Comparação de Técnicas de Predição de Links em Sub-redes de Coautoria Formada por Currículos da Plataforma Lattes

Douglas V. Santos[1], **Thiago C. Cunha**[1], **Antônio B. O. Silva**[2], **Fernando S. Parreiras**[1], **Orlando A. Gomes**[1]

[1]Laboratório de Sistemas de Informação Avançados (LAIS)
Universidade FUMEC – Belo Horizonte – MG

[2]IBGE – Instituto Brasileiro de Geografia e Estatística
Belo Horizonte – MG

orlando.gomes@fumec.br

Abstract. *The study of Lattes platform allows addressing and analyzing Brazil researchers network which could be useful for defining politics to improve science, technology, and innovation. This work evaluated Lattes Platform coauthorship network. This network evolves over time, which means that new coauthorships will arise in future. Therefore, using link prediction methods in this network would help to identify growing knowledge areas in Brazil. The used technics were Spectral Evolution, wich is new in this context, Common Neighbors, Adamic-Adar and Jaccard. The main goal was to evaluate the link prediction accuracy with different methods at the coauthorship network of Lattes Platform. The Spectral Evolution was worse than the others. Adamic-Adar method presented the best result – 817 times better than the random link prediction.*
Keywords*: Link Prediction, Co-authorship, Spectral Evolution*

Resumo. *O estudo da plataforma Lattes permite mapear e analisar a rede de pesquisadores no Brasil, o qual pode ser relevante para a adoção de políticas de incentivo ao progresso em ciência, tecnologia e inovação. Neste trabalho foi investigada a rede de coautoria da plataforma Lattes. Essa rede de coautoria evolui temporalmente, ou seja, novas colaborações entre pesquisadores surgem ao longo do tempo. Portanto, empregando-se técnicas de predição de links nessa rede, pode-se prever o crescimento de novas áreas de conhecimento no Brasil. As técnicas analisadas foram Evolução Espectral, uso inédito nesse contexto, Vizinhos Comuns, Adamic-Adar e Jaccard. O objetivo principal foi analisar e avaliar a eficácia desses métodos de predição de links na rede de coautoria da plataforma Lattes. A performance da Evolução Espectral foi inferior às outras técnicas. O melhor resultado obtido foi do método Adamic-Adar – 817 vezes superior à predição aleatória.*
Palavras-chave*: Predição de Links, Redes de Coautoria, Evolução Espectral.*

1. Introdução

A coautoria de artigos é uma das formas tangíveis e documentadas de colaboração científica. Vários aspectos da colaboração científica podem ser investigados por meio de análise de redes de coautoria, empregando-se a metodologia de análise de redes

sociais [Newman 2004, Barabási et al. 2002, Da Silva et al. 2012, Brandão et al. 2007, Parreiras et al. 2006, Liu et al. 2005]. Avaliar as redes de pesquisadores é uma atividade relevante para a adoção de políticas de incentivo ao progresso em ciência, tecnologia e inovação, pois permite o mapeamento e a mensuração das áreas de conhecimento no Brasil.

As redes de coautoria têm os vértices (nós) como os autores dos artigos analisados e as arestas (links) são as coautorias dos mesmos [Barabási and Albert 1999]. É natural supor que as redes evoluam no tempo. Considerando um instantâneo de uma rede de coautoria científica, por exemplo, a predição de links é uma ferramenta que possibilita, em um momento posterior a esse instantâneo, a recomendação de uma nova parceria entre cientistas dessa rede. A predição de links é desafiadora, pois a rede que surge desse processo, usualmente, é esparsa, gerando uma grande dificuldade em encontrar os mais prováveis candidatos a serem os próximos links da rede analisada [Getoor and Diehl 2005].

Neste contexto, emerge a seguinte questão de pesquisa: Qual é a precisão das técnicas existentes de predição de link na rede de coautoria formada pelos currículos da plataforma Lattes?

