Predição de Relacionamentos Baseada em Eventos Temporais

Paulo R. S. Soares , Ricardo B. C. Prudêncio[1]

[1]Centro de Informática, Universidade Federal de Pernambuco, Recife.

{prss,rbcp}@cin.ufpe.br

Abstract. *Link prediction deals with the occurrence of connections in a network. In this work, we propose a new proximity measure for link prediction based on the concept of temporal events. We defined a temporal event related to a pair of nodes according to the creation, maintenance or interruption of the relationship between the nodes in consecutive periods of time. We proposed an event-based score which is updated along time by rewarding the temporal events observed between the pair of nodes under analysis and their neighborhood. The assigned rewards depend on the type of temporal event observed (e.g., if a link is conserved, a positive reward is assigned). In the performed experiments, we evaluated the proposed event-based measure for link prediction in co-authorship networks. Promising results were observed when the proposed measure was compared to previous work developed in the literature of link prediction.*

Resumo. *Predição de relacionamentos lida com a ocorrência de conexões em uma rede. Nesse trabalho, propomos uma nova medida de proximidade para predição de relacionamentos baseada no conceito de eventos temporais. Definimos um evento temporal relacionado a um par de nós de acordo com a criação, manutenção ou interrupção do relacionamento entre os nós em intervalos consecutivos no tempo. Nós propomos uma medida baseada em eventos que é atualizada ao longo do tempo através de recompensa para os eventos temporais observados nos nós sob análise e nos seus vizinhos. As recompensas assinaladas dependem do tipo de evento observado (e.g., se um link é conservado, uma recompensa positiva é definida). Nos experimentos realizados, avaliamos a medida proposta para predição de relacionamentos em redes de coautoria. Resultados promissores foram obtidos quando a medida proposta foi comparada com métodos anteriores propostos na literatura de predição de relacionamentos.*

1. Introdução

À medida que a interação entre indivíduos aumenta em ambientes virtuais, mais dados de redes sociais se tornam disponíveis, servindo como base para diferentes estudos. Redes sociais são estruturas compostas por indivíduos ou entidades que podem estar conectados por diferentes formas de relacionamento (tais como amizade e colaboração). Análise de Redes Sociais (ARS) é um campo de estudo que tenta lidar com a complexidade existente nas redes sociais (Wasserman and Faust 1994). Diversas tarefas podem estar associadas à ARS. Neste trabalho, focamos no problema de predizer as conexões mais prováveis entre indivíduos baseado em estados prévios da rede. Esse um problema é conhecido na literatura como *predição de relacionamentos* (Xiang 2008; Hasan and Zaki 2011).

Boa parte dos trabalhos anteriores sobre predição de relacionamentos se baseia na aplicação de medidas de proximidade entre nós da rede no presente momento para predizer conexões em tempos futuros. As medidas de proximidade associadas aos pares de nós podem ser usadas para predição de relacionamentos em duas estratégias principais: (1) na abordagem não-supervisionada, em que uma medida pré-definida é utilizada para ordenar os pares de nós da rede e os mais bem ordenados são considerados como os links preditos; (2) na abordagem supervisionada, em que o problema de predição de relacionamentos é trato como um problema de classificação e as medidas de proximidade de nós são usadas como atributos preditores por um algoritmo de aprendizado. Uma limitação que pode ser apontada em diferentes trabalhos é que as medidas de proximidade são calculadas sem levar em consideração a evolução temporal da rede. As medidas de proximidade são computadas usando todos os dados da rede até o presente momento (i.e., o estado corrente completo da rede) sem considerar quando os links foram criados. Assim, uma fonte de informação potencialmente útil não é aproveitada adequadamente.

No presente trabalho, propomos uma nova medida de proximidade em que informação temporal é levada em consideração. Na medida proposta, nós utilizamos o conceito de eventos temporais que são atividades específicas (e.g., criação ou remoção de um link) observadas entre um par de nós em intervalos consecutivos de tempo. Por exemplo, um evento *inovativo* ocorre quando dois nós não estão conectados em um dado intervalo de tempo, mas um novo link é criado entre eles no próximo intervalo. Um valor de proximidade para um dado par de nós é computado monitorando ao longo do tempo eventos observados nos nós e nos seus vizinhos imediatos. Cada categoria de evento temporal definida no nosso trabalho (inovativo, conservativo e regressivo) é associada a uma recompensa numérica. A proximidade entre nós aumenta ou diminui ao longo do tempo dependendo dos eventos observados e respectivas recompensas. A força de uma conexão entre indivíduos é relacionada com sua frequencia de interação (Homans 1951), que foi modelada no nosso trabalho através de eventos temporais.

