ver Algoritmo 1, Página 26) para enumerar todas as tríades não vazias presentes. Após essa enumeração, para cada tempo $t_0 < t \leq T$, será feita uma análise dois a dois entre dois tempos consecutivos (t_1 será analisado em relação a t_0, t_2 em relação a t_1 e assim por diante), similar à contagem feita na Seção 2.2.3, Página 27, e exemplificada pela Figura 2.5, Página 28, de modo a calcular todas as probabilidades de transições. Essas probabilidades serão armazenadas em seus respectivos índices dentro do TTT (as transições entre t_0 e t_1 serão armazenadas em TTT(.,.,t_1), as transições entre t_1 e t_2 em TTT(.,.,t_2), até as transições entre T - 1 e T, que serão armazenadas em TTT(.,.,T));

4. **Cálculo de *scores* intermediários**: além do TTT calculado, esta tarefa poderá receber também como entrada um par de vértices, caso seja necessário realizar os cálculos para aquele par apenas. Caso contrário, o cálculo será feito para a rede inteira. Nesta etapa, serão calculados os *scores* intermediários referentes a cada índice TTT$(.,.,t)$ do tensor de transições. Para cada par de vértices (i,j) do grafo o qual deseja-se realizar a predição, e para cada $t_0 < t \leq T$, serão enumeradas todas as tríades não vazias que contém i e j como vértices. Para cada elemento deste conjunto, será determinado o tipo da tríade e serão analisadas todas as transições similares (transições de tríades que não necessariamente contém i e j), onde a tríade resultante da transição possui uma aresta entre as posições de i e j. Cada *score* intermediário será então dado pelo somatório de todas as probabilidades de aquelas transições

similares ocorrerem. É importante ressaltar que os *scores* intermediários não mudam, pois são relativos às observações anteriores da rede, e só precisam ser calculados apenas uma vez. Visto isso, os dados serão armazenados para posterior acesso direto, sem a necessidade de recálculo;

5. **Predição temporal estatística**: recebendo como entrada um par (i, j) de vértices para o qual deseja-se realizar a predição de *links*, esta tarefa consiste em criar uma série temporal baseada nos *scores* intermediários relativos ao par (i, j) de entrada. A partir daí, um método de predição estatística (no caso deste trabalho, a predição será feita utilizando três tipos diferentes de alisamento exponencial) é aplicado sobre esta série temporal, de modo a construir um modelo estatístico que descreva como ocorreu a evolução dos *scores* intermediários dentro da série temporal. Com este modelo em mãos, realiza-se uma predição para descobrir o *score* que seria atribuído ao par (i, j) em um tempo $T + \Delta$ posterior aos dados de entrada. Este valor pode então ser utilizado de diversas formas, como indicar os n pares de vértices com maiores *scores* preditos como futuros relacionamentos ou servir de entrada para algum método de aprendizagem de máquina.

4.2 Modelagem temporal dos dados

Entrada: Conjunto V de vértices do grafo; $lista\_arestas$
Saída: Tensor Z que representa $G_t(V,A_t)$, para todo $t_0 \leq t \leq T$

Para construir uma representação temporal que visa não perder informação, isto é, que tente minimizar o problema da detecção de curtos períodos de atividade dentro de longos períodos de inatividade, os dados do histórico de observações da rede serão primeiramente divididos em T períodos. Definir em quantos períodos os dados serão divididos não é trivial, pois como foi explicado no Capítulo 3, intervalos de tempo muito grandes fazem com que curtos períodos em que a rede está ativa (surgimento ou desaparecimento de relacionamentos) passem despercebidos. Em contrapartida, intervalos pequenos fazem com que qualquer mudança seja considerada como significativa.

Inicialmente é definido um tensor Z, de tamanho $n * n * T$, onde $n = |V|$, que armazena os dados de entrada, da seguinte maneira:

$$Z(i,\ j,\ t) = \begin{cases} 1,\ se\ (i,\ j,\ t)\ \in\ lista\_arestas; \\ 0,\ caso\ contrário. \end{cases} \tag{4.1}$$

As linhas 4 a 6 do Algoritmo 2, Página 69, mostram a construção do Tensor Z. Z, definido acima, é então retornado como saída desta tarefa.

4.3 Indexação

Entrada: Tensor Z
Saída: Vetor v de índices para os tipos de tríades

Quando as arestas de um grafo não são direcionadas, existem 64 ($8 * 8$) possíveis transições de tríades, e quando as arestas são direcionadas, existem 4096 ($64 * 64$) possibilidades dessas transições. As tríades vazias não serão contabilizadas inicialmente, porém serão contabilizadas caso apareçam como resultado de uma transição, por exemplo, uma tríade do tipo 2 (que possui apenas uma aresta) passar a ser uma tríade vazia.

Primeiramente deve-se enumerar todos os tipos de tríades (ver linhas 8 a 15 do Algoritmo 2, Página 69), pois o TTT que será criado não é indexado pelos vértices do grafos, mas sim pelos tipos de tríades e pelo tempo. Assim, determina-se o tipo da aresta do grafo em análise e em seguida, determina-se a numeração dos tipos de tríades como mostrada na Figura 4.1:

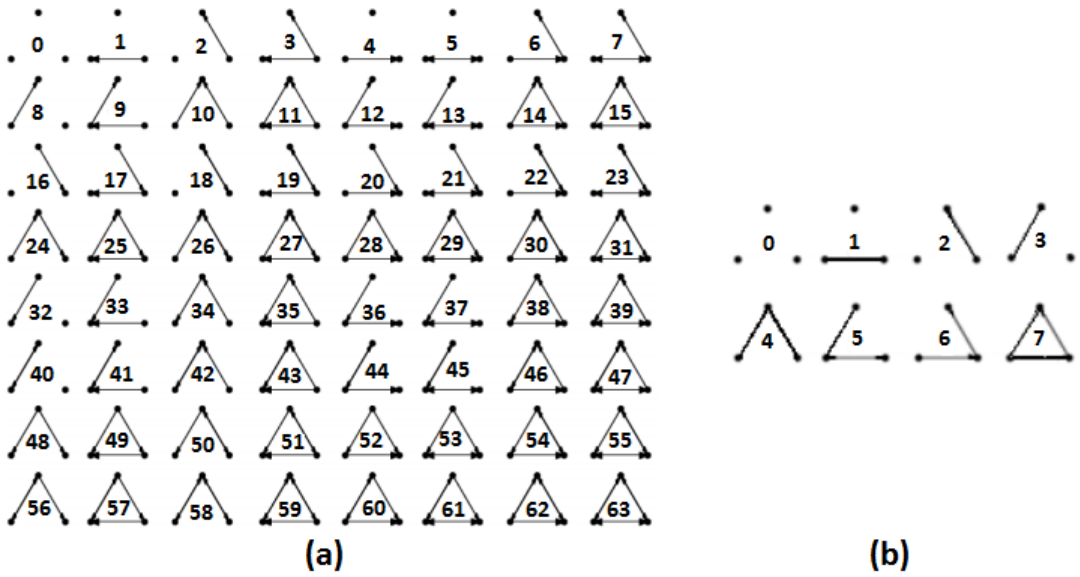

Figura 4.1: (a) Índices dos tipos de tríades de um grafo direcionado e (b) índices dos tipos de tríade para um grafo não direcionado.

A partir desta numeração, cria-se um vetor v de tamanho 8 (caso as arestas não sejam direcionadas) ou tamanho 64 (caso sejam direcionadas), em que cada índice do vetor guarda as posições das arestas (ver Figura 4.1) correspondentes ao seu respectivo tipo (o índice 0 guarda as posições de arestas de uma tríade vazia, o índice 1 guarda as posições das arestas de uma tríade do tipo 1, o índice 2 guarda as posições do tipo 2 e assim por diante). De posse dessa indexação, a próxima tarefa será responsável por criar o Tensor de Transições de Tríades a partir de v, identificar as transições que ocorrem e calcular as probabilidades dessas ocorrências.

4.4 Cálculo das probabilidades de transições

Entrada: Tensor Z; vetor v
Saída: TTT com as probabilidades de transições

Esta etapa do algoritmo é responsável por criar o TTT (linhas 17 e 18 do Algoritmo 2, Página 69), que armazena as probabilidades de transições que ocorrem ao longo dos períodos de observação, e calcular as probabilidades de transições de tríades. Para poder definir o TTT formalmente, primeiro deve-se definir uma transição. Dada uma tríade qualquer formada pelos vértices i, j e k, uma transição entre os tempos t - 1 e t é definida como $tr(a,b,t)$, onde a é o tipo da tríade formada por i, j e k no tempo t - 1 e b o tipo da mesma tríade observada no tempo t.

A forma de calcular a probabilidade de uma transição ocorrer é similar ao método mostrado na Seção 2.2.3, Página 27. Primeiramente, seja $\Delta_{a,t}$ o número de tríades do tipo a observadas em um tempo t qualquer. Pode-se definir a probabilidade $P(tr(a,\ b,\ t))$ de ocorrência da transição de uma tríade do tipo a, no tempo t - 1, para o tipo b no tempo t como se segue:

$$P(tr(a,\ b,\ t)) = \frac{|tr(a,\ b,\ t)|}{\Delta_{a,t-1}}, \tag{4.2}$$

onde $|tr(a,\ b,\ t)|$ é o número de transições de tríades do tipo a no tempo t - 1 para o tipo b no tempo t observadas. Para determinar $\Delta_{a,t}$, é aplicado o algoritmo *Triad Census* para cada fatia de tempo t, onde $t_0 \leq t \leq T$, retornando os valores de $\Delta_{a,t}$, para todo $a \in v$, determinando assim os denominadores das frações que representam as probabilidades (ver Equação 4.2). Essa definição é uma probabilidade baseada em uma contagem. Dado que foram observadas $\Delta_{a,t-1}$ tríades do tipo a no tempo t - 1, verifica-se quantas delas passaram a ser do tipo b no tempo t ($|tr(a,\ b,\ t)|$) e diz-se que a probabilidade é o segundo fator dividido pelo primeiro.

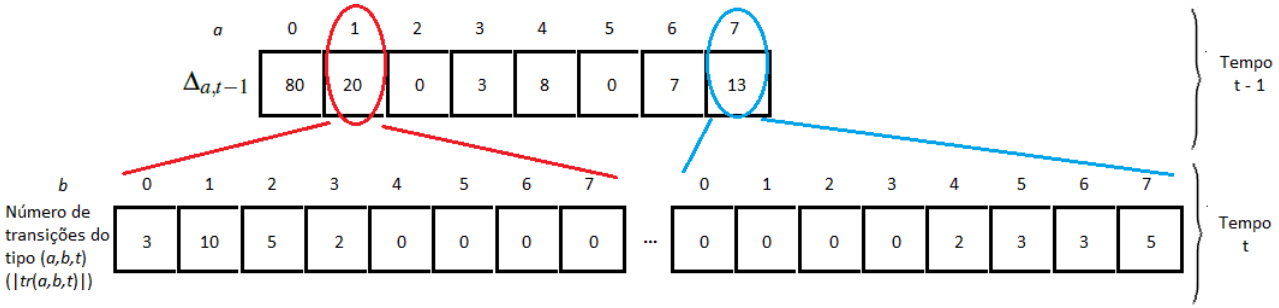

Figura 4.2: Exemplo do cálculo do número de transições do tipo (a,b,t), isto é, $|tr(a,b,t)|$, para um grafo com arestas não direcionadas. Em vermelho, está marcado o tipo de tríade por onde se inicia o prodedimento de contagem e em azul, o último tipo em que a contagem é feita.

A Figura 4.2 mostra um exemplo de como são determinados $|tr(a,\ b,\ t)|$ e $P(tr(a,\ b,\ t))$. Em vermelho, verifica-se (utilizando o algoritmo *Triad Census*) que no tempo t - 1 existiam 20 ($\Delta_{1,t-1} = 20$) tríades do tipo 1 e que, no tempo t, 3 dessas 20 tríades passaram a ser do tipo 0, logo, $|tr(1,\ 0,\ t)| = 3$ e $P(tr(1,\ 0,\ t)) = 3/20$, pela Equação 4.2. A mesma lógica é então aplicada

para calcular $P(tr(1,\ 1,\ t))$: verifica-se que das 20 tríades do tipo 1 no tempo t - 1, 10 ainda são do tipo 1 no tempo t e portanto, $|tr(1,\ 1,\ t)| = 10$ e $P(tr(1,\ 1,\ t)) = 10/20$. Se for identificado que existe um número maior que zero de tríades de um determinado tipo no tempo t - 1, esse procedimento é aplicado, caso contrário, todas as probabilidades de transição são admitidas como 0, já que não existia nenhuma tríade daquele tipo (pela Figura 4.2, Página 73, por exemplo, todas as $P(tr(2,\ .,\ t))$ são iguais à 0, já que no tempo t - 1 não existiam tríades do tipo 2).

O procedimento de contagem é então feito para todos os tipos das tríades observadas em t - 1. Apesar da Figura 4.2, Página 73, mostrar uma contagem feita em um grafo com arestas não direcionadas, o raciocínio é análogo para um grafo de arestas direcionadas, e a diferença é que o procedimento é aplicado para cada um dos 64 tipos de tríades possíveis, ao invés dos 8 tipos mostrados no exemplo. É importante ressaltar que este procedimento não é aplicado para transições a partir de tríades do tipo 0, pois o índice 0 representa tríades vazias, que são desconsideradas neste algoritmo.

Sabendo-se calcular as probabilidades de transições, e dado Z o tensor (fornecido como entrada) de tamanho $n * n * T$ para esta tarefa e v o vetor de índices de tríades (também fornecido como entrada), um TTT terá dimensões $|v| * |v| * (T$ - 1). A partir da definição de probabilidade de transição, pode-se então definir o TTT:

$$\mathrm{TTT}(a,\ b,\ t) = P(tr(a,\ b,\ t)). \tag{4.3}$$

Como uma transição $tr(a,b,t)$ qualquer ocorre entre dois tempos consecutivos (no caso, entre t - 1 e t), tem-se que a dimensão do tempo do TTT terá tamanho T - 1, pois TTT(.,.,t_1) compreende as transições entre t_0 e t_1, TTT(.,.,t_2) compreende as transições entre t_1 e t_2 e assim por diante, até o último índice, TTT(.,.,T), que compreende as transições entre T - 1 e T. A Figura 4.3 mostra graficamente o esquema de armazenamento das probabilidades de transições dentro do TTT.

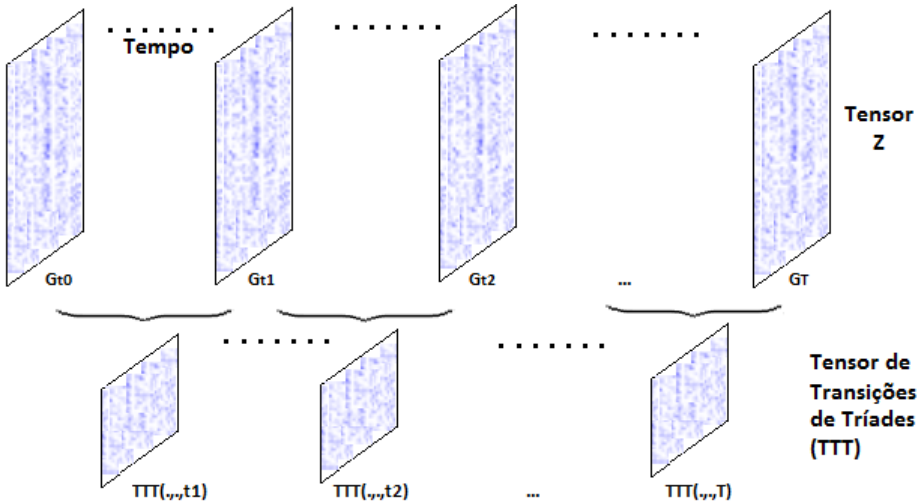

Figura 4.3: Esquema de criação de um TTT. As probabilidades de transições vistas entre os tempos t_0 e t_1 são armazenadas em TTT(.,.,t_1), entre t_1 e t_2 em TTT(.,.,t_2) e assim por diante.