A rede escolhida para este trabalho foi construída a partir dos dados estruturados dos currículos dos pesquisadores da Plataforma Lattes.[1] Nesses currículos, encontram-se informações que foram usadas para a construção das redes de coautoria. Trata-se de uma rede composta por pesquisadores e o problema de pesquisa é avaliar qual técnica de predição de novos links tem a melhor precisão na rede de coautoria. Para tal, foram desenvolvidos softwares para realizar o cálculo da predição em que cada um se baseia na técnica escolhida. As técnicas avaliadas foram: Evolução Espectral [Kunegis et al. 2010], Vizinhos Comuns [Newman 2001], Adamic-Adar [Adamic and Adar 2003] e Jaccard [Liben-Nowell and Kleinberg 2007]. É importante ressaltar que a técnica de Evolução Espectral [Kunegis et al. 2010] ainda não foi utilizada para realizar predições de links na plataforma Lattes. As técnicas de predição de links investigadas nesse trabalho utilizam como informação para a previsão de um novo link entre dois autores somente a configuração da rede de coautoria no passado. Colaborações em projetos de pesquisa, orientações, participação em bancas, dentre outros, podem influenciar na formação de uma nova coautoria. Enfatiza-se que, ao desconsiderar estas informações, procura-se evidenciar novas coautorias que não são facilmente previsíveis por essas informações dos autores. Por exemplo, espera-se que orientador e orientado sejam coautores em uma publicação no futuro.

Este trabalho está organizado da seguinte forma: A Seção 2 apresenta a fundamentação teórica e uma análise dos trabalhos relacionados. A Seção 3 descreve a comparação entre as técnicas de predição de links. A análise dos resultados é descrita na Seção 4. A Seção 5 apresenta as conclusões.

[1] http://lattes.cnpq.br

2. Fundamentação Teórica

Redes estão em todos os lugares. Os participantes das redes são os seus "nós" ou "vértices". No caso das redes sociais, os participantes são, normalmente, denominados atores e podem ser individuais (uma pessoa) ou coletivos (uma organização ou coletividade). O laço relacional, ou simplesmente laço ou link, é responsável por estabelecer a ligação entre pares de atores. No caso das redes de coautoria, o link se dá pela existência de um trabalho científico produzido pelos dois autores.

Em seu trabalho, Liben-Nowell e Kleinberg (2007) apresentaram uma análise do desempenho de técnicas de predição de links. Ao avaliar a técnica de predição aleatória, que foi utilizada nesse trabalho como referência, observou-se uma assertividade muito pequena. Uma maneira de realizar esse experimento é considerar uma rede fictícia com o mesmo número de nós de uma rede real. Por uma simulação computacional, é possível prever onde surgirá o próximo link. Possuindo o resultado da simulação, verifica-se a precisão da técnica. O desempenho ficou abaixo de 0,5%, ou seja, 99,5% dos links aleatórios não corresponderam aos links reais. Liben-Nowell e Kleinberg (2007) testaram outras técnicas para avaliar o desempenho na predição de links considerando redes de coautoria. Dentre elas, encontram-se as técnicas baseadas no conjunto de todos os caminhos, como uma medida para a soma dos caminhos entre dois nós e, exponencialmente, exibe a medida ponderada pelo número de caminhos mais curtos para interligá-los. Essa medida exponencial indicaria entre quais nós existe a maior probabilidade de surgir uma conexão [Liben-Nowell and Kleinberg 2007].

Newman (2001) identificou uma relação entre a quantidade de conexões em comum entre dois nós quaisquer e a probabilidade de haver uma conexão entre esses dois nós. A métrica acabou denominada "Vizinhos Comuns" (CN, Common Neighboors)[Newman 2001]. Segundo Barabási *et al.* (2002), a seleção de nós seria guiada pela anexação preferencial. Esse princípio afirma que a possibilidade de conexões é maior para nós que possuam um grau elevado [Barabási et al. 2002]. Segundo Perez-Cervantes *et al.* (2013), o coeficiente de Jaccard é uma normalização dos Vizinhos Comuns [Perez-Cervantes et al. 2013]. Liben-Nowell e Kleinberg (2007) afirmam que essa é uma técnica utilizada como métrica de similaridade na ciência da informação, desde sua definição em 1983 [Liben-Nowell and Kleinberg 2007]. O coeficiente é calculado dividindo a interseção dos dois conjuntos pela união dos mesmos [Digiampietri et al. 2015]. Adamic e Adar (2003) definiram uma métrica para calcular a similaridade de páginas na internet [Adamic and Adar 2003]. Posteriormente, ela passou a ser utilizada no estudo de redes sociais. Ela é referenciada em trabalhos como índice Adamic-Adar (*AA*), como visto em Lü e Zhou (2011) e Liben-Nowell e Kleinberg (2007) [Lü and Zhou 2011, Liben-Nowell and Kleinberg 2007].