Para verificar a viabilidade da medida proposta, realizamos experimentos em diferentes redes de coautoria. A medida proposta foi avaliada em diferentes cenários considerando, por exemplo, diferentes valores de recompensa para os eventos temporais e adotando uma função de ponderação para dar maior importância a eventos mais recentes. Como base de comparação, experimentos foram realizados com medidas de proximidade tradicionais aplicadas de forma estática. Além disso, para uma comparação mais justa, realizamos experimentos com uma medida de proximidade baseada em séries temporais (Potgieter et al. 2009) que, assim como nossa medida, também considera informação temporal na predição de links. Em todos os experimentos, a estratégia de predição não-supervisionada foi adotada uma vez que as medidas de proximidade são calculadas. A medida AUC foi adotada como medida de desempenho dos métodos avaliados. O desempenho atingido pela medida baseada em eventos temporais superou o desempenho dos métodos de comparação em todas as redes consideradas nos experimentos.

A Seção 2 apresenta brevemente o problema de predição de relacionamentos, seguida pela Seção 3 que descreve nossa proposta em detalhe. A Seção 4 apresenta os experimentos e resultados obtidos. Finalmente, a Seção 5 conclui o trabalho com algumas considerações finais e trabalhos futuros.

2. Predição de Relacionamentos

Predição de relacionamentos consiste em predizer novas conexões ou detectar links ocultos em uma rede. Dentre as várias abordagens para tratar esse problema, as mais difundidas requerem o uso de medidas de proximidade entre pares de nós (Xiang 2008; Lu and Zhou 2011). A partir dos valores de proximidade de uma ou mais medidas diferentes, a predição de links pode ser feita de forma não-supervisionada (e.g., (Liben-Nowell and Kleinberg 2003; Lu and Zhou 2011)) ou supervisionada (e.g., (Hasan et al. 2006; de Sá and Prudêncio 2011)).

As diversas medidas de proximidade propostas na literatura podem ser categorizadas em: (1) medidas baseadas em vizinhança e (2) medidas baseadas em caminhos (Hasan and Zaki 2011). As medidas baseadas em vizinhança levam em consideração os vizinhos imediatos dos nós. Em geral, elas consideram que dois nós são mais prováveis de formar uma conexão se seus conjuntos de vizinhos têm uma interseção alta (Xiang 2008). Dentre essas medidas, podemos citar Common Neighbors (CN) (Newman 2001), Preferential Attachment (Newman 2001), Adamic/Adar (Adamic and Adar 2003) e o coeficiente de Jaccard (Salton and McGill 1986). As medidas baseadas em caminho, por sua vez, definem proximidade entre nós considerando caminhos existentes entre eles. A ideia básica é que dois nós tendem a se conectar se existem caminhos curtos entre eles. Essas medidas variam de medidas simples como a distância geodésica a definições mais sofisticadas que consideram combinações de caminhos de tamanho variável como a medida de Katz (Katz 1953). Em termos comparativos, as medidas baseadas em vizinhança são mais difundidas, devido tanto ao custo computacional mais baixo como bom desempenho em experimentos (Liben-Nowell and Kleinberg 2003; Huan 2006; Murata and Moriyasu 2008). A medida proposta no presente trabalho pode ser vista como uma medida baseada em vizinhança (como será visto na Seção 3).

A maioria das medidas de proximidade usadas para predição de links é aplicada de forma estática, i.e., nenhuma informação temporal é levada em consideração. Contudo, informação temporal, como os instantes passados de tempo em que dois nós interagiram ou quando uma conexão entre dois nós foi observada pela primeira vez, é um aspecto importante que deveria ser considerado (Hasan and Zaki 2011). Por exemplo, atividade recente entre vizinhos em comum de dois nós pode ser mais importante que atividade mais antiga. Alguns trabalhos anteriores podem ser mencionados a respeito do uso de informação temporal. Em (Tylenda et al. 2009), os autores modelam uma rede como um grafo com pesos, em que o peso de um link entre dois nós corresponde ao tempo da atividade mais recente observada entre os nós. A predição de links foi feita nesse trabalho usando as extensões das medidas de proximidade clássicas adequadas para redes com pesos (e.g., weighted Adamic Adar).

Uma abordagem alternativa para definir proximidade usando informação temporal é através de previsão de séries temporais (Huang and Lin 2009; Potgieter et al. 2009; Soares and Prudêncio 2012; Qiu et al. 2011). Em (Huang and Lin 2009), os autores construíram uma série temporal para cada par de nós, de forma que cada observação da série correspondia a frequencia de conexões entre os nós observados durante intervalos de tempo específicos. Previsões das séries produzidas por modelos ARIMA foram utilizadas como medidas de proximidade. Em (Potgieter et al. 2009; Soares and Prudêncio 2012), os autores adotaram ideias similares, no entanto, as séries temporais são construídas a par-

tir dos valores de uma medida de proximidade clássica previamente escolhida, aplicada a intervalos de tempo consecutivos na rede. Modelos de séries temporais são utilizados para produzir previsões para a proximidade de cada par de nós. Embora bons resultados possam ser obtidos com a abordagem de séries temporais, uma limitação dessa abordagem é a escolha adequada da medida de proximidade utilizada para construir as séries temporais, assim como a escolha do modelo de previsão, dentre a variedade de modelos que podem ser aplicados (Soares and Prudêncio 2012).

