Revista Brasileira de Computação Aplicada, Julho, 2021

DOI: 10.5335/rbca.v13i2.12493
Vol. 13, Nº 2, pp. 48–57
Homepage: seer.upf.br/index.php/rbca/index

RBCA
ISSN 2176-6649

ORIGINAL PAPER

Combinando agrupamento e classificação para a predição de coautorias na Plataforma Lattes

Combining grouping and classification to predict co-authorship in the Lattes Platform

William Takahiro Maruyama [1] and Luciano Antonio Digiampietri [1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

*williammaruyama@usp.br; digiampietri@usp.br

Recebido: 16/04/2021. **Revisado:** 15/06/2021. **Aceito:** 04/07/2021.

Resumo

As Redes Sociais Online desempenham um papel importante na sociedade moderna, são um modelo e um reflexo das redes sociais do mundo real. Com as informações disponíveis na Plataforma Lattes é possível construir uma rede social acadêmica, na qual as relações entre os pesquisadores representam, por exemplo, uma parceria na produção de uma publicação. A tarefa de predição de relacionamentos (ou de links) para identificar possíveis colaboradores é uma tarefa complexa que pode favorecer a comunicação entre os usuários. O objetivo deste trabalho é propor a utilização da técnica de agrupamento e a inclusão de novos atributos que usam informações de comunidade para melhorar a previsão de relações de coautoria nas redes sociais acadêmicas.

Palavras-Chave: Redes sociais acadêmicas; Agrupamento; Redes de coautoria; Predição de Links; Redes Sociais

Abstract

Online Social Networks play an important role in modern society. They are a model and a reflection of social networks from the real world. With the information available on the Lattes Platform, it is possible to build academic social networks in which relationships between researchers represent, for example, a partnership in the production of a publication. The link prediction task to identify potential collaborations is a complex activity that can favor communication among users. The objective of this work is to propose the use of a clustering technique and the inclusion of new attributes that use community information to improve the prediction of co-authorship relationships in academic social networks.

Keywords: Academic social network; Clustering; Co-authorship networks; Link Prediction; Social Networks

1 Introdução

Nas popularizadas plataformas de Redes Sociais Online, pessoas de diferentes localidades podem se expressar e se comunicar via Internet. Essas plataformas projetam as interações da nossa sociedade atual e permitem que as pessoas rompam os limites físicos. Com o passar dos anos houve um crescimento na quantidade de sistemas e um aumento na disponibilidade de dados que podem ser representados por uma rede. Assim, surgiram novos desafios no campo da mineração de dados, especialmente em rede sociais.

Os relacionamentos ou interações encontradas nas redes sociais podem ser representados por uma ligação chamada, genericamente, de *link*. Um *link* corresponde a conexão entre duas pessoas (ou nós) em uma rede. Um tipo de Rede Social homogênea que pode ser estudada é a rede de coautorias científicas. Nesta rede, os nós são os autores e

os *links* são os relacionamentos de coautoria estabelecidos. Isto é, a colaboração entre os autores em uma publicação. Um dos problemas relacionados a redes sociais é a predição de relacionamentos (*links*), que consiste em prever a existência de relacionamentos ou interações entre pares de pessoas (Liben-Nowell and Kleinberg, 2007). Isto é, identificar uma relação que existe no mundo real, mas não está representada na rede (identificação de *link* faltante) ou a previsão de um relacionamento/ligação que não existe atualmente, mas existirá no futuro.

A pesquisa em predição de relacionamentos possui uma grande variedade de aplicações práticas. Por exemplo, as populares redes sociais online sugerem uma lista de pessoas que você talvez conheça. No comércio eletrônico é possível ser aplicada para a recomendação de produtos, já que podem ser interpretadas como previsões de *links* em grafos bipartidos entre produtos e pessoas) (Kaya, 2020). A predição de *links* também pode ajudar na reconstrução de redes biológicas (simulando/prevendo ligações proteicas) (Lü and Zhou, 2011).