Quando as arestas do grafo são direcionadas, existem 64^2 transições possíves, enquanto para arestas não direcionadas, este número cai para 8^2. Visto isso, entre dois tempos t - 1 e t consecutivos, serão calculadas $63 * 64$ probabilidades de transições ($P(tr(1,0,t))$, ..., $P(tr(1,63,t))$, $P(tr(2,0,t))$, ..., $P(tr(2,63,t))$, ..., $P(tr(63,63,t))$) para arestas direcionadas e $7 * 8$ probabilidades de transições ($P(tr(1,0,t))$, ..., $P(tr(1,7,t))$, $P(tr(2,0,t))$, ..., $P(tr(2,7,t))$, ..., $P(tr(7,7,t))$) para arestas não direcionadas (ver exemplo da Figura 4.4). Vale lembrar que as probabilidades de transições a partir de tríades vazias ($P(tr(0,.,t))$) não são calculadas, já que interessa apenas a análise de tríades não vazias. A Figura 4.4 mostra graficamente um exemplo do que guarda um índice TTT(.,.,t).

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	1/3	2/3	0	0	0	0
3	0	0	1/4	3/4	0	0	0	0
4	0	0	0	0	1/2	0	0	1/2
5	0	0	0	0	0	1/1	0	0
6	0	0	0	0	0	0	1/1	0
7	0	0	0	0	0	0	1/3	2/3

Figura 4.4: Exemplo de um índice TTT(.,.,t) de um TTT. Pelo exemplo, verifica-se que $P(tr(2,3,t))$ = 66,66%. É possível ver também que a primeira linha da matriz TTT(.,.,t) é sempre 0, pois as tríades vazias são desconsideradas, e que as somas das probabilidades em cada linha (exceto a primeira) é sempre 1.

O uso de um tensor (o TTT especificamente) para armazenar as probabilidades de transições é uma melhoria feita devido à primeira limitação, vista no início deste capítulo, na modelagem de dados utilizando TTMs que foi vista no Capítulo 3.

4.5 Cálculo de *scores* intermediários

Entrada: Tensor Z; Tensor TTT; par de vértices (i, j) (opcional)
Saída: Vetor $scoresTTT(i, j)$ de scores intermediários

Antes de detalhar como ocorre o cálculo dos *scores* intermediários, Algumas particularidades sobre esta tarefa são enumeradas a seguir:

1. O par (i, j) de vértices (onde $i \neq j$) de entrada é opcional;

2. Caso (i, j) não seja fornecido, o vetor *scoresTTT* será calculado para cada par de vértices (i, j) possível (desde que $i \neq j$), baseado nos vértices armazenados em Z;

3. Esta tarefa é a que demanda mais tempo para ser concluída. Já que as probabilidades de transições e os *scores* intermediários não mudam (será explicado mais à frente), os vetores *scoresTTT* de saída para cada par (i, j) são persistidos, de modo que a próxima tarefa possa acessar diretamente os dados e evitando que estes dados tenham de ser recalculados (o algoritmo pode começar diretamente a partir da próxima tarefa, isto é, a predição temporal estatística), pois o custo computacional de tempo é alto.

Para gerar os *scores* intermediários a partir de Z e TTT, deve-se tomar como base um par de vértices (i, j). Se (i, j) for fornecido como entrada, esta tarefa é executada apenas uma vez. Caso (i, j) não seja fornecido, os valores de i e j ($i \neq j$) serão iterados com base nos vértices armazenados em Z. A tarefa será descrita adiante assumindo que se tem um par (i, j) como referência.

Seja (i, j) um par de vértices, onde i, $j \in V$ (conjunto V de vértices do grafo $G(V,A)$ que representa a rede) e $i \neq j$, e t uma fatia de tempo qualquer, onde $t_0 \leq t \leq T$ - 1, o cálculo do *score* intermediário entre t e t + 1 ocorre da seguinte maneira: primeiramente deve-se definir o conjunto *triades_nao_vazias(i, j, t)* de tríades não vazias, isto é, que contém pelo menos uma aresta, que tem i e j como dois dos três vértices e que foram observadas no tempo t. Em resumo, *triades_nao_vazias(i, j, t)* enumera todas as tríades não vazias e que contém i e j como vértices, identificadas em $Z(., ., t)$.

O conjunto *triades_nao_vazias(i, j, t)* é então utilizado para criar outro conjunto, agora chamado de *transicoes(i, j, t, e)*. Para um elemento $e \in$ *triades_nao_vazias(i, j, t)*, este conjunto é definido pelas transições de e que resultam em uma aresta entre as posições de i e j, no tempo t + 1. Para isto, deve-se analisar a tríade e e ver quais tipos apresentariam uma aresta entre i e j. A Figura 4.5 mostra como fazer esta identificação:

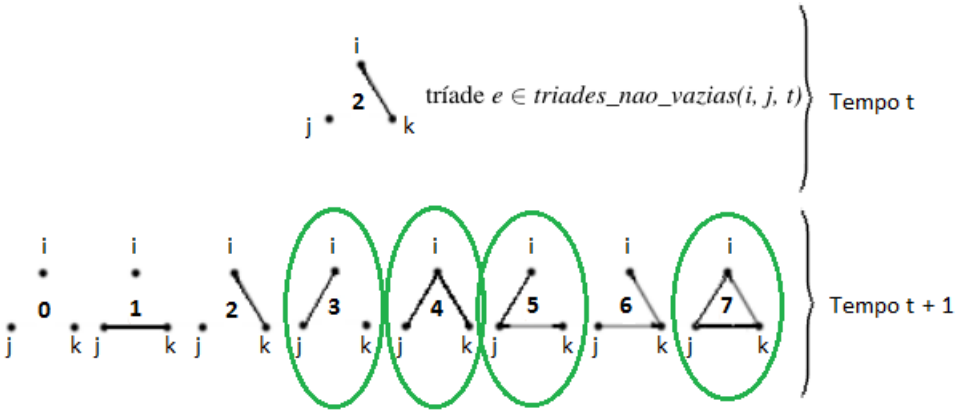

Figura 4.5: Exemplo de como identificar transições que resultam em uma aresta entre i e j (marcadas em verde).

Pela Figura 4.5, a tríade $e \in$ *triades_nao_vazias(i, j, t)* mostrada foi identificada no tempo t e contém i e j como vértices. No tempo t + 1, existem 8 transições possíveis para e, mas apenas

quatro das oito transições, sendo elas: $tr(2,3,t+1)$, $tr(2,4,t+1)$, $tr(2,5,t+1)$ e $tr(2,7,t+1)$ (marcadas em verde na Figura 4.5, Página 76) resultariam em e com uma configuração de arestas que tivesse uma aresta entre i e j. O conjunto *transicoes(i, j, t, e)* é então composto por estas transições. Pelo exemplo da Figura 4.5, Página 76, *transicoes(i, j, t, e)* = $\{tr(2,3,t+1)$, $tr(2,4,t+1)$, $tr(2,5,t+1)$, $tr(2,7,t+1)\}$.

Para toda tríade $e \in$ *triades_nao_vazias(i, j, t)*, define-se o seu respectivo conjunto *transicoes(i, j, t, e)*. O *score* intermediário é calculado da seguinte maneira: a partir de cada conjunto *transicoes(i, j, t, e)*, um valor *TD* que corresponde à soma das probabilidades de ocorrência de cada transição *tr(a, b, t + 1)* pertencente à *transicoes(i, j, t, e)* é calculado, isto é:

$$TD(e) = \sum_{tr(a,b,t+1) \in transicoes(i,j,t,e)} \text{TTT}(a,b,t+1). \tag{4.4}$$

Vale lembrar que as probabilidades de ocorrência das transições do conjunto *transicoes(i, j, t, e)* já foram calculadas e estão armazenadas no TTT construído pela tarefa anterior. Por fim, para determinar o *score* intermediário, para cada $e \in$ *triades_nao_vazias(i, j, t)*, somam-se os valores $TD(e)$ encontrados:

$$score\_intermediario(i,j,t) = \sum_{e \in triades\_nao\_vazias(i,j,t)} TD(e). \tag{4.5}$$

A fórmula direta para o cálculo do *score* intermediário é dada a seguir:

$$score\_intermediario(i,j,t) = \sum_{e \in triades\_nao\_vazias(i,j,t)} \sum_{tr(a,b,t+1) \in transicoes(i,j,t,e)} \text{TTT}(a,b,t+1). \tag{4.6}$$

A Equação 4.6 mostra como calcular o *score* intermediário para um tempo t qualquer, onde $t_0 \leq t \leq T$ - 1. O vetor *scoresTTT*(i, j) de saída é composto pelos *scores* intermediários para todos os valores de t, isto é:

$$scoresTTT(i,j) = \{score\_intermediario(i,j,t_0),\ ...,\ score\_intermediario(i,j,T-1)\}.$$

Como citado no início desta seção, se o par (i, j) for fornecido como entrada, esta tarefa se encerra após a construção de *scoresTTT*(i, j) para o par (i, j). Caso não seja fornecido, i e j serão iterados baseados nos vértices do tensor Z, e o vetor *scoresTTT*(i, j) será calculado para cada par. Outro ponto citado foi que os vetores *scoresTTT*(i, j) serão persistidos. Como os dados de Z não mudam, já que são informações sobre um período de tempo passado, não há a necessidade de se recalcular o TTT e os vetores *scoresTTT*(i, j) a cada chamada do algoritmo, já que essas informações também sempre serão as mesmas. Dado que o cálculo dos vetores *scoresTTT*(i, j) foi feito para todos os pares de vértices (i, j), os dados são persistidos para a próxima chamada deste algoritmo poder iniciar a partir da última tarefa, a predição temporal

estatística, acessando diretamente os dados calculados.

O algoritmo para o cálculo dos *scores* intermediários apresentado, para um par de vértices (i, j), pode ser descrito pelo Algoritmo 3 a seguir.

Algorithm 3 Scores_Intermediarios()

1: **Entrada:** Tensor Z; Tensor TTT; (i, j)
2: **Saída:** $scoresTTT(i, j)$
3: **for** $t \leftarrow t_0$ **to** T **do**
4: Construir o conjunto *triades_nao_vazias(i, j, t)*
5: $score\_intermediario(i, j, t) \leftarrow 0$
6: **for all** $e \in$ *triades_nao_vazias(i, j, t)* **do**
7: Construir o conjunto *transicoes(i, j, t, e)*
8: $TD(e) \leftarrow 0$
9: **for all** $tr(a, b, t+1) \in transicoes(i, j, t, e)$ **do**
10: $TD(e) \leftarrow TD(e) + \mathrm{TTT}(a, b, t+1)$
11: $score\_intermediario(i, j, t) \leftarrow score\_intermediario(i, j, t) + TD(e)$
12: Adicionar *score_intermediario(i, j, t)* ao vetor $scoresTTT(i, j)$
13: Persistir $scoresTTT(i, j)$

A lógica que compreende a escolha dos pares de vértices (i, j), bem como o Algoritmo 3, é descrita pelas linhas 20 a 40 do Algoritmo 2, Página 69.

4.6 Predição temporal estatística

Entrada: $scoresTTT(i, j)$; Δ; $tipo\_modelo$
Saída: $score\_intermediario(i, j, T + \Delta - 1)$

A tarefa de predição temporal estatística consiste em predizer o valor de *score_intermediario(i, j, T + Δ - 1)* baseado no vetor *scoresTTT(i, j)* de entrada. Como foi definido na Seção 4.4, Página 73, $scoresTTT(i, j) = \{score\_intermediario(i, j, t_0)$, ..., $score\_intermediario(i, j, T-1)\}$. Portanto, esta tarefa irá, basicamente, predizer o valor do *score* intermediário relativo a (i, j) no tempo T + Δ - 1, isto é, que valor *score_intermediario* assume Δ fatias de tempo à frente. Os modelos de predição que serão explicados adiante foram utilizados como melhoria para a segunda limitação da abordagem proposta por JUSZCZYSZYN; MUSIAL; BUDKA (2011), onde é sugerido que o *score* final seja calculado pela média aritmética dos *scores* intermediários calculados utilizando TTMs. Como foi dito no início do capítulo, a média aritmética não descreve satisfatoriamente a evolução dos valores intermediários encontrados nem identifica nenhum padrão que possa ser usado para melhorar os resultados da abordagem.

Antes de poder fazer a predição do *score_intermediario(i, j, T + Δ - 1)*, é necessário criar um modelo estatístico. O modelo criado por este trabalho é baseado em médias móveis ponderadas exponencialmente, mais conhecido como Alisamento Exponencial. Um modelo de alisamento exponencial é baseado em uma série temporal, já discutida no Capítulo 3. A

série temporal que servirá de base para o modelo estatístico é construída utilizando-se o vetor *scoresTTT(i, j)*. Mais precisamente, os valores do vetor, isto é, {*score_intermediario(i,j,*t_0*), ..., score_intermediario(i,j,T - 1)*}, serão considerados como uma série temporal. Observando-se como os *scores* intermediários são obtidos, percebe-se que faz sentido considerar *scoresTTT(i, j)* como uma série temporal.

A seguir, é calculado um modelo de alisamento exponencial baseado em *scoresTTT(i, j)*. Para o escopo deste trabalho, foram escolhidos três métodos de alisamento, sendo eles o Alisamento Exponencial Simples, o método linear de Holt (ou Alisamento Exponencial Duplo) (HOLT, 2004) e o método preditivo de Holt-Winters (ASSIS et al., 2013). Para cada *scoresTTT(i, j)*, é criado um modelo utilizando cada um desses métodos, a fim de analisar qual oferece a melhor predição.

4.6.1 Alisamento Exponencial

É um método que tenta isolar componentes da informação (série temporal) apresentada. Baseado nos valores observados (os valores do vetor *scoresTTT*, para o caso desta tarefa), este método utiliza médias móveis ponderadas exponencialmente de modo a corrigir as estimativas da média sazonal (chamada de nível) e ruído. Versões mais avançadas, como o método linear de Holt (HOLT, 2004), incluem a inclinação (mudanças de nível), e outros, ainda, a sazonalidade (ASSIS et al., 2013). É bastante usado nas áreas de administração, finanças, entre outros, por modelar informações, sazonais (ou seja, eventos que ocorrem em determinadas épocas) ou não, de maneira adequada. Dada a série temporal S_t com frequência f (os valores da série são observados de f em f medidas de tempo, como semanas, meses ou anos), a predição para Δ periodos à frente é calculada pela equação a seguir:

$$\widehat{S}_{t+\Delta|t} = a_t + \Delta * b_t + s_{t+\Delta-f}, \; com \; \Delta \leq f, \tag{4.7}$$

onde a_t é o nível estimado, b_t é a inclinação estimada, $a_t + k * b_t$ é o nível estimado no tempo $t + k$ e $s_{t+k-\Delta}$ é a estimativa do peso exponencializado para o efeito sazonal no tempo $t = k - \Delta$. Para determinar essas estimativas, definem-se primeiramente três parâmetros de suavização α, β e γ para então resolver o sistema de equações abaixo:

$$\begin{cases} a_t = \alpha(S_t - s_{t-f}) + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\ s_t = \gamma(S_t - a_t) + (1-\gamma)S_{t-f}, \end{cases} \tag{4.8}$$

onde a_t, b_t e s_t as estimativas para o nível, inclinação e sazonalidade no período t e α, β e γ os parâmetros de suavização. Este cálculo de predição sazonal é chamado de Predição Aditiva de Holt-Winters (ASSIS et al., 2013). Existe também uma versão multiplicativa, mas não será abordada neste trabalho. Versões mais simples do alisamento exponencial, como o Alisamento

Exponencial Duplo (HOLT, 2004) e o Alisamento Exponencial Simples, desconsideram os componentes de sazonalidade e inclinação, respectivamente. Os respectivos parâmetros de suavização também deixam de ser usados, logo não precisam ser definidos.