Na predição de links, o grau de separação entre os nós é um dificultador [Papadimitriou et al. 2011]. Quando o grau de separação é maior que dois, a tarefa de se prever links torna-se complexa. Métodos de predição baseados em vizinhança comum utilizam a característica de interligação dos nós, o que pode negligenciar as conexões de grau maior que dois.

Geralmente, as redes crescem pela adição de nós e interconexões à sua estrutura ao longo do tempo. Para Kunegis *et al.* (2010), a rede, no início de sua construção, evolui em várias direções e se fossem traduzir a evolução em espaços vetoriais, esses estariam representando bases de subespaços. Quando a rede torna-se complexa, contendo milhares ou, até mesmo, milhões de nós e links, o subespaço vetorial representando a direção do crescimento da rede permanece aproximadamente constante. Na decomposição das matrizes de adjacência que representam a rede objeto de análise em cada momento do tempo, nota-se que os autovalores variam enquanto os autovetores permanecem aproximadamente constantes [Kunegis et al. 2010].

O desempenho de técnicas de predição de links já foram avaliados em redes de coautoria brasileiras por vários pesquisadores. Brandão e Moro (2012) estudaram uma rede com 169 nós extraída da Ciência Brasil.[2] A avaliação da técnica denominada Affin é baseada no princípio da homofilia. O princípio da homofilia, segundo os autores, procura pesos nas características externas à rede e como essas influenciam o surgimento de colaborações. As características consideradas foram afiliação institucional e proximidade social. O acerto na previsão do surgimento de colaborações atingiu 2% [Brandão and Moro 2012]. Digiampietri *et al.* (2012) asseveram que os currículos da Plataforma Lattes são uma fonte de informação para criação e análise de redes sociais de pesquisadores.

Mena-Chalco, Junior e Marcondes (2009) desenvolveram uma ferramenta de exploração de dados dos currículos da Plataforma Lattes: o Script Lattes. Essa ferramenta, desenvolvida em linguagem de programação Python, oferece uma gama de recursos para obter informações consolidadas dos currículos [Mena-Chalco et al. 2009]. Luna, Revoredo e Cozman (2013) aplicam as técnicas de predição de links avaliadas por Liben-Nowell e Kleinberg (2003) em sub-redes originadas dos currículos cadastrados na Plataforma Lattes. Esses autores consideraram informações ontológicas na predição de links. Por exemplo, orientador e orientado, em um futuro próximo, têm grandes chances de publicar um trabalho em colaboração, o qual é denominado de ontologia probabilística. O desempenho de acerto das conexões chegou a 93,9% [Luna et al. 2013, Liben-Nowell and Kleinberg 2007].

Em uma rede construída a partir da Plataforma Lattes, Perez-Cervantes *et al.* (2013) identificaram os nós influentes no surgimento de colaborações utilizando as técnicas de predição de links avaliadas por Liben-Nowell e Kleinberg (2003) [Perez-Cervantes et al. 2013, Liben-Nowell and Kleinberg 2007]. Digiampietri, Santiago e Alves (2012) analisaram 657 pesquisadores cadastrados na Plataforma Lattes. Por meio do que chamaram de análise de classes, consideraram atributos comuns nos currículos para proporem futuras colaborações. Uma ressalva é que o método não conseguiu prever colaborações inéditas, aquelas que demostram uma produção de trabalho em conjunto pela primeira vez; as colaborações futuras foram indicadas entre colaboradores já parceiros na produção de trabalhos. A taxa de acerto foi superior a

[2] http://pbct.inweb.org.br/cienciabrasil/

94% [Digiampietri et al. 2012].