Em redes sociais acadêmicas, a predição de relacionamentos tem sido utilizada principalmente para a predição de coautorias, atividade que indica se um par de pesquisadores poderá/irá colaborar na produção de um artigo, podendo assim otimizar a produção destes pesquisadores por meio da indicação de pesquisadores cujas parcerias são mais promissoras. Assim, esse tipo de predição pode ser utilizada para favorecer a comunicação entre os pesquisadores por meio da sugestão de possíveis relacionamentos, almejando potencializar o processo de produção científica. Cada vez mais as pesquisas científicas lidam com problemas complexos que, para sua resolução, exigem a colaboração de vários especialistas. A formação de equipes adequadas bem como a identificação das expertises necessárias são desafios complexos e necessários no processo da produção científica.

A predição de relacionamentos possui alguns fatores que a torna complexa e desafiadora. Por exemplo, como conseguir um bom equilíbrio entre a quantidade de informações e a complexidade dos algoritmos? A eficiência dos algoritmos é importante, já que as redes reais são geralmente formadas por milhares ou até milhões de nós. Contudo, considerar apenas informações locais pode levar a predições ruins (Martínez et al., 2016). Além disso, quase todas as redes reais são esparsas, isto é, o número de *links* existente é extremamente menor do que o número de *links* possíveis. Logo, ao utilizar uma abordagem de aprendizado supervisionado tem-se o desafio de lidar com o grande desbalanceamento entre classes, o qual prejudica o desempenho de muitos algoritmos.

Uma estratégia para tratar a predição de relacionamentos é a abordagem híbrida. Nela são encontradas combinações de um ou mais métodos (Daud et al., 2020). Com essa abordagem, Aghabozorgi and Khayyambashi (2018) propõem uma nova métrica topológica local, que é calculada a partir das distribuições dos vértices em tríades. Além disso, utilizam dois algoritmos de aprendizado supervisionado o *Gradient Boosting Machine* (GBM) e o *Linear Discriminant Analysis* (LDA). Srilatha and Manjula (2017) usaram a inteligência de enxame para detectar semelhanças estruturais na rede social. Para isso combinam o Coeficiente de Agrupamento com o algoritmo de otimização *Firefly*. O algoritmo é inspirado no comportamento dos vaga-lumes, no qual vaga-lumes tendem a se mover em direção a outros de maior intensidade de luz (i.e, nós com alto grau). Maruyama and Digiampietri (2016) estudaram uma rede de coautoria que foi construída com informações extraídas da Plataforma Lattes. Utilizam algoritmos de classificação incorporando diferentes características, como topologia e informações contextuais da rede.

Esse trabalho, diferentemente de outras abordagens híbridas, propõem uma estratégia de aprendizado supervisionado que combina:

i. técnicas de classificação *ensembles* e de agrupamento;
ii. características de domínio, de estrutura local, estrutura global e estrutura com informações de comunidade.

O presente trabalho visa a responder às seguintes perguntas de pesquisa:

i. Usar uma estratégia de agrupamento prévio pode melhorar a predição?
ii. A inclusão de atributos que levam em consideração informações de comunidade pode contribuir na predição?

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os conceitos fundamentais. Na Seção 3 os principais trabalhos correlatos são sumarizados. A Seção 4 contém a descrição dos materiais e métodos utilizados. Na Seção 5 os resultados são apresentados e discutidos. Por fim, a Seção 6 contém as conclusões.

2 Conceitos Fundamentais

2.1 Predição de relacionamentos temporal

Formalmente, a tarefa de predição de *link* temporal pode ser formulada como: Considerando um grafo não ponderado e não direcionado $G = (V, E)$ que representa uma estrutura topológica de uma rede, como redes sociais, onde V é o conjunto de nós em G, e E é o conjunto de arestas existentes em G. Cada aresta $e = (u, v) \in E$ entre os nós $u, v \in V$ representa uma interação (ou relacionamento) entre u e v que ocorreu em um determinado tempo $t(e)$. Para duas instâncias de tempo, t e t', onde $t' > t$, considerando que $G[t, t']$ denota um subgrafo de G que consiste em todas as arestas com data/hora entre t e t'. Então, para uma dada rede $G[t_0, t'_0]$, temos uma lista de arestas não presentes em $G[t_0, t'_0]$ que estão previstas para aparecer na rede $G[t_1, t'_1]$, onde $t_0 < t'_0 \leq t_1 < t'_1$. O valor de V permanece o mesmo com o tempo (Fig. 1).