Além dos três modelos de alisamento exponencial citados, existem outros mais complexos que também são utilizados na predição de séries temporais. Um modelo similar ao Holt-Winters aditivo é o Holt-Winters multiplicativo, também usado para predições de dados sazonais. Outros modelos incluem modelos de regressão linear, usados, por exemplo, para reconhecimento facial (NASEEM; TOGNERI; BENNAMOUN, 2010), e ARMA e ARIMA, ou modelos Box-Jenkins, usados, por exemplo, para predições em sistemas hídricos (VALIPOUR; BANIHABIB; BEHBAHANI, 2013).

Finalmente, a tarefa de predição temporal estatística (linhas 42 a 45 do Algoritmo 2, Página 69) utiliza os modelos de alisamento exponencial (o modelo é escolhido através da variável $tipo\_modelo$) e resolvem o sistema de equações da Equação 4.8 e a Equação 4.7, Página 79, determinando a predição para *score_intermediario(i, j, T + Δ - 1)* como $\widehat{S}_{t+\Delta-1|t}$ e fornecendo-a como a saída final do algoritmo, persistindo o valor obtido (linha 45 do Algoritmo 2, Página 69) e terminando assim a predição temporal estatística utilizando TTTs.

4.7 Comparativo

A Tabela 4.1, Página 81, contém um quadro comparativo entre as técnicas de predição estáticas apresentadas no Capítulo 3 e a estratégia proposta, do ponto de vista de algumas questões importantes, como desempenho.

Por fim, a Tabela 4.2, Página 82, mostra uma comparação entre as métricas temporais vistas no Capítulo 3 e o modelo proposto.

Estratégias	Custo computacional	Desempenho em redes esparsas	Modelagem temporal
Padrões estruturais baseados na vizinhança	Alto, se forem usadas as formas matriciais, devido ao alto custo de operações entre matrizes.	Ruim. A grande quantidade de vértices desconectados ou isolados influencia negativamente nessas métricas. Necessário pré-processamento dos dados para melhorar a predição.	Não há, pois são métricas puramente estáticas.
Padrões estruturais baseados em caminhos			
Low-rank approximations	Médio. Fatoram matrizes, reduzindo a ordem delas.	Em geral, melhora os resultados pois condensa as informações da rede e remove ruídos.	Depende da métrica utilizada como base para o cálculo das fatorações.
Bigramas Escondidos	Baixo, pois consiste na análise do tamanho dos conjuntos de vértices (ou na soma ponderada dos scores de pares de vértices) similares.	Depende do desempenho da métrica (utilizada como base para esta técnica) em redes esparsas.	Não há, pois são métricas puramente estáticas.
Clustering Factorization	Baixo. Consiste na remoção de conexões com os piores scores (pouco relevantes) e recálculo da métrica utilizada.	Bom, pois remove conexões com scores baixos, o que é comum entre vértices isolados dentro da rede.	
Similaridades entre vértices	Baixo. Scores são calculados baseado em conjuntos de características de pares de vértices.	Bom. Considera atributos individuais e comuns de vértices. Vértices distantes ou isolados podem ter atributos diferentes se comparados ao resto da rede.	
Modelos probabilísticos	Médio. Consistem no cálculo de distribuições de probabilidade que necessitam ser mantidas em memória.	-	
Tensor de Transições de Tríades (proposto)	**Alto, devido à contagem de tríades, computação de probabilidades e uso de modelos de dados de 3 dimensões (tensores).**	**Bom, pois desconsidera tríades vazias.**	**Uso de tensores, análise das probabilidades de transições de tríades e predição de séries temporais.**

Tabela 4.1: Quadro comparativo das abordagens estáticas apresentadas no Capítulo 3 e a proposta por este trabalho (última linha).

Estratégias	Custo computacional	Desempenho em redes esparsas	Modelagem temporal
Fatoração de tensores	Alto, devido às operações custosas sobre matrizes (como o cálculo da transposta ou produto entre matrizes).	Em geral, melhora os resultados pois condensa as informações da rede e remove ruídos.	Matrizes/tensores contém snapshots da rede ao longo do tempo. O score é computado utilizando uma heurística (como o produto externo) ou uma predição estatística.
Decomposição canônica			
Séries temporais	Alto. Modelo regressivo que utiliza dados tridimensionais e calcula séries baseadas no histórico da rede.	-	Criação de séries matemáticas, baseadas no histórico de informações (podem ser scores para um par de vértices) da rede, para a predição de valores futuros.
Evolução de subgrafos	Alto, principalmente devido à contagem de tríades ($O(n^3)$ operações) e de operações sobre matrizes.	Bom, pois desconsidera tríades vazias.	Cálculo de TTMs, computação da TTM média e soma de probabilidades para gerar o score final.
Distância temporal média	Alto. Faz muitas buscas em profundidade em redes de larga escala.	Ruim, pois não consegue computar distâncias ou caminhos entre vértices isolados temporalmente.	Uso do modelo de grafos temporais, que preserva explicitamente as informações temporais nas arestas.
Tensor de Transições de Tríades (proposto)	**Alto, devido à contagem de tríades, computação de probabilidades e uso de modelos de dados de 3 dimensões (tensores).**	**Bom, pois desconsidera tríades vazias.**	**Uso de tensores, análise das probabilidades de transições de tríades e predição de séries temporais.**

Tabela 4.2: Quadro comparativo das abordagens temporais apresentadas no Capítulo 3 e a proposta por este trabalho (última linha).

5

Experimentos e Discussão

Este capítulo tratará da avaliação da abordagem proposta neste trabalho. Para tanto, as redes que foram escolhidas serão descritas, a metodologia de experimentação e avaliação será detalhada e os resultados obtidos com a metodologia proposta serão comparados com os resultados obtidos a partir de algumas outras métricas explicadas no Capítulo 3, a fim de verificar se os resultados obtidos pela metodologia proposta são melhores.

5.1 Dados

Para testar o desempenho da métrica apresentada nesse trabalho, foram escolhidas duas redes corporativas. A importância dessa escolha se dá pois são exemplos de redes reais e os dados são de empresas privadas. Além de não existir quase nenhuma base de dados sobre empresas privadas, essas redes permitem fazer uma análise em um contexto fora do comum, observando-se características que nem sempre aparecem em outros tipos de redes: ambas as redes são de envio e recepção de *e-mails*.

Em ambas as bases, um vértice representa uma conta de *e-mail* válida, enquanto uma aresta, que nesse caso é direcionada, representa que um *e-mail* foi enviado do remetente (origem da aresta) ao destinatário.

5.1.1 Base de dados da empresa Enron

A primeira rede[1] engloba os *e-mails* de executivos da empresa americana Enron, uma companhia de energia que trabalhava no ramo de distribuição de energia elétrica e gás natural e no ramo de comunicações. Por causa de um escândalo financeiro, esta empresa foi à falência, e os dados sobre sua rede de *e-mails* foram disponibilizados publicamente para uso em pesquisas.

Os dados sobre a rede estão compreendidos entre os anos de 1999 e 2002. Esses dados foram divididos em meses, partindo do mês de maio de 1999, continuamente até o mês de junho de 2002 (3 anos e 1 mês).

[1] Fonte dos dados: http://konect.uni-koblenz.de/networks/enron

5.1.2 Base de dados da Empresa de Manufatura

Já a segunda rede contém dados de *e-mails* de empregados de uma empresa de manufatura de médio porte (MICHALSKI; PALUS; KAZIENKO, 2011). Os dados dessa rede também foram disponibilizados publicamente, porém, não há como saber quais são os usuários das contas de *e-mails* que pertencem à rede, bem como dados sobre a empresa. Os *e-mails* foram enviados dentro de um período de nove meses, partindo do dia 1º de janeiro de 2010 até o dia 30 de setembro de 2010. Esses dados foram divididos semanalmente, onde os *e-mails* foram enviados em cada semana do ano (*e-mails* enviados durante a primeira semana do ano, segunda semana do ano e assim por diante).

Os dados dessa rede foram divididos semanalmente pois, além do curto período (apenas nove meses), o volume de *e-mails* também é maior, possibilitando uma análise de nível temporal mais detalhado.

A Tabela 5.1 quantifica algumas informações básicas sobre as duas base de dados apresentadas.

	Enron	**Empresa de Manufatura**
Contas de *e-mail*	151	167
***E-mails* enviados (total)**	8.413	82.927
***E-mails* enviados (média mensal)**	221,39	9214,1
Período de observação (meses)	38	9
Total de *links* possíveis	22.650	27.772

Tabela 5.1: Informações gerais sobre as redes abordadas.

5.1.3 Critérios de filtragem dos *e-mails*

Em ambas as redes, não existe distinção entre os tipos de destinatários ("To:", "CC:"ou "BCC:"), tratando cada *e-mail* incluido em um desses campos como um *link* diferente e de mesma importância. É importante frisar que vários *e-mails* podem ter sido enviados de um mesmo remetente para um mesmo destinatário, dentro de uma determinada janela de tempo. Nesse caso, é considerado que apenas uma aresta foi observada no respectivo intervalo de tempo.

Por fim, *links* que levam a contas de *e-mails* de fora da empresa foram excluídos, e *e-mails* enviados para a própria conta (remetente é o mesmo que o destinatário) também foram removidos, pois não refletem uma relação com outros indivíduos da rede.

A Figura 5.1, Página 85, representa graficamente duas séries: a série vermelha, referente à quantidade bruta de *e-mails* por mês da empresa Enron, e a azul, referente à quantidade filtrada a partir da série vermelha, seguindo os critérios de filtragem já explicados. A Figura 5.2, Página 85, mostra graficamente outra série vermelha, dessa vez com a quantidade bruta de *e-mails*, por

semana, da Empresa de Manufatura, e a série azul, com a quantidade filtrada de *e-mails* da série vermelha, seguindo os mesmos critérios de filtragem.

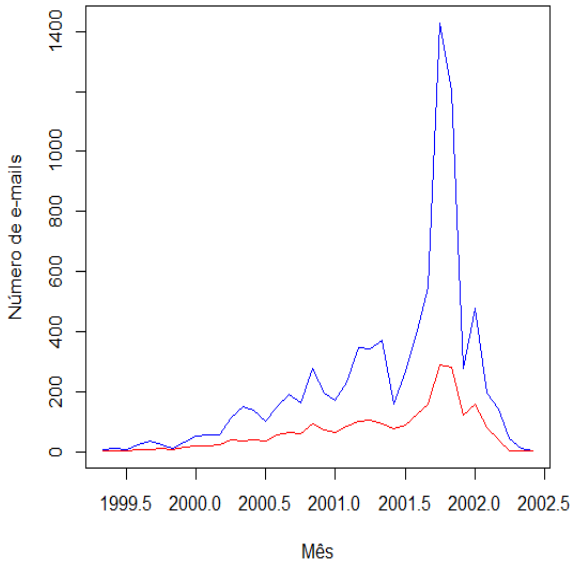

Figura 5.1: Distribuição de e-mails enviados, por mês, da rede da empresa Enron. A série azul representa a quantidade bruta de e-mails, enquanto a série vermelha representa a quantidade filtrada.

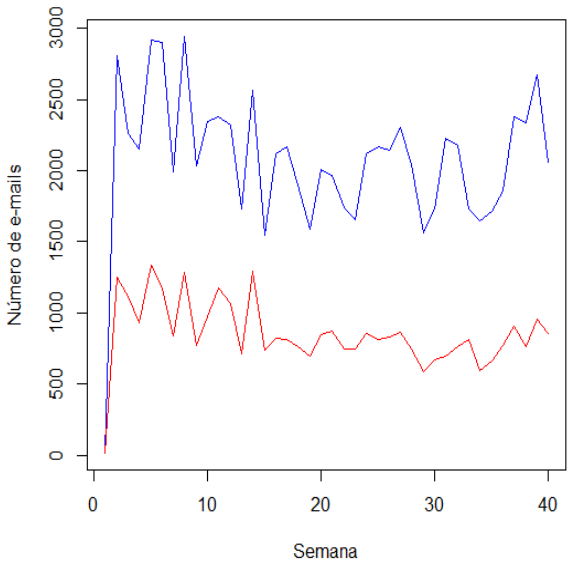

Figura 5.2: Distribuição de e-mails enviados, por semana, da rede da Empresa de Manufatura. A série azul representa a quantidade bruta de e-mails, enquanto a série vermelha representa a quantidade filtrada.

5.2 Desenho Experimental

Para a construção do modelo, assim como a aplicação das outras métricas que serão comparadas, foi utilizado um determinado período de tempo para o treinamento, enquanto os conjuntos de testes são compostos da predição para o mês seguinte, dois meses seguintes e três

meses seguintes, para o caso da rede da Enron, e a semana seguinte, duas semanas seguintes e três semanas seguintes para o caso da Empresa de Manufatura. Os experimentos e avaliações serão detalhados adiante.

Os experimentos realizados neste trabalho tem por objetivos principais mostrar a viabilidade da métrica proposta e verificar se há ganho de qualidade nos resultados obtidos, se comparada a algumas métricas (as métricas escolhidas serão detalhadas adiante) da literatura. Antes de discriminar as técnicas utilizadas para essa comparação, primeiro serão descritas algumas prerrogativas para os experimentos:

- **Amostra**: de um par (i, j) de vértices pertencentes à rede. A amostra, além de conter o par, contém todos os dados compreendidos no período de tempo considerado.
- **Amostras aleatórias**: cada experimento consistirá de amostras escolhidas aleatoriamente, tanto para o conjunto da classe positiva (isto é, pares de vértices que apresentam um *link* entre eles) quanto para o conjunto da negativa (pares de vértices que não possuem *link* entre eles). Além disso, não existirão amostras repetidas dentro de um mesmo experimento.
- **Reamostragem estratificada**: para cada experimento feito, esse tipo de subamostragem aleatória consiste em, primeiramente, descobrir a proporção entre o total de amostras da classe positiva e o total da classe negativa. Essa enumeração é importante pois será usada para manter o balanceamento original da rede quando a reamostragem for feita. A importância de se manter o balanceamento original é de que qualquer alteração pode influenciar, positiva ou negativamente, os resultados das métricas de avaliação que serão utilizadas, assim invalidando toda a experimentação.
- **Medidas de desempenho**: as avaliações serão feitas utilizando duas metodologias: a curva ROC e a curva PR (YANG; LICHTENWALTER; CHAWLA, 2015), utilizadas em trabalhos relacionados (WANG et al., 2015). Primeiramente, são definidas três métricas, sendo elas especificidade, precisão e sensibilidade (ou *recall*, em inglês):

$$Especificidade = \frac{|verdadeiros\ negativos|}{|falsos\ positivos| + |verdadeiros\ negativos|}. \tag{5.1}$$

$$Precisao = \frac{|verdadeiros\ positivos|}{|verdadeiros\ positivos| + |falsos\ positivos|}. \tag{5.2}$$

$$Recall = \frac{|verdadeiros\ positivos|}{|verdadeiros\ positivos| + |falsos\ negativos|}. \tag{5.3}$$

A curva ROC relaciona as métricas de especificidade e *recall*, ambas condicionadas aos verdadeiros positivos e verdadeiros negativos. Para a predição de *links* em redes esparsas, onde cada nova aresta em potencial é classificada, curvas ROC podem mascarar o desempenho real do classificador, pois existem muito menos exemplos da

classe positiva. O efeito prático desse fato é que o classificador pode ter uma alta taxa de falsos positivos (YANG; LICHTENWALTER; CHAWLA, 2015). A curva PR, por outro lado, é condicionada ao que é estimado pelo classificador, isto é, como o classificador classifica cada exemplo, e é mais aplicado quando há interesse em apenas uma das classes (WANG et al., 2015; YANG; LICHTENWALTER; CHAWLA, 2015). Em resumo, a curva PR encontra, do total de amostras classificadas como positivas, quais são verdadeiros positivos e quais são erros. Desse modo, os modelos criados nos experimentos são otimizados para obter o melhor desempenho possível em relação à curva PR. Por fim, para as comparações, serão calculadas as áreas abaixo das curvas ROC (AUC-ROC) e as áreas abaixo das curvas PR (AUC-PR) através do cálculo da integral de cada curva, pois são métricas de avaliação mais robustas (YANG; LICHTENWALTER; CHAWLA, 2015). Vale ressaltar que, enquanto a AUC-ROC máxima possível é igual a 1, para a AUC-PR isto não é verdade. A AUC-PR máxima varia, podendo ser menor que 1.