Maruyama e Digiampietri (2016) apresentaram um método de predição de links da plataforma Lattes utilizando-se de técnicas de inteligência artificial [Maruyama and Digiampietri 2016]. Os resultados mostraram que a acurácia obtida não foi superior à determinada para o caso no qual não há novos links na predição. Zhang (2017) realizou a predição de links em uma rede de coautoria que foi obtida no site "Web of Science" da Thomson Reuters. A técnica utilizada baseia-se em preditores com múltiplas relações no escopo de machine learning. Os melhores resultados apresentam uma precisão de 42,5% [Zhang 2017].

Neste trabalho, subconjuntos da rede de coautoria da plataforma Lattes serão analisados pela técnica de predição de links denominada de Evolução Espectral a qual é comparada com a predição de links aleatória e as técnicas Vizinhos Comuns, Adamic-Adar e Jaccard, bem como com trabalhos relacionados. É importante ressaltar que a predição de links por meio da Evolução Espectral ainda não foi utilizada na rede de coautoria da plataforma Lattes.

3. Comparação de técnicas de predição de links

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) é o detentor da Plataforma Lattes que é destinada ao registro e à disponibilização dos dados sobre progresso acadêmico dos pesquisadores de todas áreas do conhecimento, sejam esses doutores, mestres, graduados, ou técnicos de nível médio.[3] O acesso aos dados acadêmicos dos pesquisadores cadastrados na Plataforma Lattes é livre. O ID Lattes, que é composto por 16 dígitos numéricos e pode ser utilizado para isolar o acesso a um único currículo na plataforma Lattes, é um facilitador para buscar os dados no sítio da Plataforma[4] [Mena-Chalco et al. 2014].

O número de currículos na Plataforma Lattes é da ordem de milhões, sendo que a maioria é composta por currículos incompletos ou de pesquisadores que nunca escreveram um artigo científico em modo de colaboração. Para filtrar esses currículos, grupos de pesquisas foram selecionados considerando pesquisadores cadastrados que apresentam o título de doutor em seu currículo. A máquina de busca da Plataforma Lattes possui sistemas de filtros que permitem isolar um grupo de pesquisadores para ser retornado como resultado de uma pesquisa. O primeiro grupo de pesquisadores a ser retornado foi o de Bolsistas de Produtividade do CNPq. Os outros dois grupos seguiram o mesmo critério, trocando-se as características do filtro relacionado à área da ciência. O segundo grupo foi o de doutores que definiram sua área do conhecimento como Engenharias. O terceiro grupo foi o de doutores que definiram sua área do conhecimento como Ciências Exatas e da Terra. As informações sobre o número de nós (N) e links (L) de cada grupo até o ano de 2013, bem como o número de links gerados em 2013 (ΔL), encontram-se na Tabela 1. O objetivo do isolamento das áreas é presumir que, de acordo com Parreiras *et al.* (2006), esses grupos possam ser as sub-redes em que a ocorrência de produções científicas em colaborações seja de grau elevado em relação ao âmbito da rede

[3] `http://cnpq.br`

[4] `http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar`

[Parreiras et al. 2006].

Tabela 1. *Caracterização das sub-redes Lattes*

Subconjunto Lattes	N	L	ΔL
CNPq	13.790	140.014	10.955
Engenharias	21.619	70.421	1.521
Exatas	31.646	193.882	6.230

N = Número de nós até 2013, L = Número de links até 2013 e ΔL = Número de links gerados em 2013.

Os arquivos dos currículos dos pesquisadores da Plataforma Lattes foram gerados em formato simples para extração de dados dos mesmos, o XML (eXtensible Markup Language). Os currículos contêm todas as informações referentes ao progresso acadêmico dos pesquisadores e as descrições dos artigos científicos produzidos em conjunto. As informações estão todas estruturadas, o que facilitou a tarefa de construção da rede de coautoria por meio da linguagem de programação Python.