3. Predição Baseada em Eventos Temporais

No presente trabalho, propomos uma medida de proximidade que considera eventos temporais para pares de nós na rede. Essa medida combina duas ideias principais: (1) a conexão entre dois indivíduos é associada com a frequencia de interação entre eles (Homans 1951); (2) quanto maior o número de vizinhos em comum entre dois nós, maior a probabilidade de conexão (Newman 2001). A abordagem proposta atualiza o valor de proximidade entre nós dependendo dos eventos temporais observados entre eles e sua vizinhança. Um evento temporal, que será explicado melhor mais adiante, é definido a partir da aparição ou remoção de links entre os nós à medida que a rede evolui. Na abordagem proposta, inicialmente uma estrutura temporal é criada extraindo frames da rede em intervalos consecutivos de tempo. O valor de proximidade entre um par de nós é então computada agregando valores de recompensa assinalados aos eventos temporais observados durante a transição de frames.

3.1. Estrutura Temporal

Para construir uma estrutura temporal $\mathscr{N}$ para uma determinada rede, primeiramente esta deve ser dividida em uma série de *snapshots* ordenados no tempo, isto é, configurações da rede em diferentes momentos do passado. Posteriormente *snapshots* consecutivos são agrupados em *frames*. O tamanho de cada frame (número de *snapshots* que possui) é uniforme em $\mathscr{N}$ e indica o tamanho da *janela de predição*, isto é, até quantos passos no futuro deseja-se realizar a predição. Essa metodologia está detalhada logo abaixo.

Seja $G(V,A)$ um grafo representando uma rede social observada até o tempo T. Cada aresta em A é representada por uma tripla $\langle u,v,t\rangle$, indicando que os nós u e $v \in V$ possuíam um vínculo social no tempo t. Logo, pode haver mais de uma aresta em A referente a um determinado par de nós, desde que esses tenham interagido em diferentes momentos no passado. Considere Δ o tamanho da janela de predição. Portanto, a tarefa consiste em predizer novos relacionamentos entre os instantes $T+1$ a $T+\Delta$.

Seja G_t o subgrafo de G tal que seja composto apenas pelas arestas observadas no tempo t, ou seja, no t-ésimo *snapshot* temporalmente ordenado da rede. Seja ainda $[G_t, G_{t+1}, ..., G_{t+m}]$ o frame formado pela união disjunta dos grafos do instante t até $t+m$. Neste trabalho, um conjunto de n frames consecutivos $\mathscr{N} = \{F_1, \ldots, F_k, \ldots, F_n\}$ de tamanho Δ é extraído do grafo G. Logo, formalmente $\mathscr{N}$ pode ser definida como:

$$\begin{aligned}\mathscr{N} = \{&[G_{T-nw+1}, G_{T-nw+2}, \ldots, G_{T-(n-1)w}], \ldots,\\ &[G_{T-2w+1}, G_{T-2w+2}, \ldots, G_{T-w}],\\ &[G_{T-w+1}, G_{T-w+2}, \ldots, G_T]\}\end{aligned} \tag{1}$$

Logo, o frame $F_k = [G_{T-(n-k+1)\Delta+1}, ..., G_{T-(n-k)\Delta}]$ é o subgrafo de G que contém todos os links observados no k-ésimo intervalo de tempo de tamanho Δ. Por exemplo, considere uma rede observada até o ano $T = 2012$ e uma janela de predição de tamanho 2 (isto é, o objetivo é predizer novos links entre 2013 e 2014). Se $n = 3$ frames são extraídos da rede ($\mathcal{N} = \{F_1, F_2, F_3\}$), a seguinte estrutura será gerada:

$$\mathcal{N} = \{[G_{2007}, G_{2008}], [G_{2009}, G_{2010}], [G_{2011}, G_{2012}]\} \tag{2}$$

Essa estrutura seria usada para analisar como a rede evolui a cada 2 anos desde 2007. Na próxima seção, introduzimos o conceito de eventos temporais observados entre frames consecutivos na rede.

3.2. Eventos Temporais

Um evento indica o que aconteceu com o relacionamento entre dois nós (ainda estão conectados? não estão mais conectados? um link foi criado entre eles?) à medida que a rede evolui, ou seja, é a ação que leva um par de nós de um estado (conectado ou não conectado) para outro. Eventos podem ser categorizados em um de três tipos mutuamente exclusivos: *conservativo*, *inovativo* ou *regressivo*, definidos abaixo.