- **Conjuntos de treinamento e testes** (ver Figura 5.3, Página 88): Como a abordagem proposta neste trabalho é temporal, os conjuntos de treinamento e testes serão divididos de forma um pouco diferente do comum. Normalmente, em aprendizagem de máquina, os dados são organizados do modo como foi descrito na Seção 3.3.1, Página 60. Já os dados sobre as redes das empresas foram divididos em períodos de tempo (mensais, para a Enron, e semanais para a de manufatura). Os conjuntos de treinamento e testes serão formados por conjuntos independentes de amostras diferentes, extraídas do total da rede, e também por intervalos de tempos diferentes dentro dos conjuntos considerados. Por exemplo, o conjunto de treinamento, para a rede da Enron, pode ser definido como todos os dados de amostras aleatórias observados nos primeiros 24 meses, enquanto o conjunto de testes compreenderia os dados observados para os meses subsequentes. Para o caso da Empresa de Manufatura, a ideia é análoga, mudando apenas o fato de que o conjunto de treinamento compreenderia as primeiras semanas, e os de testes, as semanas subsequentes. O conjunto de treinamento será usado para criar os respectivos modelos a serem comparados, e que serão avaliados, utilizando as medidas de desempenho já explicadas, com base no seu desempenho quando aplicado ao conjunto de testes.

- **Métodos competidores**: A três métricas escolhidas para servir de base para a avaliação de desempenho do modelo proposto são: Katz (SA; PRUDENCIO, 2011) e SVD (DUNLAVY; KOLDA; ACAR, 2011) (estáticas) e a TTM média (JUSZCZYSZYN; MUSIAL; BUDKA, 2011) (temporal). Dado que as informações das duas bases das empresas são divididas em T períodos, para o caso das métricas estáticas, foram calculadas as versões matriciais dessas métricas referentes à última observação da rede, ou seja, os valores de Katz no tempo T e os valores de SVD no tempo T.

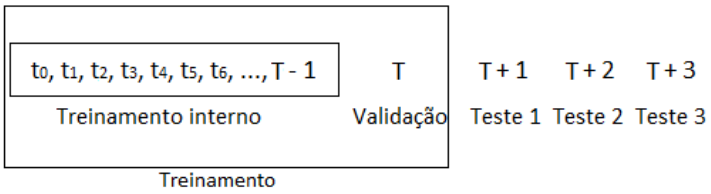

Figura 5.3: Esquema da construção dos conjuntos de treinamento, validação e testes.

Os experimentos propostos tem por base um conjunto obtido através de uma reamostragem estratificada da rede total. Esse conjunto é então dividido em conjunto de treinamento/validação e conjunto de testes. O conjunto de treinamento/validação servirá de base para treinar o modelo que se quer avaliar. O modelo, após treinado, será então testado utilizando o conjunto de testes, e seu desempenho será computado. O desempenho será avaliado segundo as AUC-ROC e AUC-PR obtidas.

Para cada método de predição a ser comparado, foram gerados 30 experimentos diferentes. A avaliação final do método de predição se dá pela obtenção das médias das AUC-ROCs e AUC-PRs observadas em cada experimento, e a comparação entre métodos é feita em termos quantitativos (as maiores médias indicam melhores desempenhos).

Para a avaliação, foram escolhidos dois métodos de predição estáticos e um método de predição temporal. As métricas estáticas escolhidas são o Katz (SA; PRUDENCIO, 2011) e o SVD (DUNLAVY; KOLDA; ACAR, 2011) (referentes à última observação dos dados de treinamento), ambas utilizadas para comparações em experimentos relacionados e com bom desempenho. A métrica de predição temporal escolhida foi a TTM, pois é o modelo de base para o desenvolvimento deste trabalho. Para este caso, foi computada a TTM média das redes. Por fim, para a metodologia proposta, comparou-se também os resultados utilizando três métodos de alisamento exponencial como métodos de predição estatística, sendo eles o Alisamento Exponencial Simples (*Simple Exponential Smoothing* (SES)), Alisamento Exponencial Duplo (*Double Exponential Smoothing* (DES)) (HOLT, 2004) e método preditivo de Holt-Winters (WINTERS, 1976; HOLT, 2004).

Normalmente, quando se tem um modelo de alisamento exponencial com base em uma série temporal, os parâmetros de suavização utilizados são escolhidos de modo a maximizar algum critério de otimização. O critério mais utilizado é minimizar a diferença entre os valores observados na série temporal e os valores preditos pelo modelo. Existem algumas métricas que medem essa diferença, como, por exemplo, o erro médio quadrático (*Mean Square Error* (MSE)), a média dos erros percentuais absolutos (*Mean Absolute Percentage Error* (MAPE)), a raiz quadrada do MSE, entre outros, como visto em (VALIPOUR; BANIHABIB; BEHBAHANI, 2013). Minimizar a diferença entre os valores observados e os calculados pelo modelo visa obter um modelo de predição com os dados mais fiéis à série temporal observada.

No caso deste trabalho, o critério de otimização escolhido não está baseado nos erros da predição dos dados da série temporal. Isso é justificado pelo fato de que os dados observados podem possuir *outliers*, ou dados muito distantes do padrão visto, o que altera os resultados

das predições, já que os parâmetros de suavização escolhidos serão diferentes. Como deseja-se obter o melhor desempenho possível em relação a AUC-PR durante os testes, o critério de otimização utilizado será a AUC-PR calculada a partir do conjunto de treinamento. O conjunto de treinamento, que vai da observação no tempo t_0 até a observação no tempo T, será dividido novamente em outro conjunto de treinamento, que vai do tempo t_0 ao tempo $T-1$, sobre o qual o modelo será construído, e a observação referente ao tempo T servirá como um conjunto de validação, ou seja, as AUC-PRs médias serão calculadas baseadas no tempo T, e os parâmetros de suavização que resultarem nas melhores médias serão aplicados aos conjuntos de testes (ver Figura 5.3, Página 88).

Os parâmetros de suavização, α, β e γ, são definidos como: $0 < \alpha < 1$, $0 < \beta < 1$ e $0 < \gamma < 1$. Existem também os casos onde algum parâmetro não é utilizado, neste caso, γ para os modelos de alisamento exponencial duplo e β e γ para os modelos de alisamento exponencial simples. É interessante observar que quanto mais baixos forem os valores atribuídos a cada parâmetro, de acordo com o sistema de equações mostrado no capítulo anterior, menor será a influência de novas observações adicionadas ao modelo estatístico, a medida que ele é construído. Por existirem muitas combinações possíveis de parâmetros a serem validadas, já que os intervalos a que pertencem os parâmetros são reais, serão mostrados os valores que obtiveram as maiores AUC-PRs médias para o conjunto de validação, juntamente com os parâmetros que maximizam a AUC-PR média.

5.3 Experimentos e resultados

Esta seção apresentará os dados sobre os experimentos realizados nas duas redes, além das avaliações dos resultados obtidos. As métricas que serão comparadas são: Katz e SVD, a TTM média (métrica temporal) e o modelo proposto, o TTT, utilizando os três tipos de alisamento exponencial, explicados no Capítulo 4, para a predição estatística, a fim de descobrir qual apresenta os melhores resultados.

5.3.1 Avaliação de hiperparâmetros

A experimentação da rede da Enron consiste na execução de 30 experimentos, para cada métrica escolhida. Cada experimento consiste em uma reamostragem estratificada de aproximadamente 2% do total de arestas da rede. O conjunto de treinamento compreende o período de Maio de 1999 a Junho de 2001, e os três conjuntos de testes são relativos à Julho, Agosto e Setembro de 2001. Algumas informações sobre os testes estão na Tabela 5.2, Página 90.

A Tabela 5.3, Página 90, apresenta alguns dos resultados da validação feita e os parâmetros utilizados, para o modelo de TTT + SES. O melhor parâmetro está em negrito.

Pelos resultados da validação, verificou-se que α próximo de 0,3 resulta na melhor AUC-PR média. A partir desse valor, uma segunda busca, considerando a casa dos centésimos,

	Julho-2001	**Agosto-2001**	**Setembro-2001**
Número de pares	445	445	445
Pares conectados	2	3	4
Pares não conectados	443	442	441
Proporção	0,005%	0,007%	0,009%

Tabela 5.2: Dados amostrais de um experimento - Enron.

α	**AUC-PR média**	α	**AUC-PR média**
0,1	0,2493	0,6	0,3641
0,2	0,3366	0,7	0,3599
0,3	**0,4370**	0,8	0,3224
0,4	0,3692	0,9	0,3410
0,5	0,3667	-	-

Tabela 5.3: Validação para o modelo TTT + SES - Enron.

foi feita para alfas maiores que 0,3. Entretanto, verificou-se que não houve ganho de desempenho com relação à AUC-PR média e a busca foi encerrada. Logo, 0,3 será o parâmetro utilizado nos testes. Os valores das AUC-PRs médias encontrados na validação e nos testes dificilmente serão os mesmos, mas espera-se que um bom resultado seja obtido com o melhor parâmetro encontrado durante a validação.

A Tabela 5.4 mostra a validação feita para o modelo TTT + DES. O melhor valor encontrado está em negrito.

	β	0,1	0,2	**0,3**	0,4	0,5	0,6	0,7	0,8	0,9
α	-	-	-	-	-	-	-	-	-	-
0,1	|	0,2846	-	-	-	-	-	-	-	-
0,2	|	0,3025	0,3041	-	-	-	-	-	-	-
0,3	|	0,3345	0,3617	0,3815	-	-	-	-	-	-
0,4	|	0,3918	0,3912	**0,4107**	0,3855	-	-	-	-	-
0,5	|	0,3682	0,3640	0,3618	0,3261	0,3453	-	-	-	-
0,6	|	0,3615	0,3511	0,3450	0,3150	0,3195	0,3114	-	-	-
0,7	|	0,3480	0,3179	0,3148	0,3126	0,3321	0,2860	0,2787	-	-
0,8	|	0,3205	0,3133	0,3118	0,3087	0,3069	0,2757	0,2670	0,2592	-
0,9	|	0,3362	0,3340	0,3317	0,2992	0,2957	0,2652	0,2617	0,2328	0,2320

Tabela 5.4: Validação para o modelo TTT + DES - Enron. Os valores da tabela são as AUC-PRs médias encontradas.

A Tabela 5.4 condensa os valores de AUC-PRs médias obtidas ao combinar-se diferentes alfas e betas e realizando o mesmo esquema de buscas feito para o modelo TTT + SES, lembrando que $0 < \alpha < 1$ e $0 < \beta < 1$. Percebe-se que, quando $\alpha = 0{,}4$ e $\beta = 0{,}3$, obtem-se uma média de 0,41. Portanto, os testes utilizarão os valores 0,4 e 0,3, respectivamente. Por fim, a Tabela 5.5, Página 91, mostra os resultados de validação para o modelo TTT + Holt-Winters.

α	β	γ	**AUC-PR média**	α	β	γ	**AUC-PR média**
0,1	0,1	0,1	0,3828	0,2	0,1	0,1	0,4104
0,1	0,1	0,2	0,3835	0,3	0,1	0,1	0,4118
0,1	0,1	0,3	0,4333	0,4	0,1	0,1	0,3827
0,1	**0,1**	**0,4**	**0,4334**	0,5	0,1	0,1	0,3762
0,1	0,1	0,5	0,4314	0,6	0,1	0,1	0,3767

Tabela 5.5: Validação para o modelo TTT + Holt-Winters - Enron.

Para o caso do modelo TTT + Holt-Winters, foram avaliadas menos combinações de parâmetros (mostradas na Tabela 5.5), pois o parâmetro γ aumenta as possibilidades de combinações, lembrando que $0 < \alpha < 1$, $0 < \beta < 1$ e $0 < \gamma < 1$. Além disso, dentre os três modelos utilizados, é o mais custoso de se calcular, já que é calculado a partir do sistema de equações completo descrito pela Equação 4.8, Página 79. Novamente, a combinação de parâmetros que resultou na melhor média encontra-se destacado em negrito na Tabela 5.5.

Já os experimentos da rede da Empresa de Manufatura também consistem na avaliação de 30 experimentos, para cada métrica escolhida. Cada experimento contém uma amostra estratificada correspondente a aproximadamente 1% do total de arestas da rede. O conjunto de treinamento compreende as 26 primeiras semanas do ano de 2010, e os três conjuntos de testes são relativos as 27ª, 28ª e 29ª semanas. Alguns dados sobre os conjuntos de testes se encontram na Tabela 5.6.

	Semana 27	**Semana 28**	**Semana 29**
Número de pares	277	277	277
Pares conectados	9	7	6
Pares não conectados	268	270	271
Proporção	0,034%	0,026%	0,022%

Tabela 5.6: Dados amostrais de um experimento - Empresa de Manufatura.

A Tabela 5.7, Página 92, apresenta alguns dos resultados da validação feita e os parâmetros utilizados, para o modelo de TTT + SES. Percebe-se que, desta vez, o valor de α que produz a melhor AUC-PR média não está entre 0,1 e 0,9, mas sim entre 0,01 e 0,09. Para encontrar esse valor, foi feita uma busca no conjunto inicial de parâmetros. Ao verificar que 0,1 era o melhor parâmetro, uma segunda busca foi feita considerando a casa dos centésimos, para alfas menores que 0,1. Como, durante essa nova busca, foi visto que o ganho não foi significativo (menos de 0,01), a busca foi encerrada. O melhor parâmetro está em negrito.

A Tabela 5.8, Página 92, mostra os valores dos parâmetros α e β e as AUC-PRs médias obtidas.

O mesmo esquema de busca para otimizar os parâmetros foi feito. Desta vez, ao se expandir a busca para a casa dos centésimos, verificou-se que quando α e β são iguais à 0,01, tem-se o melhor valor. Portanto, tanto α como β serão iguais à 0,01 durantes os testes.

α	AUC-PR média	α	AUC-PR média
0,01	0,4698	0,1	0,4663
0,02	0,4703	0,3	0,4454
0,03	0,4696	0,4	0,4315
0,04	**0,4715**	0,5	0,4194
0,05	0,4714	0,7	0,3870
0,06	0,4692	0,8	0,3678
0,07	0,4672	0,9	0,3491

Tabela 5.7: Validação para o modelo TTT + SES - Empresa de Manufatura.

α	β	AUC-PR média	α	β	AUC-PR média
0,01	**0,01**	**0,4495**	0,1	0,1	0.4208
0,02	0,02	0,4491	0,2	0,2	0,3994
0,03	0,03	0,4414	0,3	0,3	0,3704
0,04	0,04	0,4309	0,4	0,4	0,3502
0,05	0,05	0,4257	0,5	0,5	0,3273
0,06	0,06	0,4201	0,6	0,6	0,3068
0,07	0,07	0,4206	0,7	0,7	0,2773
0,08	0,08	0,4198	0,8	0,8	0,2515
0,09	0,09	0,4180	0,9	0,9	0,2375

Tabela 5.8: Validação para o modelo TTT + DES - Empresa de Manufatura.

α	β	γ	AUC-PR média	α	β	γ	AUC-PR média
0,1	0,1	0,1	0,3005	0,2	0,1	0,1	0,3112
0,1	0,1	0,2	0,3068	0,3	0,1	0,1	0,3242
0,1	0,1	0,3	0,3068	0,4	0,1	0,1	0,3235
0,1	0,1	0,4	0,3107	**0,5**	**0,1**	**0,1**	**0,3261**
0,1	0,1	0,5	0,3103	0,6	0,1	0,1	0,3224

Tabela 5.9: Validação para o modelo TTT + Holt-Winters - Empresa de Manufatura.

Finalmente, a Tabela 5.9 mostra a validação para o modelo TTT + Holt-Winters. Os parâmetros em negrito serão os utilizados nos testes, já que produziram a melhor média durante a validação.