3.1. A construção das matrizes de adjacência das redes de Coautoria

Possuindo os currículos agrupados, parte-se para a construção das redes. Os nós das redes são os donos dos currículos, cada um com sua identificação definida pelo ID Lattes. As interligações entre os nós são as indicações da existência da coautoria na produção de artigos científicos publicados em jornais, revistas ou anais de congressos da comunidade científica. Apesar de os currículos estarem estruturados em formato XML, é necessário limpar os dados e encontrar a estratégia para definir como a coautoria pode estar ocorrendo. O fluxograma que serviu de base para a lógica de programação encontra-se na Figura 1.

O programa tem início realizando a leitura da seleção de currículos XML. Ao ler os currículos, são filtrados os nomes dos autores que representam os donos dos currículos. Esse filtro é utilizado para que seja aplicada uma "limpeza" nas palavras que compõem os nomes (retiram-se acentos, vírgulas e caracteres especiais). Estando limpos, os nomes são organizados em uma lista e os IDs Lattes de cada um são localizados nos currículos. O critério de selecionar os currículos de doutores foi aplicado. Os currículos válidos são submetidos a uma filtragem de seus IDs Lattes e as informações sobre os artigos publicados analisadas.

A primeira etapa da análise das informações dos artigos busca verificar se o ano de publicação corresponde ao intervalo definido para obter a matriz de adjacência. Estando no intervalo, é necessário saber quais são os autores e os coautores dos artigos. A informação do autor foi obtida, pois esse é o dono do currículo. Todas as informações dos nomes dos autores, coautores e IDs Lattes são carregados em uma lista. Por fim, chega-se ao momento em que as informações das listas obtidas podem ser cruzadas para gerar as matrizes de adjacência. Os nomes em referências bibliográficas são conhecidos assim como, os IDs Lattes e as relações de coautoria. O pesquisador, ao cadastrar o currículo na