(a) Conservativo: Um evento conservativo ocorre quando um relacionamento entre dois nós não deixa de existir quando a rede evolui, isto é, quando dois nós possuem um vínculo em um frame e esse laço é preservado no próximo frame. Para cada par de nós $\langle u, v\rangle$, define-se um peso $\mathscr{C}(u, v, k)$ relacionado ao frame F_k, para levar em consideração eventos conservativos durante a transição entre o $(k-1)$-ésimo e o k-ésimo frame. Formalmente:

$$\mathscr{C}(u, v, k) = \begin{cases} c, \text{ se } \langle u, v\rangle \in E_{k-1} \cap E_k \\ 0, \text{ caso contrário} \end{cases} \tag{3}$$

Na equação, E_{k-1} e E_k são os conjuntos de arestas observadas nos frames F_{k-1} e F_k respectivamente. A constante c indica o peso relacionado a eventos conservativos, que deve assumir um valor não negativo já que a conexão entre os nós foi preservada.

(b) Inovativo: Eventos inovativos representam a criação de um novo link entre um par de nós em diferentes frames. Ocorre quando dois nós não estão conectados em um frame e um link é observado no próximo frame. O peso de um evento inovativo $\mathscr{I}(u, v, k)$ associado ao par $\langle u, v\rangle$ e ao frame F_k é:

$$\mathscr{I}(u, v, k) = \begin{cases} i, \text{ se } \langle u, v\rangle \in E_k \backslash E_{k-1} \\ 0, \text{ caso contrário} \end{cases} \tag{4}$$

A constante i na equação acima indica o peso dos eventos inovativos. Seu valor deve ser positivo, pois se entende que laço entre dois nós foi fortalecido.

(c) Regressivo: Eventos regressivos possuem ideia oposta aos inovativos. Representa a remoção de um link existente entre dois pares de nós durante a transição entre frames. O peso associado a este tipo de evento $\mathscr{R}(u, v, k)$ é definido da seguinte forma:

$$\mathscr{R}(u,v,k) = \begin{cases} r, \text{ se } \langle u,v\rangle \in E_{k-1} \backslash E_k \\ 0, \text{ caso contrário} \end{cases} \tag{5}$$

Neste evento, r deve assumir um valor negativo, pois conota perda de força no vínculo entre dois nós.

As constantes c, i e r podem ser determinadas de maneira empírica através da avaliação do desempenho em conjunto de validação. Neste trabalho, por simplicidade, o número de pesos foi reduzido a dois, tornando os valores de c e r proporcionais a i, ao qual foi atribuído o valor 1.0 fixo.

3.3. Proximidade Baseada em Eventos

A medida proposta combina: (1) as recompensas associadas a *eventos primários*, que são os eventos estritamente relacionados ao par de nós sob análise; e (2) as recompensas associadas a *eventos secundários*, que são aqueles que ocorrem entre um nó do par e a vizinhança do outro nó. A proximidade entre um dado par de nós (u,v) é definida como:

$$score(u,v) = \sum_{k=2}^{n} P(u,v,k) + \alpha.S(u,v,k) \tag{6}$$

$$P(u,v,k) = \mathscr{C}(u,v,k) + \mathscr{I}(u,v,k) + \mathscr{R}(u,v,k) \tag{6a}$$

$$S(u,v,k) = \sum_{x \in \Gamma(u) \cap \Gamma(v)} P(u,x,k) + P(v,x,k) \tag{6b}$$

Na Eq. 6a, $P(u,v,k)$ representa o peso de um evento (conservativo, inovativo ou regressivo) para um par de nós $\langle u,v\rangle$ observado na transição entre o frame $k-1$ e o frame k. Na Eq. 6b, $S(u,v,k)$ indica o peso agregado dos eventos secundários associados ao par $\langle u,v\rangle$, isto é, os eventos primários observados na vizinhança desses nós. Nessa equação, $\Gamma(x)$ é o conjunto de vizinhos do nó x na rede. Na Eq. 6, o parâmetro α é um fator de amortização que representa quão fortemente eventos secundários afetam o vínculo entre u e v. $P(u,v,k)$ e $S(u,v,k)$ associados ao primeiro frame (isto é, $k=1$) apresentam valores nulos, visto que a estrutura desse frame não foi gerada por nenhum conjunto de eventos, logo, não foram considerados na Eq. 6.

A Figura 1 ilustra o processo para cálculo da proximidade. Os nós 1 e 3 têm o nó 2 como vizinho em comum na rede (estrutura completa $G(V,A)$, antes da divisão em frames). Logo, o score para o par de nós $\langle 1,3\rangle$ será calculado em função dos eventos ocorridos entre eles (eventos primários) e aqueles que aconteceram nos pares $\langle 1,2\rangle$ e $\langle 2,3\rangle$ (eventos secundários) ao longo dos frames de $\mathscr{N}$.