5.3.2 Resultados para a rede Enron

A Tabela 5.10, Página 93, mostra os resultados obtidos para cada métrica no primeiro conjunto de testes. A métrica com melhor desempenho está destacada em negrito.

Observando os resultados apresentados na Tabela 5.10, Página 93, é possível verificar um ganho perceptível quando se compara as métricas utilizando a AUC-ROC. As métricas estáticas obtiveram resultados mais próximos, e um pouco distantes da métrica proposta (pelo menos 8 pontos percentuais de diferença), para qualquer um dos modelos de alisamento testados. Com

	AUC-ROC (média)	desvio padrão	AUC-PR (média)	desvio padrão
Katz	0,6218	0,1584	0,1813	0,2887
SVD	0,5941	0,1402	0,0892	0,1810
TTM média	0,6856	0,2236	0,0332	0,0903
TTT + SES	0,7207	0,2433	0,1655	0,2335
TTT + DES	**0,7052**	**0,2445**	**0,2505**	**0,3226**
TTT + Holt-Winters	0,7324	0,2370	0,1272	0,1983

Tabela 5.10: Resultados para o primeiro conjunto de testes - Enron.

relação à TTM média, os resultados foram mais próximos, com 2 pontos percentuais de diferença. Isso quer dizer que a análise estrutural proposta produz *scores* mais altos para amostras da classe positiva, quando comparadas a amostras aleatórias da classe negativa, ou seja, o conjunto de amostras selecionadas pelo classificador (amostras classificadas como pertencentes à classe positiva) é maior.

Ao analisar os valores obtidos pelo cálculo das AUC-PRs, que é o foco deste trabalho, verifica-se que para todas as métricas, os resultados indicam que, em termos absolutos (o quanto a AUC-PR se aproxima de 1), todos os classificadores erram bastante ao sugerir *links*. Em termos relativos (comparando os valores obtidos entre si), observa-se que dentre as métricas avaliadas, o TTT + DES e o Katz possuem as maiores médias, enquanto as outras métricas mostraram um desempenho mais baixo.

Foi feito o teste estatístico de Welch, disponível na linguagem estatística R[2], para comparar as médias de duas amostras independentes e com variâncias desconhecidas: as amostras das AUC-PRs do TTT + DES (média μ_1) e o Katz (média μ_2). Seja H_0 a hipótese nula, em que $\mu_1 > \mu_2$, e H_1 a hipótese alternativa, ou seja, $\mu_1 \leq \mu_2$, deseja-se testar se a hipótese nula deve ser rejeitada para um intervalo de confiança de 99%. A partir do teste estatístico de Welch, foi observado que o *p-value* = 0,8077, maior que o nível de significância de 0,01. Portanto, a hipótese nula não deve ser rejeitada. Com base na aceitação de H_0, é possível concluir que $\mu_1 > \mu_2$. Logo, estatisticamente, modelo de TTT + DES obteve um desempenho melhor quando comparado ao Katz.

Do ponto de vista dos modelos de alisamento testados, os valores obtidos pelas AUC-ROCs e AUC-PRs mostram que a rede é melhor caracterizada quando se consideram apenas a presença dos componentes de nível, inclinação e ruído, discutidos no capítulo anterior, sem considerar mudanças de sazonalidade dos dados. Analisar apenas o ruído e o nível, como é feito no alisamento exponencial simples, resulta em um modelo insuficiente para descrever os dados observados, enquanto tentar considerar a sazonalidade introduz erros na predição (isso indica que os dados não possuem sazonalidade, ou que ela não é bem definida dentro do intervalo de tempo considerado), o que acaba diminuindo o desempenho do seu respectivo modelo. Vale observar também que, apesar do Katz ser uma métrica estática, obteve um bom desempenho com

[2]cran.r-project.org

relação às duas avaliações feitas.

A Tabela 5.11 mostra os resultados obtidos para o conjunto de testes 2:

	AUC-ROC (média)	desvio padrão	AUC-PR (média)	desvio padrão
Katz	**0,6024**	**0,1270**	**0,1211**	**0,1480**
SVD	0,5728	0,1135	0,0694	0,1158
TTM média	0,5846	0,1454	0,0131	0,0075
TTT + SES	0,5962	0,1670	0,0440	0,0813
TTT + DES	0,5662	0,1987	0,0577	0,1099
TTT + Holt-Winters	0,5870	0,1856	0,0519	0,1028

Tabela 5.11: Resultados para o segundo conjunto de testes - Enron.

A medida que se utilizam os modelos treinados para predizer relacionamentos em um futuro cada vez mais distante, é esperado que o desempenho caia. Porém, o que realmente ocorre depende das particularidades de cada rede social. Os resultados obtidos para o segundo conjunto de testes mostra que houve uma queda das médias das AUC-ROCs e AUC-PRs, para todas as métricas testadas. O desempenho de Katz caiu um pouco em relação aos dados do primeiro conjunto de testes. Já para todas as outras métricas, verifica-se uma queda perceptível nas médias das AUC-ROCs, em que as maiores perdas ocorreram nos modelos que utilizam alisamento exponencial.

Verificando os valores das médias das AUC-PRs, a queda de desempenho mais brusca ocorreu exatamente com o modelo de TTT + DES. O modelo de alisamento exponencial duplo calcula o nível e inclinação para cada período de tempo do treinamento, mas aplica o último nível + inclinação calculados nas predições, de forma linear. Verifica-se então que a inclinação calculada não condiz com a observada na Figura 5.1, Página 85. O mesmo ocorre com o modelo de alisamento simples, mas dessa vez, apenas o nível é repetido para todas as predições. Nota-se então que o nível calculado é muito diferente do que pode ser observado na Figura 5.1, Página 85. As métricas restantes não apresentaram quedas tão significativas, porém seus resultados ainda continuam baixos quando comparados ao Katz.

O aumento no número de *links* e outros fatores ajudam a entender a diferença entre os níveis e inclinações calculados e observados e justifica a queda de desempenho para o segundo conjunto de testes. Desta vez, o modelo que obteve o melhor desempenho foi o Katz.

O último conjunto de testes tem seus dados sumarizados na Tabela 5.12, Página 95.

Os dados apresentados referentes ao último conjunto de testes mostram um comportamento diferente do que foi apresentado até o momento. Do ponto de vista das AUC-ROCs, as métricas testadas tiveram desempenho similar, com exceção da métrica SVD, que teve um desempenho um pouco pior. Ao estudar os valores das AUC-PRs, verifica-se que houve uma melhora com relação ao conjunto de testes anterior. Apenas a métrica SVD teve desempenho pior se comparado ao segundo conjunto de testes, mesmo não sendo uma queda muito significativa. Os aumentos mais significativos podem ser justificados pelo aumento no número de *links*

	AUC-ROC (média)	desvio padrão	AUC-PR (média)	desvio padrão
Katz	**0,5922**	**0,0977**	**0,1710**	**0,1698**
SVD	0,5404	0,0620	0,0450	0,0768
TTM média	0,5991	0,1850	0,0215	0,0157
TTT + SES	0,5882	0,1813	0,1369	0,1674
TTT + DES	0,5942	0,1929	0,1057	0,1325
TTT + Holt-Winters	0,6032	0,1851	0,0921	0,1251

Tabela 5.12: Resultados para o terceiro conjunto de testes - Enron.

observados por mês (ver Figura 5.1, Página 85), entre os meses de maio e dezembro de 2001 (com pico próximo justamente à setembro de 2001, mês referente ao terceiro conjunto de testes), o que aumenta a probabilidade de novos *links* terem surgido, bem como a probabilidade de *links* anteriores terem se mantido.

Como a diferença entre as médias das AUC-ROCs dos modelos TTT + SES e TTT + DES é ínfima (< 0,1), ao comparar as médias das AUC-PRs, verifica-se que o modelo TTT + SES obtém um melhor desempenho, tendo desempenho inferior apenas à metrica de Katz. Para o terceiro conjunto de testes, novamente a métrica Katz obteve o melhor desempenho.

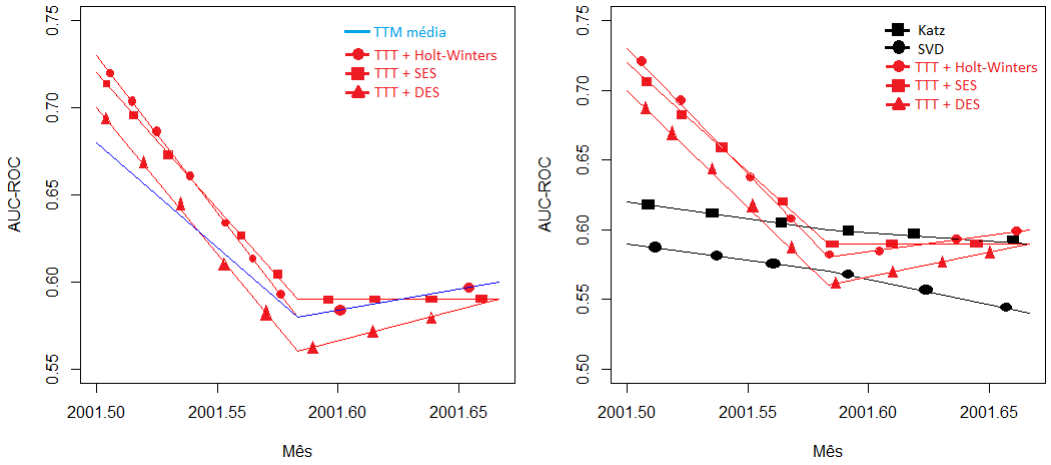

Figura 5.4: Evolução dos valores das AUC-ROCs médias. O gráfico à esquerda compara os modelos de alisamento (vermelho) com a métrica TTM média (azul). O gráfico à direita compara os modelos de alisamento com as métricas estáticas (preto).

A Figura 5.4 mostra uma visão da evolução dos resultados obtidos nos conjuntos de testes, do ponto de vista da avaliação com a curva ROC.

Como foi visto nas tabelas, para a avaliação com a curva ROC, o modelo proposto neste trabalho obteve um desempenho superior em relação às outras métricas, mas fica consideravelmente defasado à medida que é aplicado a observações de tempos posteriores. Esta defasagem ocorre devido à diferença (erro) entre os valores preditos e os observados.

A Figura 5.5, Página 96, mostra a mesma evolução, mas do ponto de vista dos resultados referentes à avaliação com a curva PR.

Novamente, pode-se notar que o modelo proposto, no caso o TTT + DES, consegue um

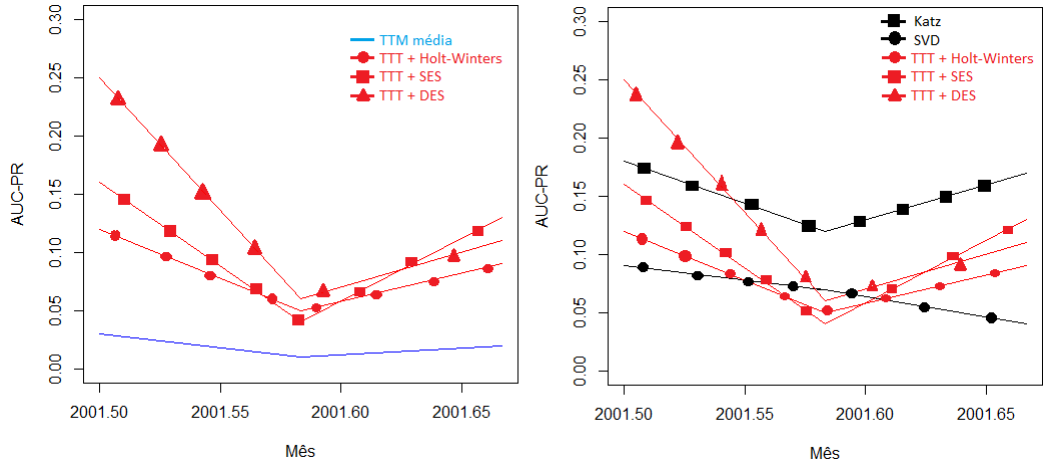

Figura 5.5: Evolução dos valores das AUC-PRs médias. O gráfico à esquerda compara os modelos de alisamento (vermelho) com a métrica TTM média (azul). O gráfico à direita compara os modelos de alisamento com as métricas estáticas (preto).

ótimo desempenho para o primeiro conjunto de testes, quando comparado ao resto das métricas. Entretanto, o mesmo comportamento da curva ROC é observado: os modelos propostos ficam rapidamente defasados ao longo do tempo, e o modelo TTT + DES é superado pelo Katz.

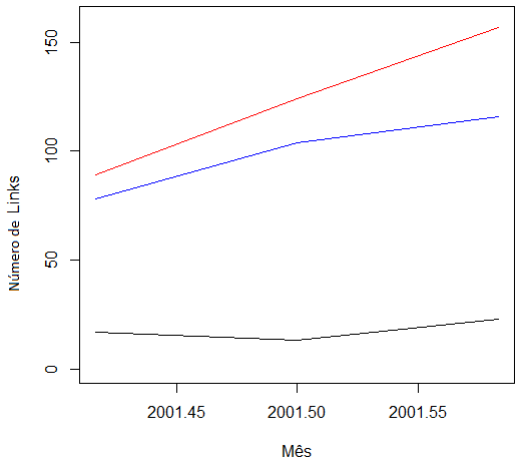

Figura 5.6: Análise dos *links* observados a partir de Julho/2001, com relação a Junho/2001. A série vermelha mostra o total de *links* no mês, a preta mostra o número de *links* que se manteram e a azul, o número de *links* novos que contém vértices observados em Junho/2001

O motivo da métrica Katz ter tido um desempenho melhor para os testes da rede Enron pode ser explicado pelas particularidades da rede. Além do aumento do total de *links* (ver Figura 5.1, Página 85), fazendo uma análise entre o último período de treinamento (Junho de 2001) e o período de testes, verifica-se que há uma manuntenção de uma porcentagem razoável do número de *links* observados em Junho, como visto na Figura 5.6.

Aliado a isso, observou-se que a maioria dos novos *links* formados nos períodos de testes contém vértices que formaram *links* no último mês do treinamento (ver Figura 5.6), o que indica um possível *cluster* em Junho, e que esse *cluster* cresceu (em termos de número de vértices) e ficou mais denso (mais *links*), aumentando a probabilidade de existerem novos caminhos (ou

caminhos cada vez menores) entre vértices conectados e a probabilidade de se criar um caminho entre vértices que antes eram isolados. Isso pode ter influenciado positivamente a métrica Katz, pois ela calcula o tamanho dos caminhos entre dois vértices, dando maior peso aos menores caminhos.

5.3.3 Resultados para a rede da Empresa de Manufatura

Agora serão analisados os resultados obtidos para a Empresa de Manufatura. Os resultados para a base da Enron foram diferentes do esperado, devido às particularidades da rede da empresa. A Tabela 5.13 sumariza os resultados obtidos com as avaliações baseadas nas curvas ROC e PR, com o modelo mais bem avaliado destacado em negrito.

	AUC-ROC (média)	**desvio padrão**	**AUC-PR (média)**	**desvio padrão**
Katz	0,5090	0,1354	0,2352	0,1521
SVD	0,8404	0,0811	0,3399	0,1303
TTM média	0,9437	0,0277	0,4363	0,1448
TTT + SES	**0,9440**	**0,0275**	**0,4370**	**0,1425**
TTT + DES	0,9378	0,0354	0,4250	0,1428
TTT + Holt-Winters	0,8833	0,0443	0,2886	0,1026

Tabela 5.13: Resultados para o primeiro conjunto de testes - Empresa de Manufatura.

Ao contrário do que foi visto na base da Enron, o desempenho das métricas foi melhor em termos quantitativos. O modelo proposto, utilizando os alisamentos exponenciais simples e duplo, obteve ótimos resultados, e similares à TTM média. As métricas estáticas, entretanto, obtiveram um desempenho mais fraco, e o Katz, em particular, ruim.