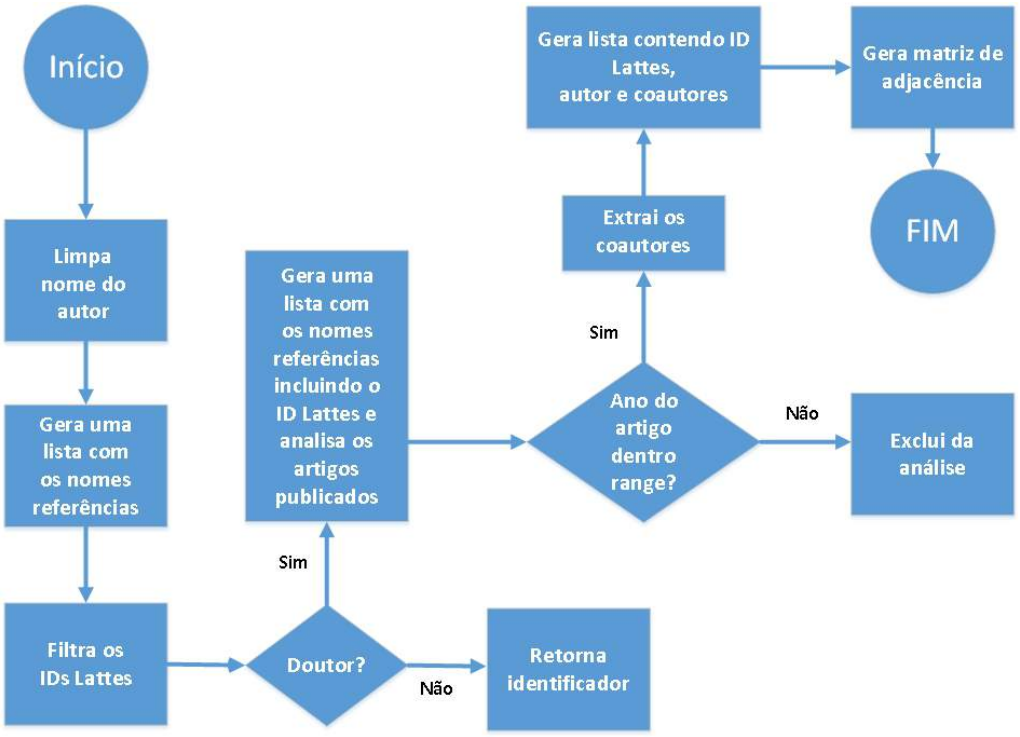

Figura 1. Fluxograma do programa gerador das Matrizes de Adjacência

Plataforma Lattes, pode escolher a maneira conveniente para ser nomeado em suas referências bibliográficas. Isso gera dificuldade na filtragem dos dados devido à existência de homônimos em que um autor pode escolher mais de uma maneira para ser citado. O programa gerador das matrizes minimizou esse problema considerando os nomes válidos, se esses forem representados nos currículos dos pesquisadores em formato de permuta. Se "A" referenciou "B", "B" tem que ter referenciado "A", o que minimiza os erros. Foram considerados os artigos publicados. As matrizes geradas não consideram o peso das ligações. Em todas as matrizes, o número que significa uma colaboração entre dois autores é o número "1", que indica uma evolução no surgimento de colaborações inéditas entre os autores. Se "A" colaborou com "B" em 2011 e voltou a colaborar com esse em 2012, em ambas as matrizes o número correspondente será o "1". Paralelamente, se "A" não colaborou com "C" em 2011, mas colaborou com esse em 2012, os números correspondentes nas matrizes serão, respectivamente, "0" e "1".

3.2. Predição de Links

Quatro técnicas de predição de links foram escolhidas, conforme descrito na *Introdução*. As matrizes de adjacências geradas foram analisadas obedecendo a característica de cada método. É importante esclarecer que para realizar os cálculos da técnica de Evolução Espectral utilizou-se do software Octave[5] que, por se tratar do contexto de álgebra linear, realiza as operações matriciais. Portanto, não se utilizou uma lista encadeada ou tabela hash para os dados da rede de coautoria, porque seria inviável o processamento algébrico por essa técnica.

[5]https://www.gnu.org/software/octave/

3.2.1. A teoria da Evolução Espectral

Considere uma linha temporal subdividida em três momentos ($t_1 < t_2 < t_3$), onde a distância temporal entre instantes de tempo é de um ano. Para cada momento, uma matriz correspondente foi mapeada e nomeada uma matriz de adjacência A_1, A_2 e A_3, respectivamente. U_2 representa a matriz resultante da decomposição de A_2 em seus autovetores. Para prever a matriz de adjacência A_3, deve-se calcular a matriz dos autovalores Λ_3 e conservar os autovetores da matriz U_2:

$$A_3 = U_2 \cdot \Lambda_3 \cdot U_2^t$$

.

Para se determinar a matriz de autovalores Λ_3, Kunegis *et al.* (2010) sugerem utilizar uma equação linear (regressão linear): .

$$\Lambda_3 = 2\Lambda_2 - \Lambda_1,$$

onde os valores de Λ_1 e Λ_2 são obtidos pelas decomposições das matrizes de adjacência A_1 e de $A2$ em seus autovalores e autovetores, respectivamente [Kunegis et al. 2010].

3.2.2. Vizinhos Comuns

Para dois nós x e y, obtém-se o número de vizinhos comuns a partir do número de nós do conjunto resultante da intersecção dos conjuntos $\Gamma\ (x)$ e $\Gamma\ (y)$. Por representar uma métrica não normalizada, essa técnica geralmente reflete a similaridade relativa entre pares de nós. VC (Vizinhos Comuns) pode ser representada como

$$VC(x, y) =\mid \Gamma\ (x) \cap \Gamma\ (y)\ \mid,$$

onde $\Gamma\ (x)$ é o conjunto de vizinhos do nó x e $\Gamma\ (y)$ é o conjunto de vizinhos do nó y. Fazendo-se o produto dos elementos da matriz de adjacência correspondentes da linha x pela linha y, determina-se o número de vizinhos comuns dos nós x e ym[Liben-Nowell and Kleinberg 2007].