Do frame F_1 ao frame F_2, percebe-se que um evento conservativo e um regressivo ocorreram nas díades $\langle 1,2\rangle$ e $\langle 2,3\rangle$ respectivamente. Até o momento, nenhum evento está associado ao par $\langle 1,3\rangle$. Portanto, até o frame F_2, o score parcial do par $\langle 1,3\rangle$ é dado pela combinação amortizada dos pesos referentes aos eventos secundários descritos acima, ou seja, $\alpha(c+r)$. Olhando para o próximo passo evolutivo (frame F_3), um evento regressivo está associado ao par $\langle 1,2\rangle$, enquanto um evento inovativo ocorreu no par $\langle 2,3\rangle$. Há também um evento inovativo relacionado ao par em análise $\langle 1,3\rangle$; sua ocorrência também

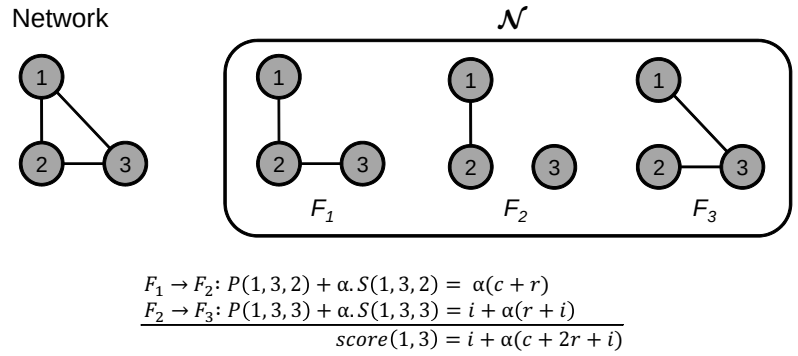

Figura. 1. Exemplo de cálculo do score baseado em evento. $\mathcal{N}$ representa a estrutura temporal da rede dividida em três frames (F_1, F_2 e F_3).

deve ser levada em consideração na computação do score final do par (o peso referente a esse evento, contudo, não deve ser amortizado, visto que se trata de um evento primário). Logo, o score resultante deste frame é dado por $i+\alpha(r+i)$. Por definição, os scores são cumulativos no tempo. O score final para um par de nós é formado pela agregação de todos seus scores parciais ao longo de $\mathcal{N}$. Portanto, isso resulta em um score final igual a $i+\alpha(c+2r+i)$ para o par $\langle 1,3\rangle$.

No nosso trabalho, propomos ainda uma medida de proximidade em que eventos mais recentes observados são considerados mais relevantes na predição. Para isso, uma nova medida (ver Eq. 7) foi derivada a partir da Eq. 6. Uma função monótona crescente $\beta : \mathbb{Z} \rightarrow \mathbb{R}$ foi adicionada a equação básica para ponderar mais fortemente eventos mais recentes. Por simplicidade, uma função logarítmica foi utilizada neste trabalho. Todavia, outras funções podem ser adotadas. Nota-se, que a equação básica é um caso particular dessa nova fórmula onde $\beta = 1$.

$$score(u,v) = \sum_{k=2}^{n} \beta(k).[P(u,v,k) + \alpha.S(u,v,k)] \tag{7}$$

$$\beta(k) = log(k) \tag{7a}$$

4. Experimentos e Resultados

Nessa seção, descrevemos os experimentos realizados para avaliar a abordagem proposta no trabalho assim como apresentamos os resultados obtidos.

4.1. Dados e Metodologia de Experimentos

Para avaliar o desempenho da predição de links, quatro redes de coautoria foram utilizadas nos experimentos. Nas redes de coautoria, um nó representa um autor e uma aresta indica uma colaboração, isto é, indica que dois nós específicos foram coautores de um artigo científico. Como visto, o método proposto faz uso de uma estrutura temporal para a predição de novos relacionamentos, por isso cada aresta armazena o tempo em que a conexão foi observada. Mais especificamente, nas redes adotadas aqui, cada aresta armazena o ano de publicação do artigo relacionado.

As redes utilizadas foram coletadas do arXiv[1] em quatro subáreas distintas: *i) astro-physics* (astro-ph), *ii) condensed matter* (cond-mat), *iii) high energy physics - lattice* (hep-lat) e *iv) theoretical high energy physics* (het-th). Colaborações do ano 1993 até 2000 foram usadas para astro-ph e cond-mat, e de 1992 até 2010, para hep-lat e hep-th. Informações sumarizadas sobre as redes estão expostas na Tabela 1.