O problema de predição de *links* pode ser encarado com uma abordagem de aprendizado supervisionado, mais especificamente modelado com uma classificação binária. Neste caso, cada dado de treinamento corresponde a um par de vértices da rede e o rótulo do ponto representa a presença ou ausência de uma aresta entre o par. Dado um par de vértices (u, v) no grafo $G = (V, E)$ e a classe sugerida pelo modelo de classificação $A(u, v, t)$ (Kumar et al., 2020).

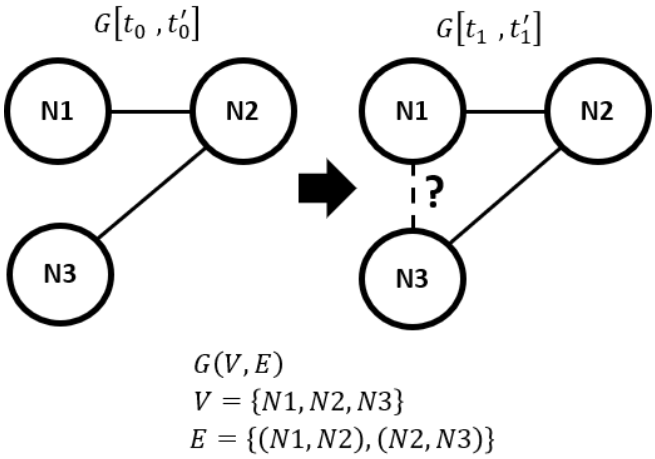


Figura 1: Exemplo de predição em um grafo.

$$\text{A(u,v,t)} = \begin{cases} 1, & \text{se } (u,v) \in E \text{ no tempo t.} \\ 0, & \text{se } (u,v) \notin E \text{ no tempo t.} \end{cases} \quad (1)$$

Assim, para uma dada janela de tempo de uma rede t_0 a predição de *links* encontra *links* que surgem no próximo intervalo de tempo t_1.

2.2 Engenharia de características

Na abordagem supervisionada um dos maiores desafios é a identificação apropriada de características (Hasan and Zaki, 2011). Esse processo é conhecido como Engenharia de Características, no qual é vital a transformação dos dados brutos para uma representação do problema de aprendizagem (Prado and Digiampietri, 2020). Para transformar dados brutos em características diferentes funções podem ser utilizadas, como soma, média, máximo, mínimo e contagem (Hasan et al., 2006). Por exemplo, podemos verificar a utilização da função de contagem nos seguinte atributos da Tabela 1: "Artigos em anais do pesquisador 1", "Artigos em anais do pesquisador 2", "Artigos em periódicos do pesquisador 1" e "Artigos em periódicos do pesquisador 2".

Pode-se observar, na Tabela 1, a existência de duas categorias de características, as métricas baseadas no grafo/estrutural e as métricas baseadas em conteúdo/domínio. As medidas baseadas em grafos são as abordagens mais comuns e usam a estrutura topológica das redes. Sua vantagem é a aplicabilidade em qualquer domínio. As medidas baseadas em conteúdo consideram os atributos dos vértices e arestas. Essas características têm a vantagem de utilizar padrões intrínsecos ao domínio da rede.

2.3 Atributos de comunidade

A detecção de comunidade refere-se à atividade de identificar grupos em redes sociais de acordo com suas propriedades estruturais. Um método conhecido é o algoritmo de agrupamento hierárquico chamado Louvain (Blondel et al., 2008), com implementações disponíveis para diversas linguagens de programação, como para Python na biblioteca community. Alguns trabalhos propõem a incorporação dessa informação da comunidade para a criação de atributos utilizados para a predição de *links*.