Analisando as AUC-ROCs médias, percebe-se que o modelo de TTT + SES (proposto) e TTM média ficam empatados e seguidos de perto pelo modelo de TTT + DES, e obtendo um desempenho superior ao Katz. Para comparar os dados das AUC-PRs, o foco deste trabalho, foi feito o mesmo teste estatístico de Welch descrito na Seção 5.3.2, Página 92. Neste caso, foram feitos três testes estatísticos, com intevalo de confiança de 99%, comparando as três amostras: TTM média (μ_1), TTT + SES (μ_2) e TTT + DES (μ_3). No primeiro teste, a hipótese nula (H_0) diz que $\mu_1 = \mu_2$ e a hipótese alternativa (H_1), que $\mu_1 \neq \mu_2$. Pelo teste estatístico, o *p-value* = 0,9855 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. No segundo teste, H_0 diz que $\mu_2 = \mu_3$ e H_1, que $\mu_2 \neq \mu_3$. Pelo teste estatístico, o *p-value* = 0,7455 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. No terceiro teste, H_0 diz que $\mu_3 = \mu_1$ e H_1, que $\mu_3 \neq \mu_1$. Pelo teste estatístico, o *p-value* = 0,7613 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. Finalmente, pelos testes feitos, pode-se dizer que as três amostras (da TTM média, da TTT + SES e da TTT + DES) possuem a mesma média e que os três modelos obtiveram o mesmo desempenho.

Os resultados desse primeiro conjunto de testes evidenciam uma mesma característica observada na Enron: um modelo estatístico baseado na evolução da estrutura da rede provê uma

ótima base para predições, desde que não considere o componente de sazonalidade. Isso indica, como ocorreu na Enron, que os dados não possuem sazonalidade, ou que a sazonalidade não está bem definida para o período observado.

A Tabela 5.14 mostra as informações para o segundo conjunto de testes, referente à 28ª semana do ano de 2010, com o melhores resultados em negrito:

	AUC-ROC (média)	**desvio padrão**	**AUC-PR (média)**	**desvio padrão**
Katz	0,4271	0,1647	0,1364	0,1177
SVD	0,7991	0,1287	0,2979	0,1705
TTM média	0,9330	0,0390	0,4012	0,1660
TTT + SES	**0,9340**	**0,0387**	**0,4064**	**0,1679**
TTT + DES	0,9349	0,0375	0,3760	0,1738
TTT + Holt-Winters	0,9038	0,0723	0,3655	0,1740

Tabela 5.14: Resultados para o segundo conjunto de testes - Empresa de Manufatura.

Para comparar as AUC-PRs do segundo conjunto de testes, foi utilizado o mesmo teste estatístico de Welch de forma similar a como foi feito no primeiro conjunto de testes. Novamente, foram feitos três testes estatísticos com intevalo de confiança de 99% e comparando três amostras: TTM média (μ_1), TTT + SES (μ_2) e TTT + DES (μ_3). No primeiro teste, a hipótese nula (H_0) diz que $\mu_1 = \mu_2$ e a hipótese alternativa (H_1), que $\mu_1 \neq \mu_2$. Pelo teste estatístico, o *p-value* = 0,9033 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. No segundo teste, H_0 diz que $\mu_2 = \mu_3$ e H_1, que $\mu_2 \neq \mu_3$. Pelo teste estatístico, o *p-value* = 0,4933 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. No terceiro teste, H_0 diz que $\mu_3 = \mu_1$ e H_1, que $\mu_3 \neq \mu_1$. Pelo teste estatístico, o *p-value* = 0,5687 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. Finalmente, pelos testes estatísticos feitos, as três amostras (da TTM média, da TTT + SES e da TTT + DES) possuem a mesma média e os três modelos obtiveram o mesmo desempenho.

Finalmente, a Tabela 5.15 mostra os resultados obtidos para o último conjunto de testes.

	AUC-ROC (média)	**desvio padrão**	**AUC-PR (média)**	**desvio padrão**
Katz	0,4265	0,1818	0,1461	0,1428
SVD	0,8389	0,1019	0,3311	0,1711
TTM média	0,9558	0,0349	0,4718	0,1737
TTT + SES	**0,9557**	**0,0349**	**0,4737**	**0,1812**
TTT + DES	0.9515	0.0343	0.4166	0.1742
TTT + Holt-Winters	0,9402	0,0405	0,3753	0,1760

Tabela 5.15: Resultados para o terceiro conjunto de testes - Empresa de Manufatura.

De acordo com os dados apresentados, novamente os modelos TTT + SES, TTT + DES e a métrica TTM média obtiveram desempenhos iguais, pela avaliação com a curva ROC. Por fim, a comparação das AUC-PRs do último conjunto de testes foi feita utilizando o mesmo teste

estatístico de Welch, para a TTM média e o TTT + SES (duas maiores médias amostrais). Com intevalo de confiança de 99%, a amostra da TTM média (μ_1) foi comparada com a amostra do TTT + SES (μ_2). No teste realizado, a hipótese nula (H_0) diz que $\mu_1 = \mu_2$ e a hipótese alternativa (H_1), que $\mu_1 \neq \mu_2$. Pelo resultado do teste estatístico, o *p-value* = 0,9672 é maior que o nível de significância de 0,01 e, portanto, H_0 não deve ser rejeitada. De posse disso, as duas amostras (da TTM média e do TTT + SES) possuem, estatisticamente, a mesma média e os modelos obtiveram o mesmo desempenho. O fato dos modelos terem obtido o mesmo desempenho se dá, em parte, pela manutenção, no período de testes, de uma quantidade razoável de conexões vistas no treinamento (ver Figura 5.7).

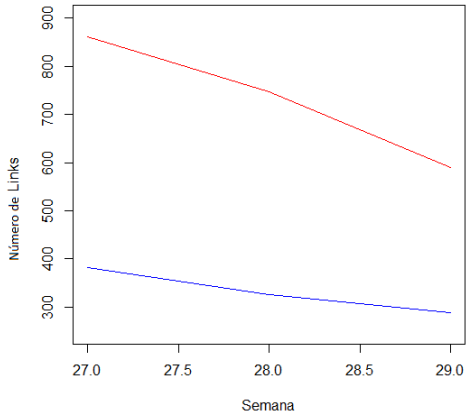

Figura 5.7: Análise dos *links* observados a partir da 27ª semana de 2010. A série vermelha mostra o total de *links* na semana e a azul, o número de *links* que se manteram desde a 26ª semana.

Para entender a razão dos modelos TTT + SES e TTT + DES e a métrica TTM média terem obtido praticamente o mesmo desempenho durante todos os testes, basta observar, como na Figura 5.8, os valores das predições de cada modelo estatístico e da média dos *scores* observados.

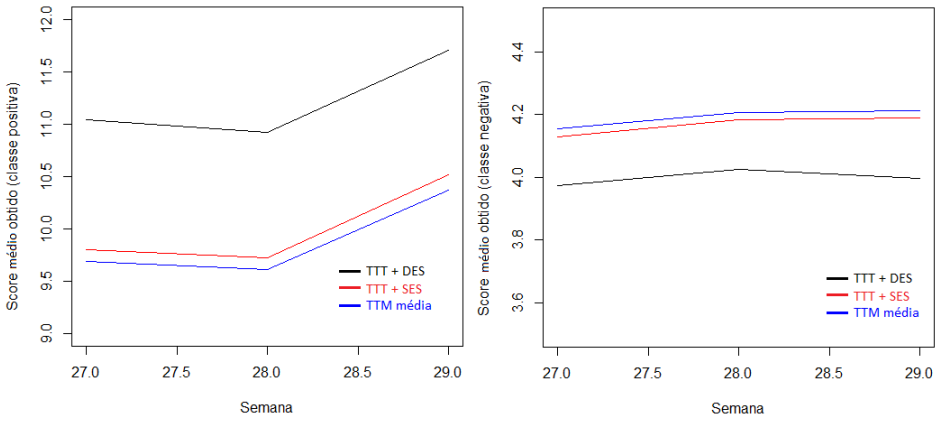

Figura 5.8: Comparação dos *scores* obtidos para as melhores métricas. A série preta representa a predição média do modelo TTT + DES, a série vermelha, a predição média do modelo TTT + SES e a série azul, a TTM média.

Tanto para os conjuntos da classe positiva tanto para os da negativa, a média dos valores obtidos com as predições do modelo TTT + SES foram praticamente iguais à TTM média, para

todos os três conjuntos de testes, o que resultou no desempenho praticamente igual, como pode ser visto na Figura 5.8, Página 99.

Do mesmo modo que foi mostrado para a rede da Enron, as Figuras 5.9 e 5.10 mostram a evolução dos desempenhos para a rede da Empresa de Manufatura, levando em conta as avaliações com a curva ROC e com a curva PR, respectivamente. As escalas das duas figuras estão diferentes para facilitar a visualização da evolução de cada modelo, principalmente dos modelos TTT + SES, TTT + DES e TTM média, que tiveram desempenhos similares.

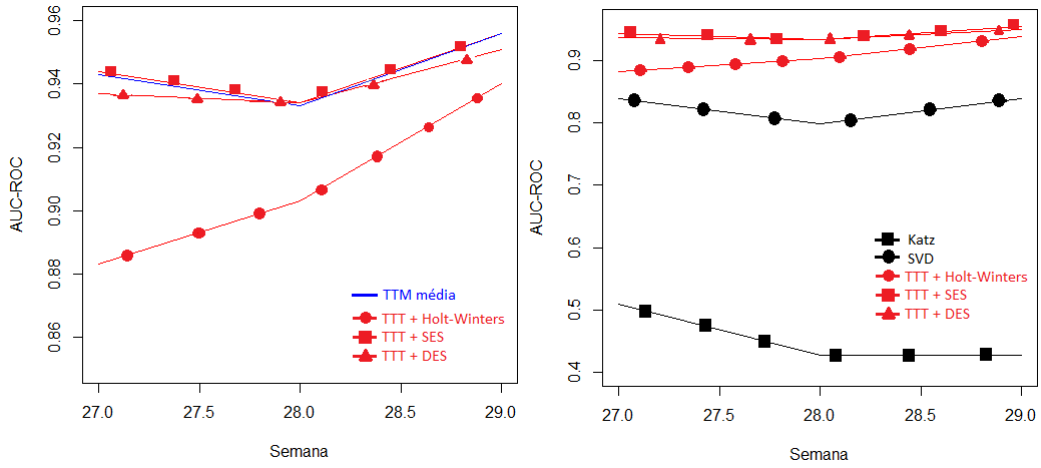

Figura 5.9: Evolução dos valores das AUC-ROCs médias. O gráfico à esquerda compara os modelos de alisamento (vermelho) com a métrica TTM média (azul). O gráfico à direita compara os modelos de alisamento com as métricas estáticas (preto).

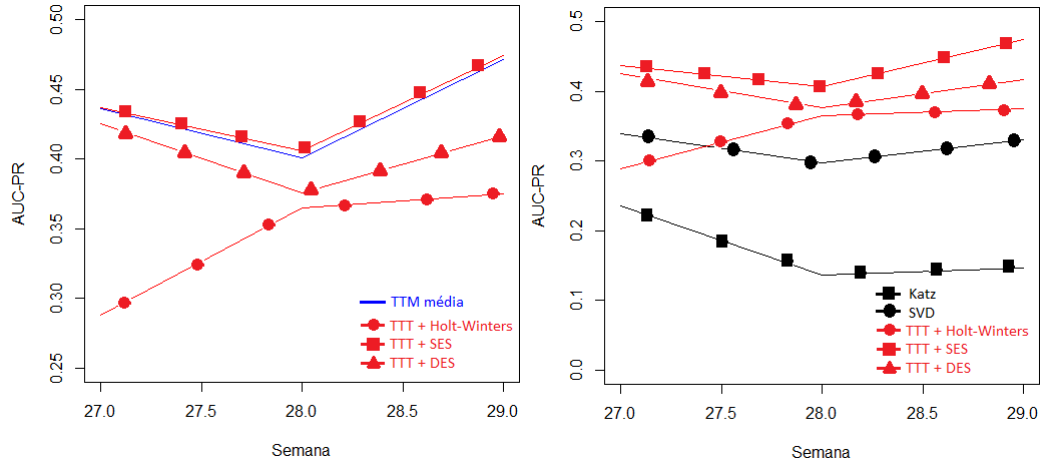

Figura 5.10: Evolução dos valores das AUC-PRs médias. O gráfico à esquerda compara os modelos de alisamento (vermelho) com a métrica TTM média (azul). O gráfico à direita compara os modelos de alisamento com as métricas estáticas (preto).

5.4 Considerações finais

Neste capítulo, foram apresentados os experimentos e testes para avaliação do modelo proposto. Para a experimentação e testes, foram escolhidas duas redes de *e-mails* corporativos.

Essas redes foram escolhidas por serem dados reais, de difícil obtenção (maioria das fontes não possuem esses dados com informações temporais) e diferentes das redes de coautoria que geralmente são utilizadas em vários trabalhos relacionados. Além disso, as métricas escolhidas para realização dos testes foram o SVD e o Katz (métricas estáticas), a TTM média como métrica temporal e o modelo proposto utilizando alisamento exponencial simples, duplo e o método preditivo de Holt-Winters. A grande diferença das métricas escolhidas é que o Katz e SVD são métricas mais robustas e a TTM média é uma métrica temporal. Trabalhos relacionados costumam comparar seus modelos apenas com as métricas estáticas mais simples, como o número de vizinhos em comum.

Também foi visto que o modelo proposto neste trabalho tem sempre o melhor desempenho para predições de curto prazo. A medida que o tempo passa, porém, verificou-se que o modelo fica rapidamente defasado caso a rede não mantenha o comportamento visto durante o treinamento. Foi visto também que isto ocorre pelo fato dos modelos de alisamento aditivos manterem o mesmo nível para as predições e modificando, linearmente, o valor final em função da inclinação e da sazonalidade, quando aplicáveis. Especificamente, para a rede da Enron, esta foi a razão do modelo TTT + DES proposto ter se destacado no primeiro conjunto de testes, mas ter ficado defasado a partir do segundo conjunto de testes, sendo superado pelo Katz. Foi explicado também a razão da métrica Katz ter conseguido um bom desempenho para a rede da Enron.

Para a rede da Empresa de Manufatura, como, durante o período de testes, o comportamento da rede permaneceu quase o mesmo, tanto o modelo TTT + SES quanto o TTT + DES obtiveram ótimos desempenhos ao longo de todos os três testes. Foi visto, também, que o motivo da TTM média conseguir o mesmo desempenho do TTT + SES se deu pelo fato de que as predições médias dos modelos estatísticos, tanto para o conjunto da classe positiva quanto para o da negativa, foram muito próximas das médias da classe positiva e negativa, respectivamente, uma coincidência que não ocorreu na rede da Enron. Por fim, a métrica Katz, que teve um bom desempenho na rede da Enron, mostrou-se insuficiente para servir como método preditor para a rede da Empresa de Manufatura.

Por fim, com base nos testes realizados, verificou-se que o modelo TTT + Holt-Winters não obteve um bom desempenho quando comparado às outras métricas testadas, visto que os dados ou não eram sazonais, ou o período de observação não foi suficiente para que a sazonalidade dos dados fosse definida. Visto os melhores resultados dos conjuntos de testes e de acordo com os testes estatísticos apresentados, pode-se afirmar que o modelo que gera melhores predições de curto prazo é o TTT + DES e que ele funciona muito bem para as duas redes, já que na rede da Empresa de Manufatura, a diferença da avaliação para o modelo TTT + SES é desprezível.

6

Considerações finais

Este capítulo apresenta algumas considerações finais sobre o trabalho desenvolvido, citando as contribuições obtidas pela pesquisa e comparando os resultados obtidos com os objetivos traçados no primeiro capítulo. Além disso, será feita uma análise das limitações da metodologia proposta, para assim então enumerar possíveis melhorias para trabalhos futuros.

A pesquisa desenvolvida neste trabalho compreendeu os seguintes passos: revisão da literatura, compreensão de conceitos, análise de trabalhos relacionados e o desenvolvimento e avaliação da métrica resultante.