Tabela. 1. Informações sumarizadas das redes.

	astro-ph	**cond-mat**	**hep-lat**	**hep-th**
Artigos	19.077	20.664	9.367	38.569
Autores	16.978	18.070	4.718	17.887
Colaborações	140.157	56.731	32.309	58.855

Para realizar as predições, uma janela de predição de tamanho um foi adotada na construção das estruturas temporais das redes, portanto o objetivo das análises realizadas foi investigar como a rede evoluiu a cada ano. Seguindo a metodologia descrita na Seção 3.1, os conjuntos de teste foram construídos da seguinte maneira: dados de 1993 até 1999 foram usados para construir $\mathscr{N}$ (com um número total de $n = 7$ frames) e o *snapshot* do ano 2000 foi usado como frame de predição para as redes astro-ph e cond-mat. Já para as redes hep-lat e hep-th, $\mathscr{N}$ foi construída a partir das colaborações do ano 1992 até 2009 (com um número total de $n = 18$ frames), enquanto o frame de predição foi composto pelas colaborações do ano 2010.

Nos nossos experimentos, os valores dos parâmetros de recompensa associados aos eventos foram definidos através de experimentos em um conjunto de validação. O último frame de $\mathscr{N}$ foi adotado como o conjunto de validação para seleção de parâmetros em cada rede: o frame de 1999 foi usado para validação nas redes astro-ph e cond-mat e o frame de 2009 foi usado para as redes hep-lat e hep-th. A Tabela 2 mostra a distribuição de classes de pares de nós (conectados ou não-conectados) para os conjuntos de validação e teste em cada rede. Como pode ser visto, a distribuição de classes é altamente desbalanceada, o que é comum em predição de links. Para minimizar o efeito deste desbalanceamento na avaliação, nos experimentos utilizados a medida AUC (Area Under ROC Curve) como medida de desempenho.

4.2. Métodos de Comparação

Nos experimentos, utilizamos inicialmente como base para comparação métricas tradicionais baseadas em vizinhança (PA, CN, AA e JC), largamente usadas na literatura. Realizamos também experimentos com a abordagem de séries temporais discutida na Seção 2. Inicialmente a rede é dividida em frames, usando o procedimento descrito na Seção 3.1. Uma série temporal é construída para cada par de nós sob análise aplicando uma dada medida de proximidade em cada frame da rede. Um modelo de previsão é então usado para prever o próximo valor da série. Essa previsão é definida como o valor de proximidade entre o par de nós. No nosso trabalho, adotados a medida AA e o modelo de Regressão Linear (RL) como método de previsão. Essa foi a melhor combinação de medida de proximidade e modelo de previsão avaliada em (Soares and Prudêncio 2012)

[1] arXiv.org e-Print archive - Cornell University Library

Tabela. 2. Distribuição de classes para os conjuntos de validação e teste.

(a) Validação

	astro-ph	**cond-mat**	**hep-lat**	**hep-th**
Nº de pares	745.683	149.769	330.137	498.782
Conectados (+)	7.797	1.737	1.099	1.232
Não-conectados (-)	737.886	148.032	329.038	497.550
Proporção ($^+/_-$)	1,06%	1,17%	0,33%	0,25%

(b) Teste

	astro-ph	**cond-mat**	**hep-lat**	**hep-th**
Nº de pares	1.337.607	254.651	373.053	553.544
Conectados (+)	9.791	2.665	608	1.626
Não-conectados (-)	1.327.816	251.986	372.445	551.918
Proporção ($^+/_-$)	0,74%	1,06%	0,24%	0,29%

que utilizaram os mesmos dados considerados no presente trabalho. Finalmente, destacamos que a medida proposta, assim como os métodos de referência, foram aplicados para predição não-supervisionada. Os pares de nós são ordenados diretamente usando os valores de proximidade e os mais bem ordenados são classificados como os links positivos.

4.3. Seleção de Parâmetros

Nessa seção, utilizamos o conjunto de validação para selecionar os melhores valores de recompensa. Como mencionado anteriormente na Seção 3.2, o parâmetro i foi fixado como 1.0 e os outros dois parâmetros c e r variaram de forma proporcional a i: $c = \{0.0, 0.25, 0.5, 1.0, 2.0\}$ e $r = \{-2.0, -1.0, -0.5, -0.25\}$. Consideramos um valor fixo para α de 0.05, definido através de uma bateria preliminar de experimentos.

A Tabela 3 mostra os valores de AUC obtidos no conjunto de validação para cada rede. Os quatro melhores valores observados não marcados em negrito (para facilitar a leitura, apresentamos os resultados apenas com duas casas decimais). Como pode ser visto, os melhores resultados foram obtidos por valores de c e r ao redor da mesma região para as quatro redes analisadas, o que sugere que as melhores combinações não variam muito de rede para rede. Embora melhores valores tenham sido gerados quando $c \leq i$, bons resultados também foram obtidos quando considerou-se pesos de eventos conservativos maiores que os de eventos inovativos. Por outro lado, se r for altamente ponderado ($r = -2.0$ e -1.0)), o método baseado em eventos não obtém bom desempenho, o que pode ser percebido através da leitura dos valores presentes na primeira coluna das tabelas. Como esperado, isso mostra que se o objetivo é predizer links em uma rede, então o histórico de eventos inovativos e conservativos são mais relevantes no processo preditivo.