No presente trabalho, é proposta a utilização de alguns atributos que incorporam a informação de comunidade. Sendo que os quatro primeiros atributos (CN_SH, RA_SH, CCPA e WIC) foram calculados utilizando a biblioteca Networkx que dispõem de suas implementações.

- CN_SH (Soundarajan and Hopcroft, 2012): calcula o número de vizinhos comuns entre dois pesquisadores e soma-se 1 para cada vizinho comum pertencente à mesma comunidade do par de pesquisadores. Onde $f(w)$ é igual a 1 se w pertence à mesma comunidade que u e v ou 0 caso contrário. E $\Gamma(u)$ denota o conjunto de vizinhos de u.

$$\text{CN\_SH} = |\Gamma(u) \cap \Gamma(v)| + \sum_{w \in \Gamma(u) \cap \Gamma(v)} f(w) \quad (2)$$

- RA_SH (Soundarajan and Hopcroft, 2012): calcula o índice de alocação de recursos considerando apenas vizinhos comuns pertencentes à mesma comunidade do par de pesquisadores.

$$\text{RA\_SH} = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{f(w)}{|\Gamma(w)|} \quad (3)$$

- CCPA (Common Neighbor Centrality (Ahmad et al., 2020)): é baseado no número de vizinhos comuns e sua centralidade. Onde α é um parâmetro que varia entre [0,1], N denota o número total de nós no grafo e d_{uv} denota a distância mais curta entre u e v.

$$\text{CCPA} = \alpha \cdot (|\Gamma(u) \cap \Gamma(v)|) + (1 - \alpha) \cdot \frac{N}{d_{uv}} \quad (4)$$

- WIC (Within Inter Cluster (Valverde-Rebaza and de Andrade Lopes, 2012)): calcula a proporção de vizinhos comuns dentro e entre clusters de todos os pares de pesquisadores na comunidade.
- CC (Comunidade em Comum): atributo binário, no qual 1 indica a correspondência da comunidade do par de pesquisadores e 0 caso contrário.

$$\text{CC(u,v)} = \begin{cases} 1, & \text{se } \textit{comunidade}(u) = \textit{comunidade}(v) \\ 0, & \text{caso contrário} \end{cases} \quad (5)$$

2.4 Métricas

A maioria das abordagens considera o problema de predição de relacionamentos como uma tarefa de classificação binária, sendo os casos positivos a existência de conexão e casos negativos a ausência. Para avaliar o desempenho dos algoritmos, algumas métricas são amplamente encontradas na predição de relacionamentos e são apresentados a seguir.

- Especificidade: mede a capacidade do modelo de detectar casos negativos.

$$\text{Especificidade} = \frac{\#\text{Verdadeiro Negativo (VN)}}{\#\text{VN} + \#\text{Falso Positivo (FP)}} \quad (6)$$

- Revocação ou Sensibilidade: mede a proporção dos casos positivos detectados pelo modelo

$$\text{Sensibilidade} = \frac{\#\text{Verdadeiro Positivo (VP)}}{\#\text{VP} + \#\text{False Negativo (FN)}} \quad (7)$$

- Precisão: mede a proporção de casos positivos corretamente os detectados pelo modelo.

$$\text{Precisão} = \frac{\#\text{VP}}{\#\text{VP} + \#\text{FP}} \quad (8)$$

- Medida-F: representa o equilíbrio (média harmônica) entre a precisão e a revocação.

$$F = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \quad (9)$$

- Acurácia: é a proporção/taxa de acertos total da classificação.

$$\text{Acurácia} = \frac{\#\text{VP} + \#\text{VN}}{\#\text{VP} + \#\text{FP} + \#\text{FN} + \#\text{VN}} \quad (10)$$

- Acurácia Balanceada: é uma média entre a sensibilidade e especificidade.

$$\text{Acurácia Balanceada} = \frac{\text{Sensibilidade} + \text{Especificidade}}{2} \quad (11)$$