3.2.3. Adamic-Adar

Adamic e Adar (2003) definiram uma técnica para calcular a similaridade de páginas na Internet. Posteriormente, ela passou a ser utilizada no estudo de redes sociais (WANG *et al.*, 2015). Ela é referenciada em trabalhos como índice Adamic-Adar (AA), como visto em Lü e Zhou (2011) e Liben-Nowell e Kleinberg (2007). AA supõe que características semelhantes são mais relevantes para a similaridade. Para isso, considera em seu cálculo a frequência de z que corresponde à ocorrência da característica comum (intersecção) entre os conjuntos $\Gamma\ (x)$ e $\Gamma\ (y)$ de acordo com a equação

$$AA(x, y) = \sum_{z \in \Gamma\ (x) \cap \Gamma\ (y)} \frac{1}{\log \mid \Gamma\ (z)\ \mid}.$$

3.2.4. Coeficiente de Jaccard

Segundo Perez-Cervantes *et al.* (2013), o Coeficiente de Jaccard (JC) é uma normalização dos Vizinhos Comuns [Perez-Cervantes et al. 2013]. Liben-Nowell e Kleinberg (2007) afirmam que essa é uma técnica utilizada como métrica de similaridade na ciência da informação, desde sua definição em 1983 [Liben-Nowell and Kleinberg 2007]. JC assume os valores elevados para os pares de nós que compartilham uma maior proporção de vizinhos comuns em relação ao número de vizinhos que têm. O coeficiente é calculado dividindo o número de nós da interseção dos dois conjuntos pelo número de nós da união dos mesmos [Digiampietri et al. 2015], de acordo com a fórmula

$$JC(x,y) = \frac{\mid \Gamma\,(x) \cap \Gamma\,(y) \mid}{\mid \Gamma\,(x) \cup \Gamma\,(y) \mid}.$$

4. Resultados e Discussões

O percentual de acerto para todas as técnicas foi determinado utilizando-se o conceito de precisão

$$P(\%) = \frac{V_p}{V_p + F_p} \times 100,$$

onde V_p são os verdadeiros positivos – links previstos corretamente pela técnica e F_P são os falsos positivos – links previstos pela técnica incorretamente.

A Tabela 2 apresenta os resultados para cada algoritmo utilizado nesse trabalho. Os resultados da evolução espectral (*EE*), que é o algoritmo inédito investigado neste trabalho, precisam ser elucidados. Para a sub-rede da grande área de Engenharias, a *EE* não acertou nenhum dos links previstos para o ano de 2013. Já para a grande área de Exatas, a capacidade de memória do computador utilizado (24 GB) não foi suficiente. A *EE* apresentou um desempenho inferior aos outros métodos ($3,56\%$). O método Adamic-Adar foi o que atingiu o melhor desempenho que foi 817 vezes superior ao processo de predição de links aleatório.

De acordo com os resultados da Tabela 2, o desempenho de todos os métodos é baixo se comparado com os resultados apresentados por Digiampietri *et al*. (2012), cuja acurácia foi de 94% . Porém, é importante ressaltar que neste trabalho utilizou-se a medida de precisão; caso utilizasse a medida de acurácia, os resultados seriam em torno de 99% para qualquer conjunto investigado nesse trabalho.

Zhang (2017), utilizando uma técnica de machine learning com preditores, conseguiu a precisão de 42,5% no melhor dos resultados. Entretanto, é importante considerar que a base de dados analisada não é a plataforma Lattes. Em seus resultados, a técnica de Vizinhos Comuns obteve 24,4% em determinado contexto. Essa mesma técnica, nas sub-redes da plataforma Lattes analisados neste artigo, obteve o valor de 5,3%. Portanto, a mudança da base de dados pode mudar os resultados substancialmente.