A Tabela 4 mostra ainda que os valores de AUC obtidos para cada combinação de recompensas no conjunto de teste. Os resultados observados nos conjuntos de teste não variaram muito em relação ao conjunto de validação, o que indica que esses parâmetros podem ser determinados de maneira empírica com pouca perda de desempenho.

Tabela. 3. AUC para diferentes valores de recompensa - conjunto de validação.

(a) astro-ph						(b) cond-mat					
r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**	r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**
-2.0	0.34	0.35	0.36	0.40	0.46	**-2.0**	0.30	0.30	0.31	0.35	0.42
-1.0	0.63	0.66	0.67	0.69	0.69	**-1.0**	0.56	0.62	0.63	0.64	0.65
-0.5	0.79	0.79	0.80	0.80	0.79	**-0.5**	0.79	0.80	**0.81**	**0.81**	0.80
-0.25	0.79	**0.80**	**0.80**	**0.80**	**0.80**	**-0.25**	0.79	0.80	**0.81**	**0.81**	0.81
(c) hep-lat						(d) hep-th					
r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**	r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**
-2.0	0.37	0.45	0.58	0.68	0.72	**-2.0**	0.26	0.28	0.32	0.42	0.52
-1.0	0.80	0.83	0.82	0.82	0.81	**-1.0**	0.66	0.75	0.76	0.76	0.75
-0.5	**0.84**	**0.84**	**0.84**	**0.84**	0.83	**-0.5**	**0.85**	**0.85**	**0.85**	**0.85**	0.84
-0.25	0.83	0.83	0.83	0.83	0.83	**-0.25**	0.84	0.84	0.84	0.84	0.84

Tabela. 4. AUC para diferentes valores de recompensa - conjunto de teste.

(a) astro-ph						(b) cond-mat					
r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**	r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**
-2.0	0.34	0.36	0.39	0.44	0.51	**-2.0**	0.30	0.31	0.32	0.37	0.43
-1.0	0.64	0.67	0.68	0.6979	0.70	**-1.0**	0.57	0.63	0.64	0.6544	0.65
-0.5	0.78	0.79	**0.79**	0.79	0.78	**-0.5**	0.80	**0.81**	**0.81**	0.81	0.80
-0.25	0.78	**0.79**	**0.79**	**0.79**	0.79	**-0.25**	0.80	**0.81**	**0.81**	0.8123	0.80
(c) hep-lat						(d) hep-th					
r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**	r/c	**0.0**	**0.25**	**0.5**	**1.0**	**2.0**
-2.0	0.29	0.36	0.49	0.64	0.73	**-2.0**	0.22	0.23	0.27	0.35	0.46
-1.0	0.76	0.84	0.84	0.84	0.83	**-1.0**	0.66	0.73	0.73	0.73	0.72
-0.5	0.84	**0.85**	**0.85**	**0.85**	0.85	**-0.5**	**0.86**	**0.86**	**0.85**	0.85	0.84
-0.25	0.83	0.84	0.85	**0.85**	0.85	**-0.25**	0.85	**0.85**	0.85	0.85	0.84

4.4. Análise Comparativa

A Tabela 5 apresenta os valores de AUC obtidos pelas medidas propostas assim como pelos métodos de comparação. Nessa tabela, nos referimos a função definida para β para distinguir as medidas conforme equação usada para a agregação de recompensas. Os resultados mostraram que os métodos que utilizaram informação temporal (tanto os propostos como o método baseado em séries temporais) em geral foram melhores que a abordagem clássica estática utilizando PA, CN, AA e JC. Esses resultados ratificam que informação temporal é um aspecto importante a considerar para a predição de links. As medidas baseadas em eventos temporais apresentaram os melhores resultados dentre todos os métodos avaliados. O ganho de desempenho foi maior para as redes cond-mat, hep-lat e hep-th, mas menos expressivo para a rede astro-ph , no qual os resultados obtidos foram similares aos obtidos pela medida AA.

Nos resultados, observamos um ganho de desempenho em todas as redes quando a medida baseada em eventos usando $\beta = log$ foi comparada com a mesma medida uti-

lizando $\beta = 1$. Esse resultado dá apoio a expectativa que eventos mais recentes trazem maior informação sobre a emergência de links. Embora esse ganho de desempenho tenha sido em geral pequeno em valores absolutos, ele foi verificado estatisticamente usando um teste t com 95% de nível de confiança.