3 Trabalhos Correlatos

Lü and Zhou (2011) e Hasan and Zaki (2011) realizaram levantamentos de diversas abordagens baseadas na topologia de grafos para predição de *links*. O primeiro conjunto de atributos foi denominado como conjunto de métricas baseadas em vizinhança (Hasan and Zaki, 2011) ou métricas de índice local (Lü and Zhou, 2011). Elas são métricas calculadas com base na informação local de um nó, isto é, utiliza-se a informação da estrutura dos nós vizinhos. O segundo conjunto de atributos foi chamado por Hasan and Zaki (2011) de conjunto de métricas baseadas no caminho. Enquanto Lü and Zhou (2011) chamaram essa categoria de conjunto de índice global. Essas métricas utilizam informação global da rede (não apenas informações exclusivas de um par de nós), por exemplo, considera todos os caminhos possíveis entre um par de nós. Em outro trabalho sobre o estado da arte, Martínez et al. (2016) propuseram uma taxonomia para a classificação dos métodos de predição de *links*. Nessa taxonomia, as técnicas foram divididas em quatro grandes grupos: métodos baseados em similaridade, probabilísticos e estatísticos, baseado em algoritmos e métodos de pré-processamento.

Ozcan and Oguducu (2016) propuseram um método que inicialmente calcula a similaridade entre os pares de nós com base em medidas quasi-local em diferentes janelas de tempo da rede. Então, aplicam um modelo de predição de *links* baseado na Rede Neural NARX. Para avaliar a proposta usaram uma rede coautoria criada a partir do DBLP.

Fu et al. (2018) utilizaram algoritmos de aprendizado supervisionado Random Forest (RF), Gradient Boosting, Decision Tree (GBDT) e Support Vector Machine (SVM). Para realizar essa predição, foram extraídas características do grafo original e características do grafo de linha. Sendo que o primeiro grupo são características estruturais locais comumente utilizadas (como CN, PA, HPI, SO, SA, JC e etc.). O segundo grupo de características foi extraído com objetivo de considerar os pesos dos links. Para isso os pesquisadores transformam o grafo original em grafo de linha e, em seguida, extraíram as características das arestas no grafo original usando os índices de centralidade no grafo de linha.

Chuan et al. (2018) implementaram uma nova métrica para predição de *links* em redes de coautoria baseada na similaridade de conteúdo, denominada LDAcosin. Utilizaram uma abordagem supervisionada e criaram vetores de características combinando a medida proposta com métricas ponderadas de predição de *links*. Os vetores foram utilizados com o método de classificação binária SVM ponderado. Aplicaram o LDA usando o título e o resumo de cada artigo. A ideia é que quanto maior a similaridade entre dois conjuntos de artigos, maior a possibilidade de vínculo no futuro.

Fu et al. (2018) adotaram algoritmos de aprendizado supervisionado como Random Forest (RF), Gradient Boosting, Decision Tree (GBDT) e Support Vector Machine. Para realizar essa predição, foram extraídas características do grafo original e características do grafo de linha. Sendo que o primeiro grupo são características estruturais locais comumente utilizadas (como CN, PA, HPI, SO, SA, JC e etc.). O segundo grupo de características foi extraído com objetivo de considerar os pesos dos links. Para isso os pesquisadores transformam o grafo original em grafo de linha e, em seguida, extraem as características das arestas no grafo original usando os índices de centralidade no grafo de linha.

A pesquisa dos autores Li, Tu and Chai (2020) utiliza uma abordagem supervisionada, de classificação binária, chamada de *Ensemble-model-based* (EMLP). Nela são calculados quatro índices de similaridade CN, LHN-II, COS+ e MFI dos pares de nós na rede. Segundo os autores, a seleção desses diferentes índices de similaridades considera a representatividade de diferentes informações sobre a estrutura da rede. Além da integração dos índices, também realizaram a integração, com o algoritmo Stacking, do modelo de regressão logística e o modelo Xgboost.

Li, Song, Lu, Zeng, Shi and Liu (2020) propõem o algoritmo de predição de *links* de alocação de atenção inteligente (IAALPA) para calcular de maneira adaptativa o índice de alocação de atenção (AAI) de acordo com a dispersão da rede e prever as possíveis amizades entre usuários

em diferentes círculos sociais.