O levantamento dos estudos mais relevantes sobre a Predição de Links permitiu ter-se uma visão mais ampla sobre o estado da arte desse problema. Foi visto também que a Predição de Links está relacionada à ARS e que existem vários contextos onde o problema pode ser identificado, como em redes de coautoria, corporativas, de terrorismo, sistema nervoso humano, geologia, entre outros. Além disso, identificou-se que vários fatores melhoram a qualidade da predição de relacionamentos e que a adição do conceito de temporalidade, em particular, contribui positivamente neste âmbito. A revisão dos conceitos fundamentais sobre grafos e suas representações contribuiu para uma modelagem mais intuitiva dos dados e para a implementação de funções mais otimizadas. Por fim, ajudou a traçar os objetivos deste trabalho e a definir um escopo prático para a aplicação das metodologias vistas na literatura.

Além disso, foram identificadas algumas vantagens e desvantagens dos estudos abordados, comparados na Seção 4.7, Página 80, principalmente da TTM. Estas vantagens e desvantagens serviram de base para sugerir algumas melhorias para o trabalho de JUSZCZYSZYN; MUSIAL; BUDKA (2011) e ajudar a criar o modelo proposto neste trabalho.

Com relação ao desenvolvimento do modelo proposto, o trabalho consistiu primeiramente em modelar a rede como uma estrutura temporal. Isso foi possível com a utilização de uma estrutura de dados tridimensional, a TTT, que tem a vantagem de permitir o acesso direto às informações de cada tempo observado. De posse da TTT contendo os dados temporais, a análise temporal proposta constituiu em percorrer as todas as transições das tríades (menor subgrafo não trivial e objeto de análise) relevantes, ao longo de todos os tempos observados, calculando as probabilidades de transições para cada tipo de tríade e condensando essas probabilidades em uma série de *scores* parciais. Estes *scores* foram então analisados como uma série temporal,

com o objetivo de predizer os *scores* seguintes, por meio da aplicação de técnicas de alisamento exponencial aditivo. Os modelos de alisamento utilizados foram o SES, o DES e o modelo de predição Holt-Winters, e desses modelos, os *scores* finais foram obtidos.

A partir dos experimentos realizados no Capítulo 5, foi possível verificar uma melhora considerável dos resultados com a métrica proposta, avaliados sob as perspectivas da curva ROC e da curva PR, em comparação às outras abordagens testadas. A grande diferença deste trabalho para outros similares é que foi incluído, dentre os objetos de comparação, um modelo de predição temporal, além da experimentação em redes de *e-mails* corporativos, que possuem características bastante diferentes (por exemplo, o grande desbalanceamento de classes e relacionamentos unilaterais) das outras redes comumente vistas em outros trabalhos, principalmente redes de coautoria. Finalmente, foi verificado que o modelo que mais se adequa aos dados testados é o modelo de TTT + DES, o que indica que extrair informações de nível, inclinação (tendência) e ruído contribuem positivamente para o modelo preditivo. Em contrapartida, tentar identificar o componente de sazonalidade (periodicidade dos dados) contribuiu negativamente.

Com base no que foi observado acima, podem ser citados, como contribuições, os itens a seguir:

- Análise temporal para a predição de relacionamentos. Este trabalho criou uma abordagem nova, totalmente temporal e sem relação com a grande maioria das outras abordagens vistas na literatura (como predições temporais baseadas em métricas estáticas);
- Criação de uma análise estrutural, temporal e detalhada aplicadas a duas redes corporativas com um desbalanceamento de classes muito grande, diferente das redes utilizadas na maioria dos estudos sobre o assunto;
- Elaboração de um modelo flexível, que permite não só o estudo da evolução de tríades ao longo do tempo, mas também de qualquer outra sub-estrutura do grafo, além de possibilitar a fácil aplicação de outras metodologias de predição estatística às séries temporais obtidas;
- Comparação de resultados com um modelo temporal já existente na literatura, a TTM, além de outras métricas estáticas (Katz e SVD) mais robustas em relação ao que normalmente se encontra em experimentos de trabalhos similares;
- Obtenção de resultados para várias configurações de hiperparâmetros, com um método de busca pelo melhor parâmetro durante a validação. Não é possível, porém, avaliar todas as configurações possíveis porque os intervalos a que pertencem os parâmetros de suavização são reais e contínuos, ou seja, existem infinitos valores a serem testados. Isso quer dizer que podem ainda existir configurações com resultados melhores que os apresentados e que não foram testadas;

- Utilização da área sob a curva ROC, mas principalmente a área sob a curva PR, que também foi utilizada como critério de otimização durante os experimentos. Atualmente, a AUC-PR tem mais importância como método de avaliação quando comparada à AUC-ROC, pois a última apresenta algumas desvantagens, como o fato de que não lida bem com o desbalanceamento de classes, já que um classificador que retorna vários falsos positivos acaba por mascarar os resultados.

6.1 Limitações do trabalho

Este trabalho teve por objetivo principal desenvolver um método de predição temporal que resultasse em melhores resultados, se comparados aos resultados de algumas abordagens presentes na literatura e analisar como adicionar o conceito de temporalidade influencia positivamente os resultados das predições. Todavia, este trabalho foi desenvolvido com algumas restrições, sendo elas:

- Não foram utilizadas, nas comparações de metodologias, todas as metodologias presentes na literatura. Foram escolhidas três métricas, sendo duas estáticas: Katz e SVD, que têm desempenho muito bom (DUNLAVY; KOLDA; ACAR, 2011) para a Predição de Links e uma temporal, a TTM média, que serviu como base para o desenvolvimento deste trabalho. Além de existirem diversas abordagens diferentes, muitas tem um foco muito diferente do proposto por este trabalho, o que torna mais difícil fazer comparações e explicar as diferenças entre os resultados obtidos;

- Devido à limitação de recursos computacionais, as redes escolhidas para executar os experimentos não possuem muitos indivíduos (cada uma tem pouco mais de 100 vértices). Tanto a metodologia desenvolvida quanto as que foram selecionadas da literatura requerem cálculos custosos que envolvem matrizes de grandes dimensões;

- Como foi explicado no Capítulo 4, visto que as redes observadas são muito esparsas, para o método proposto, foram excluídas todas as tríades vazias, o que pode ter impactado, de alguma forma, os resultados obtidos. Não foi investigado, porém, se houve algum impacto e se ele foi positivo ou negativo;

- Para este trabalho, todas as redes observadas são equivalentes à definição de grafo simples, ou seja, um conjunto de vértices (onde cada vértice não possui nenhuma informação extra) e um conjunto de arestas (arestas simples, direcionadas ou não direcionadas, e acíclicas). Grafos com múltiplas arestas entre os mesmos vértices, entre outros tipos, não foram abordados;

- Devido a dificuldade de se conseguir dados públicos e confiáveis sobre redes, ainda mais com informações temporais (como, por exemplo, o *timestamp* de surgimento

de cada aresta), o que é um requisito chave neste trabalho, apenas duas redes foram utilizadas nos experimentos;

- Para os resultados obtidos a partir da validação de parâmetros de suavização explicada no Capítulo 5, apenas um conjunto finito de valores foi validado. Isso foi feito pelo fato de que os valores de cada parâmetro variam em um intervalo real e estes intervalos são contínuos (qualquer número real entre 0 e 1). Este fato implica em uma imensa quantidade de combinações de parâmetros que podem ser testadas. Visto que nem todas as combinações possíveis foram testadas, podem existir combinações de parâmetros que mostrem resultados melhores dos que os apresentados neste trabalho.

Analisando mais detalhadamente o algoritmo proposto, algumas desvantagens podem ser observadas, como, por exemplo:

- Seja n o número de vértices da rede, T o número de intervalos de tempo e K_t o número de tríades não vazias observadas no tempo t, e de acordo com o Algoritmo 2, Página 69, o custo de tempo pode ser dado como: $(T * n^3) + [T * (2 * K_t)]$, onde $(T * n^3)$ é o custo de tempo da tarefa 3 (linhas 17 e 18 do Algoritmo 2, Página 69) e $[T * (2 * K_t)]$ o custo da tarefa 4 (linhas 30 a 40 do Algoritmo 2, Página 69). Para o pior caso, o *big O* é dado por $O(T * (n^3 + K_t))$. Este é um custo alto e que demanda muitos recursos computacionais para que possa ser utilizado em redes de larga escala;

- Redes mais densas (com mais *links* presentes) refletem em um aumento da constante K_t, ou seja, com relação ao tempo de execução, o desempenho do algoritmo piora ao passo que a densidade da rede aumenta;

- O custo computacional de tempo requer que as representações matriciais sejam mantidas na memória, de modo que o acesso às representações seja feito em O(1). Visto isso, o custo de memória também acaba sendo alto;

- O algoritmo atual só pode ser aplicado a redes que possam ser representadas como grafos simples (vértices simples, arestas direcionadas ou não), não podendo, ainda, ser aplicado à grafos bipartidos, com múltiplas arestas, etc;

- O algoritmo ainda não comporta a adição de informações contextuais da rede, como importância de relacionamentos, informações sobre indivíduos (informações pessoais, interesses pessoais, entre outros). Adicionar essas informações (o que é feito em outras abordagens) certamente melhoraria os resultados das predições;

- O algoritmo trata todas as tríades e transições semelhantes como tendo a mesma importância. Dar mais importância (pesos maiores) às tríades e transições mais relevantes e menos importância (pesos menores) às menos relevantes pode impactar positivamente na qualidade das predições. Esta técnica de valoração (atribuição de pesos) das informações já é aplicada em várias abordagens.

6.2 Trabalhos futuros

Tanto as restrições no desenvolvimento do trabalho quanto as desvantagens do algoritmo, citadas anteriormente, permitem que se faça uma nova avaliação do que foi feito até o momento, e dessa avaliação, concluir possibilidades de melhorias e ampliar o campo de análise. Além da Predição de Links, o trabalho proposto pode ser utilizado para outros fins, como o *ranking* de vértices, sistemas de recomendação, entre outros. As seguintes melhorias podem ser aplicadas ao trabalho desenvolvido:

- Utilizar métodos de predição estatística mais complexos que o alisamento exponencial aditivo, como por exemplo, o alisamento exponencial multiplicativo ou o ARIMA (FARUK, 2010; VALIPOUR; BANIHABIB; BEHBAHANI, 2013);
- Buscar outras metodologias de predição estática ou temporal de *links* para posterior comparação e análise de possíveis melhorias;
- Investigar como diminuir o efeito do desbalanceamento entre classes presente na rede, como, por exemplo, a atribuição de pesos aos exemplos das classes positiva e negativa;
- Pesquisar como aplicar o modelo proposto em redes sociais de outros domínios, não só redes de *e-mails* corporativos, e observar o comportamento e particularidades;
- Implementar a adição de informações contextuais da rede e avaliar como essas informações podem ser utilizadas para melhorar a predição, com base em trabalhos acadêmicos similares;
- Tentar reduzir o custo computacional, tentando implementar versões matriciais (ou outras otimizações que reduzam o custo de tempo) das operações definidas no Capítulo 4. Outra alternativa é a implementação de um algoritmo paralelo e distribuído, onde cada computador da rede é responsável por uma parcela do cálculo dos *scores*.
- Avaliar um meio de dar mais importância às tríades e transições relevantes, de modo a melhorar os *scores* calculados;
- Tentar relacionar o desempenho do modelo com relação a AUC-PR (ou outro critério de otimização escolhido) e os crítérios de otimização baseados no erro de predição, para tentar diminuir o espaço de busca de parâmetros durante a otimização pelo critério escolhido. Também, variar mais os parâmetros utilizados para os experimentos deste trabalho, de modo a descobrir se existem resultados melhores ou se os apresentados neste trabalho estão otimizados ao máximo.

ADAMIC, L. A.; ADAR, E. Friends and Neighbors on the Web. **Social Networks**, [S.l.], v.25, n.3, p.211–230, 2003.

ALBERT, R.; BARABASI, A.-L. Statistical mechanics of complex networks. **Rev. Mod. Phys.**, [S.l.], v.74, n.1, p.47–97, Jan. 2002.

ALBERT, R.; JEONG, H.; BARABASI, A. L. The Diameter of the World Wide Web. **Nature**, [S.l.], v.401, p.130–131, 1999.

ASSIS, M. V. O. de et al. Holt-Winters statistical forecasting and ACO metaheuristic for traffic characterization. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC), 2013. **Anais...** [S.l.: s.n.], 2013. p.2524–2528.

BACKSTROM, L. et al. Group Formation in Large Social Networks: membership, growth, and evolution. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 12., New York, NY, USA. **Proceedings...** ACM, 2006. p.44–54. (KDD '06).

BACKSTROM, L.; LESKOVEC, J. Supervised random walks: predicting and recommending links in social networks. In: WSDM. **Anais...** ACM, 2011. p.635–644.

BARABASI, A. **Linked**: the new science of networks. Cambridge, Mass.: Perseus Pub., 2002.

BARABASI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, [S.l.], n.5439, p.509–512, 1999.

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

BATAGELJ, V.; MRVAR, A. A subquadratic triad census algorithm for large sparse networks with small maximum degree. **Social Networks**, [S.l.], v.23, n.3, p.237 – 243, 2001.

BENCHERIF, M. A. et al. Fusion of Extreme Learning Machine and Graph-Based Optimization Methods for Active Classification of Remote Sensing Images. **IEEE Geoscience and Remote Sensing Letters**, [S.l.], v.12, n.3, p.527–531, March 2015.

BENCHETTARA, N.; KANAWATI, R.; ROUVEIROL, C. A Supervised Machine Learning Link Prediction Approach for Academic Collaboration Recommendation. In: FOURTH ACM CONFERENCE ON RECOMMENDER SYSTEMS, New York, NY, USA. **Proceedings...** ACM, 2010. p.253–256. (RecSys '10).

BEUTEL, A.; AKOGLU, L.; FALOUTSOS, C. Graph-Based User Behavior Modeling: from prediction to fraud detection. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 21. **Proceedings...** ACM, 2015. p.2309–2310. (KDD '15).

BRINGMANN, B. et al. Learning and Predicting the Evolution of Social Networks. **Intelligent Systems, IEEE**, [S.l.], v.25, n.4, p.26–35, July 2010.

CARLEY, K. M. **Dynamic network analysis**. [S.l.]: Citeseer, 2003.

CARROLL, J. D.; CHANG, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. **Psychometrika**, [S.l.], v.35, n.3, p.283–319, 1970.

CHEN, H.; LI, X.; HUANG, Z. Link prediction approach to collaborative filtering. In: DIGITAL LIBRARIES, 2005. JCDL '05. PROCEEDINGS OF THE 5TH ACM/IEEE-CS JOINT CONFERENCE ON. **Anais...** [S.l.: s.n.], 2005. p.141–142.

CLAUSET, A.; MOORE, C.; NEWMAN, M. E. J. Hierarchical structure and the prediction of missing links in networks. **Nature**, [S.l.], v.453, p.98–101, 2008.

COOKE, R. J. E. **Link prediction and link detection in sequences of large social networks using temporal and local metrics.** 2006. Dissertação (Mestrado em Ciência da Computação) — Department of Computer Science, University of Cape Town.

CORMODE, G.; KRISHNAMURTHY, B. Key differences between Web 1.0 and Web 2.0. **First Monday**, [S.l.], v.13, n.6, 2008.

CRAFT, M. E. Infectious disease transmission and contact networks in wildlife and livestock. **Philosophical Transactions of the Royal Society of London B: Biological Sciences**, [S.l.], v.370, n.1669, 2015.

DAVIDSEN, J.; EBEL, H.; BORNHOLDT, S. Emergence of a Small World from Local Interactions: modeling acquaintance networks. **Phys. Rev. Lett.**, [S.l.], v.88, p.128701, Mar 2002.

DAVIS, D.; LICHTENWALTER, R.; CHAWLA, N. V. Multi-relational Link Prediction in Heterogeneous Information Networks. In: ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM), 2011 INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2011. p.281–288.