Tabela. 5. Desempenho dos Métodos (AUC).

	astro-ph	**cond-mat**	**hep-lat**	**hep-th**
Eventos ($\beta = log$)	0.7948	0.8269	0.8681	0.8830
Eventos ($\beta = 1$)	0.7929	0.8147	0.8573	0.8625
Séries Temporais (AA + RL)	0.7584	0.7402	0.8537	0.7899
AA	0.7844	0.7346	0.8049	0.7513
CN	0.7391	0.6744	0.7598	0.6838
JC	0.7299	0.6114	0.7408	0.6481
PA	0.5043	0.5455	0.5668	0.4834

Embora tanto o método de séries temporais como os métodos propostos explorem a natureza temporal do problema de predição, os métodos baseados em eventos obtiveram os melhores resultados (também confirmados estatisticamente). A abordagem proposta analisa a rede focando mais diretamente na formação de links. O método de séries temporais, por sua vez, assume que uma dada medida apresenta uma tendência ao longo do tempo. Assim, ele realiza uma análise indireta das conexões da rede. Na maior parte dos casos, a abordagem baseada em séries temporais superou os métodos estáticos, mas não foi capaz de superar os resultados baseados na análise dos eventos temporais.

5. Conclusão

Nesse trabalho, introduzimos uma nova medida de proximidade para predição de relacionamentos baseada em eventos temporais. Diferentes experimentos foram realizados em redes de coautoria. Inicialmente, avaliamos a robustez do método proposto em relação à escolha dos seus parâmetros. Mostramos resultados promissores em relação a métodos de predição anteriores da literatura. Adicionalmente, observamos que quando a idade dos eventos foi considerada, o método proposto apresentou melhores resultados. No nosso trabalho, adotamos uma função log para ponderar melhor eventos mais recentes. No entanto, outras funções (e.g., linear) podem ser consideradas em trabalhos futuros.

Destacamos que a medida baseada em eventos foi descrita aqui especificamente para redes não-direcionadas e sem pesos. Uma possível linha de pesquisa é estender as ideias discutidas aqui para outras categorias de redes mais complexas. Por exemplo, uma variedade de eventos temporais poderia ser definida se a direção do link fosse levada em consideração (e.g., um evento conservativo por ser observado em uma direção mas um evento regressivo pode ser observado na outra direção, indicando uma interação não recíproca entre os nós em um dado momento). No caso de redes com peso, eventos temporais poderiam considerar o aumento ou diminuição dos pesos entre os indivíduos ao longo do tempo, em vez de considerar apenas a existência dos links. Os valores de recompensa nesses casos devem ser definidos de forma cuidadosa a fim de refletir adequadamente a importância dos eventos no processo de predição.

Referências

[Adamic and Adar 2003] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

[de Sá and Prudêncio 2011] de Sá, H. R. and Prudêncio, R. B. C. (2011). Supervised link prediction in weighted networks. In *Proceedings of the 2011 International Joint Conference on Neural Networks*, pages 2281–2288. IEEE.

[Hasan et al. 2006] Hasan, M., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *Proceedings of SDM 06 Workshop on Link Analysis, Counterterrorism and Security*.

[Hasan and Zaki 2011] Hasan, M. and Zaki, J. (2011). A survey of link prediction in social networks. In Aggarwal, C., editor, *Social Network Data Analytics*, pages 243–275. Springer.

[Homans 1951] Homans, G. C. (1951). *The human group*. Routledge and Kegan, London.

[Huan 2006] Huan, Z. (2006). Link prediction based on graph topology: the predictive value of the generalized clustering coefficient. In *Proceedings of Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Huang and Lin 2009] Huang, Z. and Lin, D. K. J. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2):286–303.

[Katz 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

[Liben-Nowell and Kleinberg 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the 2003 International Conference on Information and Knowledge Management*, pages 556–559.

[Lu and Zhou 2011] Lu, L. and Zhou, T. (2011). Link prediction in complex networks: a survey. *Physica A*, 390(6):1150–1170.

[Murata and Moriyasu 2008] Murata, T. and Moriyasu, S. (2008). Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257.

[Newman 2001] Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64.

[Potgieter et al. 2009] Potgieter, A., April, K. A., Cooke, R. J. E., and Osunmakinde, I. O. (2009). Temporality in link prediction: understanding social complexity. *Journal of Emergence: Complexity and Organization*, 11(1):83–96.

[Qiu et al. 2011] Qiu, B., He, Q., and Yen, J. (2011). Evolution of node behavior in link prediction. In *AAAI*, pages 1810–1811.

[Salton and McGill 1986] Salton, G. and McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

[Soares and Prudêncio 2012] Soares, P. and Prudêncio, R. (2012). Time series based link prediction. In *Proc. of the Intern. Joint Conf. on Neural Networks*, pages 784–790.

[Tylenda et al. 2009] Tylenda, T., Angelova, R., and Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In *SNA-KDD '09*, pages 1–9.

[Wasserman and Faust 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.

[Xiang 2008] Xiang, E. W. (2008). *A survey on link prediction models for social network data*. PhD thesis, PhD Qualifying Exam, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.