Bastami et al. (2019) propuseram um arcabouço multinível chamado GLP (*Gravitation-based link prediction*) para predição de links. O trabalho incorpora detecção de comunidades de grafo (utilizando o método Louvain (Blondel et al., 2008)), otimização de subgrafos (utilizando gravidade newtoniana e leis de movimento, chamado de GSA ou *Gravitational Search Algorithm*) e aplicação de métodos de predição de *links* local (como o Adamic-Adar) nos subgrafos otimizados e que pode ser executado concorrentemente. Os *links* externos entre todos os subgrafos otimizados relevantes são preditos a partir de uma força de limiar calculada entre os subgrafos.

4 Materiais e Métodos

4.1 Conjunto de dados

O conjunto de dados do experimento foi extraído da Plataforma Lattes e foi construído por Digiampietri et al. (2013) e também utilizado por Digiampietri and Maruyama (2014). Nesse conjunto foram selecionados os 657 currículos de pesquisadores permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que foram avaliados nos triênios 2004-2006 e 2007-2009 pela CAPES. A partir deste conjunto foram extraídos ou calculados 32 atributos (ou características).

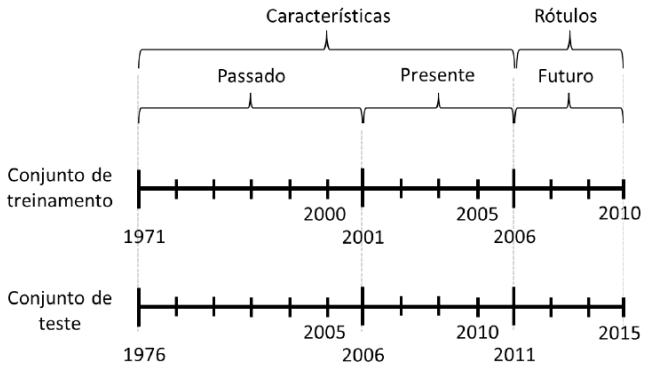


Figura 2: Janelas de tempo dos conjuntos de dados

Os dados foram coletados em uma janela de tempo de 1971 a 2015. Para a etapa de treinamento, os dados de 1971 a 2000 foram considerados passado; de 2001 a 2005 presente; e de 2006 a 2010 foi considerado futuro (ou seja, as coautorias que o sistema deveria prever). Para a etapa de teste, os dados de 1976 a 2005 foram considerados passados; de 2006 a 2010 presente e o sistema tentou prever as coautorias que ocorreram de 2011 a 2015. A Fig. 2 ilustra essa divisão.

4.2 Método proposto

O método proposto é ilustrado na Fig. 3 e é descrito a seguir:

a) Seleção dos atributos: nessa etapa são selecionados os atributos que serão utilizados. Na Tabela 1 há a indicação (com um "x") de quais atributos são encontrados em cada conjunto criado. Com o objetivo de responder as perguntas deste trabalho, foram criados quatro conjuntos de dados para os experimentos:

- Original: atributos utilizados em Digiampietri and Maruyama (2014);
- Comunidade: apenas os atributos que incluem informação de comunidade;
- Completo: todos os atributos;
- Seleção: foi obtido com a utilização do algoritmo de seleção de características RFECV (*Recursive Feature Elimination*).

b) Agrupamento: o algoritmo K-means foi utilizado para o agrupamento. Para determinar o número de grupos (ou *clusters*) ideal, utilizou-se o coeficiente de silhueta, calculado sobre o conjunto original de atributos. O algoritmo foi treinado com o conjunto de treinamento e posteriormente foi utilizado para rotular as instâncias do conjunto de teste e criar os grupos de teste.

c) Treinamento do modelo: cada grupo obtido do conjunto de treinamento foi utilizado para treinar um modelo de classificação *Random Forest*. Para encontrar os hiper parâmetros do algoritmo foi utilizada a estratégia de pesquisa em grade com validação cruzada estratificada com 10 subconjuntos;

d) Predição: Os grupos de teste são usados para predição, sendo que cada grupo é apresentado ao modelo correspondente ao rótulo de agrupamento. Nessa etapa, além de usar apenas os modelos *Random Forest*, mais duas estratégias foram testadas. Com os hiper parâmetros encontrados dos modelos *Random Forest*, eles foram combinados usando os algoritmos de *Stacking* e *Voting*. No primeiro, o classificador final foi a Regressão Logística e, no último, usou-se a classificação pela maior probabilidade;

e) Análise dos resultados: os resultados da predição são analisados segundo as métricas apresentadas anteriormente.