5. Conclusão

A técnica da evolução espectral, que foi utilizada de forma inédita em redes extraídas da plataforma Lattes nesse trabalho, apresentou um desempenho inferior aos métodos

Tabela 2. ***Precisão de acerto dos links previstos em 2013 para cada algoritmo.***

Subconjunto Lattes	*AL(%)*	*EE(%)*	*JC(%)*	*AA(%)*	*VC(%)*
CNPq	0,0079	3,5641	5,0299	6,4538	5,3094
Engenharias	0,0016	0	3,7581	5,0875	4,2579
Exatas	0,0014	–	3,1647	4,2567	3,2538

AL = Aleatória, EE = Evolução Espectral, JC = Jaccard, AA = Adamic-Adar e VC = Vizinhos Comuns.

considerados clássicos pela literatura. O melhor desempenho entre todas as técnicas foi de 6,45% pelo método Adamic-Adar – 817 vezes superior ao método aleatório.

Para trabalhos futuros sugere-se um foco nas sub-redes da Plataforma Lattes, procurando agrupar os pesquisadores pela área de formação e não pela grande área da ciência. É importante ressaltar que este trabalho teve o foco na predição de links para colaborações inéditas e não considera quaisquer outras características dos nós das redes objeto de estudo (por exemplo, filiação, relação entre orientador e orientando, dentre outras).

Uma outra oportunidade futura encontra-se no uso de uma matriz de adjacência com pesos para verificar se há melhora na predição de links pela técnica de Evolução Espectral e demais técnicas. A ideia de explorar os pesos já possui resultados interessantes na literatura [Sett et al. 2016].

Referências

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

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614.

Brandão, M. A. and Moro, M. M. (2012). Recomendação de colaboração em redes sociais acadêmicas baseada na afiliação dos pesquisadores. In *SBBD (Short Papers)*, pages 73–80.

Brandão, W. C., Parreiras, F. S., and Silva, A. B. d. O. (2007). Redes em ciência da informação: evidências comportamentais dos pesquisadores e tendências evolutivas das redes de coautoria. *Informação & Informação*.

Da Silva, A. K. A., Barbosa, R. R., and Duarte, E. N. (2012). Rede social de coautoria em ciência da informação: estudo sobre a área temática de organização e representação do conhecimento. *Informação & Sociedade*, 22(2).

Digiampietri, L., Maruyama, W. T., Santiago, C., and da Silva Lima, J. J. (2015). Um sistema de predição de relacionamentos em redes sociais. In *Brazilian Symposium on Information Systems*, volume 11.

Digiampietri, L., Mena-Chalco, J., de Jésus Pérez-Alcázar, J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012). Minerando e caracterizando dados de currículos lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

Kunegis, J., Fay, D., and Bauckhage, C. (2010). Network growth and the spectral evolution model. In *Proceedings of the 19th ACM International Conference on Information and knowledge Management*, pages 739–748. ACM.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7):1019–1031.

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170.

Luna, J. E. O., Revoredo, K., and Cozman, F. G. (2013). Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, 19(4):397–409.

Maruyama, W. and Digiampietri, L. (2016). Co-authorship prediction in academic social network.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., and Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, 65(7):1424–1445.

Mena-Chalco, J. P., Junior, C., and Marcondes, R. (2009). Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205.

Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y. (2011). Friendlink: link prediction in social networks via bounded local path traversal. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on Networks*, pages 66–71. IEEE.

Parreiras, F. S., Silva, A. d. O., Matheus, R. F., Brandão, W. C., et al. (2006). Redeci: colaboração e produção científica em ciência da informação no brasil. *Perspectivas em ciência da Informação*, 11(3):302–317.

Perez-Cervantes, E., Mena-Chalco, J. P., De Oliveira, M. C. F., and Cesar, R. M. (2013). Using link prediction to estimate the collaborative influence of researchers. In *eScience (eScience), 2013 IEEE 9th International Conference on*, pages 293–300. IEEE.

Sett, N., Singh, S. R., and Nandi, S. (2016). Influence of edge weight on node proximity based link prediction methods: an empirical analysis. *Neurocomputing*, 172:71–83.

Zhang, J. (2017). Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. *Information Processing & Management*, 53(1):42–51.