DAVIS, J. A.; LEINHARDT, S. **The Structure of Positive Interpersonal Relations in Small Groups**. [S.l.]: Distributed by ERIC Clearinghouse [Washington, D.C.], 1967. 54 p.p.

DONG, Y. et al. Link Prediction and Recommendation across Heterogeneous Social Networks. In: IEEE 12TH INTERNATIONAL CONFERENCE ON DATA MINING, 2012. **Anais...** [S.l.: s.n.], 2012. p.181–190.

DOPPA, J. et al. Learning Algorithms for Link Prediction based on Chance Constraints. In: EUROPEAN CONFERENCE ON MACHINE LEARNING (ECML). **Anais...** [S.l.: s.n.], 2010.

DUNLAVY, D. M.; KOLDA, T. G.; ACAR, E. Temporal Link Prediction Using Matrix and Tensor Factorizations. **ACM Trans. Knowl. Discov. Data**, New York, NY, USA, v.5, n.2, p.10:1–10:27, Feb. 2011.

ERDöS, P.; RéNYI, A. On Random Graphs I. **Publicationes Mathematicae Debrecen**, [S.l.], v.6, p.290, 1959.

ESLING, P.; AGON, C. Time-series Data Mining. **ACM Comput. Surv.**, New York, NY, USA, v.45, n.1, p.12:1–12:34, Dec. 2012.

FARUK, D. O. A hybrid neural network and {ARIMA} model for water quality time series prediction. **Engineering Applications of Artificial Intelligence**, [S.l.], v.23, n.4, p.586 – 594, 2010.

FREEMAN, L. C. Centrality in social networks conceptual clarification. **Social Networks**, [S.l.], v.1, n.3, p.215 – 239, 1979.

FRIEDMAN, N. et al. Learning Probabilistic Relational Models. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE - VOLUME 2, 16., San Francisco, CA, USA. **Proceedings...** Morgan Kaufmann Publishers Inc., 1999. p.1300–1307. (IJCAI'99).

GETOOR, L. **Learning Statistical Models from Relational Data**. 2001. Tese (Doutorado em Ciência da Computação) — Stanford University.

GETOOR, L.; DIEHL, C. P. Link Mining: a survey. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.7, n.2, p.3–12, Dec. 2005.

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

GRANOVETTER, M. The Strength of Weak Ties. **The American Journal of Sociology**, [S.l.], v.78, n.6, p.1360–1380, 1973.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining**: concepts and techniques. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems).

HASAN, M. A. et al. Link Prediction Using Supervised Learning. In: IN PROC. OF SDM 06 WORKSHOP ON LINK ANALYSIS, COUNTERTERRORISM AND SECURITY. **Anais...** [S.l.: s.n.], 2006.

HECKERMAN, D.; MEEK, C.; KOLLER, D. **Probabilistic Models for Relational Data**. [S.l.]: Microsoft Research, 2004. (MSR-TR-2004-30).

HECKERMAN, D.; MEEK, C.; KOLLER, D. **Probabilistic Entity-Relationship Models, PRMs and Plate Models**. [S.l.]: MIT Press, 2007. p.201–239.

HINDLE, D. Noun Classification from Predicate-argument Structures. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 28., Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 1990. p.268–275. (ACL '90).

HOLT, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. **International Journal of Forecasting**, [S.l.], v.20, n.1, p.5–10, 2004.

HUANG, Z. Link Prediction Based on Graph Topology: the predictive value of the generalized clustering coefficient. In: WORKSHOP ON LINK ANALYSIS: DYNAMICS AND STATIC OF LARGE NETWORKS (LINKKDD2006). **Anais...** [S.l.: s.n.], 2006.

HUANG, Z.; LIN, D. K. J. The Time-Series Link Prediction Problem with Applications in Communication Surveillance. **INFORMS J. on Computing**, Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA, v.21, n.2, p.286–303, Apr. 2009.

JEH, G.; WIDOM, J. SimRank: a measure of structural-context similarity. In: EIGHTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, New York, NY, USA. **Proceedings...** ACM, 2002. p.538–543. (KDD '02).

JEONG, H. et al. The large-scale organization of metabolic networks. **Nature**, [S.l.], v.407, n.6804, p.651–654, Oct. 2000.

JIN, E. M.; GIRVAN, M.; NEWMAN, M. E. J. Structure of growing social networks. **Phys. Rev. E**, [S.l.], v.64, p.046132, Sep 2001.

JUSZCZYSZYN, K.; MUSIAL, K.; BUDKA, M. Link Prediction Based on Subgraph Evolution in Dynamic Social Networks. In: PRIVACY, SECURITY, RISK AND TRUST (PASSAT) AND 2011 IEEE THIRD INERNATIONAL CONFERENCE ON SOCIAL COMPUTING (SOCIALCOM), 2011 IEEE THIRD INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2011. p.27–34.

KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, [S.l.], v.18, n.1, p.39–43, 1953.

KEELING, M. J.; EAMES, K. T. D. Networks and epidemic models. **J. R. Soc. Interface**, [S.l.], v.2, p.295, 2005.

KEMPE, D.; KLEINBERG, J.; KUMAR, A. Connectivity and Inference Problems for Temporal Networks. **Journal of Computer and System Sciences**, [S.l.], v.64, n.4, p.820 – 842, 2002.

KLEINBERG, J. M. Authoritative Sources in a Hyperlinked Environment. **J. ACM**, New York, NY, USA, v.46, n.5, p.604–632, Sept. 1999.

KOLLER, D.; PFEFFER, A. Probabilistic Frame-based Systems. In: FIFTEENTH NATIONAL/TENTH CONFERENCE ON ARTIFICIAL INTELLIGENCE/INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE, Menlo Park, CA, USA. **Proceedings...** American Association for Artificial Intelligence, 1998. p.580–587. (AAAI '98/IAAI '98).

KREBS, V. Uncloaking Terrorist Networks. **First Monday**, [S.l.], v.7, n.4, 2002.

KUMAR, R. et al. The Web As a Graph. In: NINETEENTH ACM SIGMOD-SIGACT-SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS, New York, NY, USA. **Proceedings...** ACM, 2000. p.1–10. (PODS '00).

KUNEGIS, J.; LOMMATZSCH, A.; BAUCKHAGE, C. The slashdot zoo: mining a social network with negative edges. In: WORLD WIDE WEB, 18., New York, NY, USA. **Proceedings...** ACM, 2009. p.741–750. (WWW '09).

Lü, L.; ZHOU, T. Link prediction in complex networks: a survey. **Physica A**, [S.l.], v.390, n.6, p.11501170, 2011.

LESKOVEC, J.; HUTTENLOCHER, D.; KLEINBERG, J. Predicting Positive and Negative Links in Online Social Networks. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19., New York, NY, USA. **Proceedings...** ACM, 2010. p.641–650. (WWW '10).

LIBEN-NOWELL, D.; KLEINBERG, J. The Link Prediction Problem for Social Networks. In: TWELFTH INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, New York, NY, USA. **Proceedings...** ACM, 2003. p.556–559. (CIKM '03).

LIBEN-NOWELL, D.; KLEINBERG, J. The Link-prediction Problem for Social Networks. **J. Am. Soc. Inf. Sci. Technol.**, New York, NY, USA, v.58, n.7, p.1019–1031, May 2007.

LICHTENWALTER, R.; LUSSIER, J. T.; CHAWLA, N. V. New perspectives and methods in link prediction. In: KDD. **Anais...** ACM, 2010. p.243–252.

LU, L.; ZHOU, T. Link prediction in weighted networks: the role of weak ties. **EPL (Europhysics Letters)**, [S.l.], v.89, n.1, p.18001, 2010.

MICHALSKI, R.; PALUS, S.; KAZIENKO, P. Matching Organizational Structure and Social Network Extracted from Email Communication. In: LECTURE NOTES IN BUSINESS INFORMATION PROCESSING. **Anais...** Springer Berlin Heidelberg, 2011. v.87, p.197–206.

MOODY, J. Matrix methods for calculating the triad census. **Social Networks**, [S.l.], v.20, n.4, p.291 – 299, 1998.

MURATA, T.; MORIYASU, S. Link Prediction based on Structural Properties of Online Social Networks. **New Generation Computing**, [S.l.], v.26, n.3, p.245–257, 2008.

NASEEM, I.; TOGNERI, R.; BENNAMOUN, M. Linear Regression for Face Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.32, n.11, p.2106–2112, Nov 2010.

NEVILLE, J. **Statistical models and analysis techniques for learning in relational data**. 2006. Tese (Doutorado em Ciência da Computação) — University of Massachusetts Amherst.

NEWMAN, M. Clustering and preferential attachment in growing networks. **Physical Review E**, [S.l.], v.64, n.2, p.025102, 2001.

NEWMAN, M. The Structure and Function of Complex Networks. **SIAM review**, [S.l.], v.45, n.2, p.167–256, 2003.

NEWMAN, M.; BARABASI, A.; WATTS, D. **The Structure and Dynamics of Networks**. [S.l.]: Princeton University Press, 2011. (Princeton Studies in Complexity).

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

PAGE, L. et al. The PageRank citation ranking: bringing order to the web. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 7., Brisbane, Australia. **Proceedings...** [S.l.: s.n.], 1998. p.161–172.

PAN, R. et al. One-Class Collaborative Filtering. In: ICDM '08. EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2008. **Anais...** [S.l.: s.n.], 2008. p.502–511.

POOLE, D. Probabilistic Horn abduction and Bayesian networks. **Artificial Intelligence**, [S.l.], v.64, n.1, p.81 – 129, 1993.

POTGIETER, A. et al. Temporality in Link Prediction: understanding social complexity. **E:CO**, [S.l.], v.11, 2009.

ROXIN, A.; RIECKE, H.; SOLLA, S. A. Self-Sustained Activity in a Small-World Network of Excitable Neurons. **Phys. Rev. Lett.**, [S.l.], v.92, p.198101, May 2004.

SA, H. de; PRUDENCIO, R. Supervised link prediction in weighted networks. In: NEURAL NETWORKS (IJCNN), THE 2011 INTERNATIONAL JOINT CONFERENCE ON. **Anais...** [S.l.: s.n.], 2011. p.2281–2288.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. New York, NY, USA: McGraw-Hill, Inc., 1983.

SANTORO, N. et al. Time-Varying Graphs and Social Network Analysis: temporal indicators and metrics. **ArXiv e-prints**, [S.l.], Feb. 2011.

SCELLATO, S. et al. Small-world behavior in time-varying graphs. **Phys. Rev. E**, [S.l.], v.81, p.055101, May 2010.

SOLOMONOFF, R.; RAPOPORT, A. Connectivity of random nets. **The bulletin of mathematical biophysics**, [S.l.], v.13, n.2, p.107–117, 1951.

SPIEGEL, S. et al. Link Prediction on Evolving Data Using Tensor Factorization. In: CAO, L. et al. (Ed.). **New Frontiers in Applied Data Mining**. [S.l.]: Springer Berlin Heidelberg, 2012. p.100–110. (Lecture Notes in Computer Science, v.7104).

SPITZER, F. **Principles of Random Walk**. [S.l.]: Springer New York, 2013. (Graduate Texts in Mathematics).

STROGATZ, S. H. Exploring complex networks. **Nature**, [S.l.], v.410, p.268–276, Mar. 2001.

TANG, J.; AGGARWAL, C.; LIU, H. Node Classification in Signed Social Networks. In: SIAM INTERNATIONAL CONFERENCE ON DATA MINING, 2016. **Proceedings...** [S.l.: s.n.], 2016. p.54–62.

TANG, J. et al. Temporal Distance Metrics for Social Network Analysis. In: ND ACM WORKSHOP ON ONLINE SOCIAL NETWORKS, 2., New York, NY, USA. **Proceedings...** ACM, 2009. p.31–36. (WOSN '09).

TASKAR, B.; ABBEEL, P.; KOLLER, D. Discriminative Probabilistic Models for Relational Data. In: EIGHTEENTH CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, San Francisco, CA, USA. **Proceedings...** Morgan Kaufmann Publishers Inc., 2002. p.485–492. (UAI'02).

TASKAR, B. et al. Link Prediction in Relational Data. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS) 16. **Anais...** Cambridge: MA: MIT Press, 2004.

TYLENDA, T.; ANGELOVA, R.; BEDATHUR, S. Towards Time-aware Link Prediction in Evolving Social Networks. In: WORKSHOP ON SOCIAL NETWORK MINING AND ANALYSIS, 3., New York, NY, USA. **Proceedings...** ACM, 2009. p.9:1–9:10. (SNA-KDD '09).

VALIPOUR, M.; BANIHABIB, M. E.; BEHBAHANI, S. M. R. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. **Journal of Hydrology**, [S.l.], v.476, p.433 – 441, 2013.

VARTAK, S. A survey on link prediction. **State University of New York, Binghamton, NY–13902, USA**, [S.l.], 2008.

WANG, C.; SATULURI, V.; PARTHASARATHY, S. Local Probabilistic Models for Link Prediction. In: DATA MINING, 2007. ICDM 2007. SEVENTH IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2007. p.322–331.

WANG, P. et al. Link prediction in social networks: the state-of-the-art. **Science China Information Sciences**, [S.l.], v.58, n.1, p.1–38, 2015.

WANG, S.; MANNING, C. D. Baselines and Bigrams: simple, good sentiment and topic classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SHORT PAPERS - VOLUME 2, 50., Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2012. p.90–94. (ACL '12).

WANG, Y. et al. A trust-based probabilistic recommendation model for social networks. **Journal of Network and Computer Applications**, [S.l.], v.55, p.59 – 67, 2015.

WASSERMAN, S.; FAUST, K. **Social network analysis**: methods and applications. [S.l.]: Cambridge University Press, 1994. v.506.

WATTS, D. J. Networks, Dynamics and the Small World Phenomenon. **AJS**, [S.l.], 1999.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. **nature**, [S.l.], v.393, n.6684, p.440–442, 1998.

WINTERS, P. R. Forecasting Sales by Exponentially Weighted Moving Averages. In: **Mathematical Models in Marketing**. [S.l.]: Springer Berlin Heidelberg, 1976. p.384–386. (Lecture Notes in Economics and Mathematical Systems, v.132).

XIANG, E. W. **A survey on link prediction models for social network data**. Acesso: 10-06-2015, `https://www.researchgate.net/publication/228343916_A_survey_on_link_prediction_models_for_social_network_data`.

XU, Z. et al. Infinite Hidden Relational Models. **ArXiv e-prints**, [S.l.], June 2012.

YANG, X.-S. Small-world networks in geophysics. **Geophysical Research Letters**, [S.l.], v.28, n.13, p.2549–2552, 2001.

YANG, Y.; LICHTENWALTER, R. N.; CHAWLA, N. V. Evaluating link prediction methods. **Knowledge and Information Systems**, [S.l.], v.45, n.3, p.751–782, 2015.

YIN, Z. et al. A Unified Framework for Link Recommendation Using Random Walks. In: ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM), 2010 INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2010. p.152–159.

YU, P. S.; HAN, J.; FALOUTSOS, C. **Link Mining**: models, algorithms, and applications. 1st.ed. [S.l.]: Springer Publishing Company, Incorporated, 2010.

ZANETTE, D. H. Dynamics of rumor propagation on small-world networks. **Phys. Rev. E**, [S.l.], v.65, p.041908, Mar 2002.

ZENG, Z. et al. A link prediction approach using semi-supervised learning in dynamic networks. In: SIXTH INTERNATIONAL CONFERENCE ON ADVANCED COMPUTATIONAL INTELLIGENCE (ICACI), 2013. **Anais...** [S.l.: s.n.], 2013. p.276–280.

ZHANG, Q.; LI, B. Discriminative K-SVD for dictionary learning in face recognition. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2010 IEEE CONFERENCE ON. **Anais...** [S.l.: s.n.], 2010. p.2691–2698.

ZHENG, Q.; SKILLICORN, D. B. Analysis of criminal social networks with typed and directed edges. In: INTELLIGENCE AND SECURITY INFORMATICS (ISI), 2015 IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2015. p.1–6.