5 Resultados

A Fig. 5 apresenta o gráfico da proporção de classes, destacando-se um alto desbalanceamento de classes, como esperado na maioria das redes sociais. Isto é, há muitos casos de ausência de relações de coautoria (classe 0) em comparação com a presença de coautoria (classe 1). Considerando esse cenário, a métrica de acurácia não é indicada para avaliação de um modelo, pois uma acurácia alta como 95,24% poderia ter sido obtida classificando todos os casos como negativos. Portanto, este trabalho está sendo avaliado pelas métricas de precisão, sensibilidade e a medida-F, pois são métricas que vão analisar a classe de interesse (a existência de relacionamento).

A Tabela 2 sumariza em cada linha a execução de uma configuração e os respectivos resultados obtidos. A coluna *Agrupamento* indica a utilização ou não do K-Means e a coluna *Modelos* apresenta a combinação de algoritmos de classificação aplicada. A coluna *Atributos* apresenta o con-

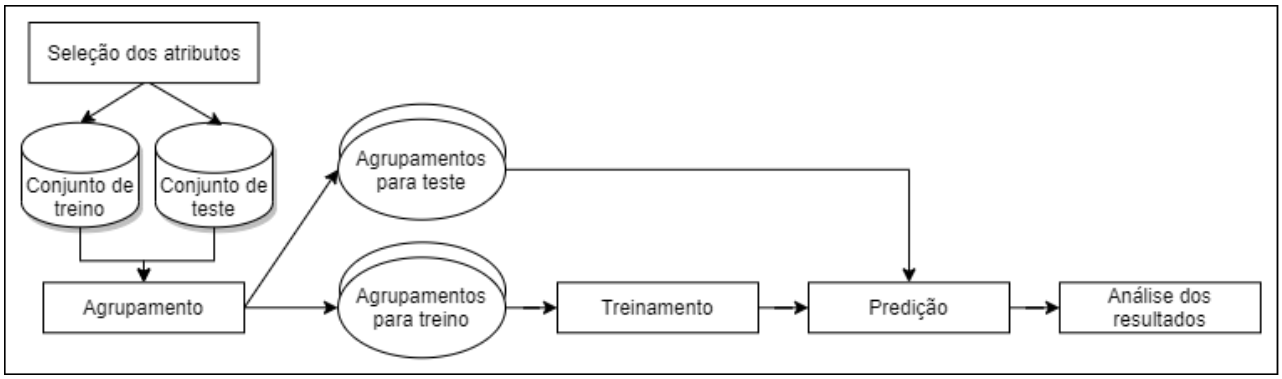


Figura 3: Fluxo do método proposto.

junto de atributos utilizado, sendo possível averiguar na Tabela 1 os atributos que formam cada conjunto.

Para definir a quantidade de grupos (*clusters*) foi analisado o valor de *silhouette* variando o parâmetro K do K-Means. Os resultados obtidos são apresentados na Fig. 4 e a partir deles foi definido o valor de K=2 para os demais experimentos.

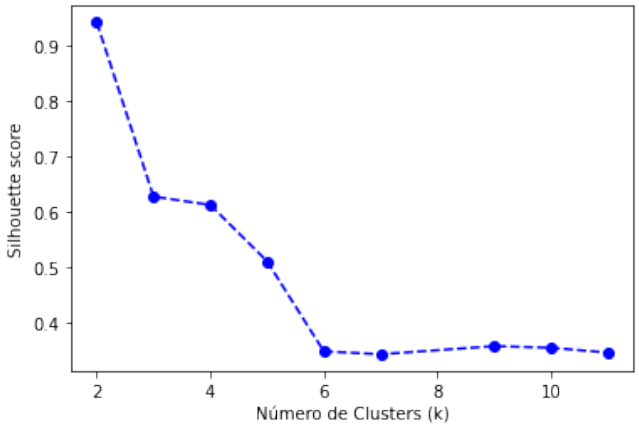


Figura 4: Score de silhouette em relação ao número de clusters do K-Means aplicado no conjunto *original*.

Para analisar os resultados, foram selecionados os melhores (Top-1) resultados das métrica de precisão, sensibilidade e medida-F de cada conjunto. Deste modo, foi possível observar, a partir da Tabela 2, os seguintes resultados:

- Precisão (linhas 2, 10, 14, 7): o modelo *Random Forest + Stacking* esteve presente nos cinco conjuntos. A maior precisão foi de 0,567 obtida com o conjunto *original*;
- Sensibilidade (linhas 9, 1, 3, 15): o modelo que mais esteve presente foi o *Random Forest*. A maior sensibilidade foi de 0,671 obtida com o conjunto "todos";
- medida-F (linhas 15, 2, 5, 11): na maioria dos casos o melhor modelo foi a combinação do *Random Forest* com *Voting* ou *Stacking*. Além disso, o conjunto *seleção* com a utilização do modelo *Random Forest + Voting* resultou na maior medida-F, 0,513.

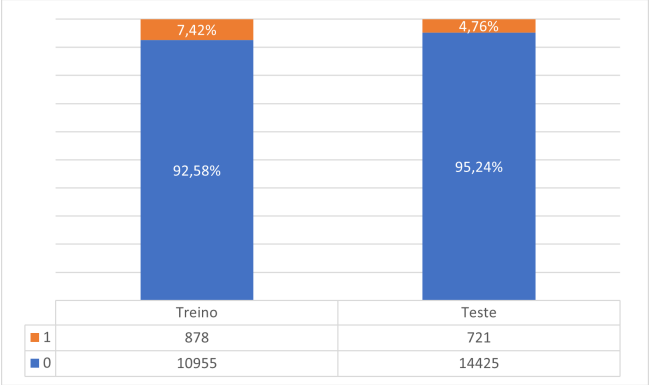


Figura 5: Proporção de classes nos conjuntos de dados.

A seguir, são apresentadas as discussões referentes às perguntas de pesquisa que motivaram o presente trabalho.

5.1 Usar uma estratégia de agrupamento prévio pode melhorar a predição?

Foi realizada uma comparação entres as abordagens de aplicar e não aplicar o agrupamento. Para isso, foram selecionados os melhores resultados obtidos em cada conjunto aplicando e não aplicando o agrupamento, independente do modelo usado. Com essa seleção, os gráficos presentes na Fig. 6 foram criados para as métricas de precisão, sensibilidade e medida-F.

Ao observar apenas o conjunto "original" o agrupamento melhorou a precisão e a medida-F dos modelos comparados, contudo a sensibilidade foi inferior. Enquanto o conjunto "comunidade" não apresentou diferenças nos resultados.

Tabela 1: Conjunto de atributos.

				Conjunto de atributos			
#	**Atributo**	**Categoria**	**Subcategoria**	**Original**	**Comunidade**	**Completo**	**Seleção**
1	Artigos em anais do pesquisador 1	Domínio	Contagem	x		x	x
2	Artigos em anais do pesquisador 2	Domínio	Contagem	x		x	x
3	Atigos em periódicos do pesquisador 1	Domínio	Contagem	x		x	
4	Artigos em periódicos do pesquisador 2	Domínio	Contagem	x		x	
5	Orientação em andamento	Domínio	Booleano	x		x	
6	Orientação passado	Domínio	Booleano	x		x	
7	Orientação presente	Domínio	Booleano	x		x	
8	Orientadores em comum	Domínio	Intersecção	x		x	
9	Orientandos em comum	Domínio	Intersecção	x		x	
10	Periódicos passado	Domínio	Soma	x		x	
11	Periódicos presente	Domínio	Soma	x		x	
12	Conferência passado	Domínio	Soma	x		x	
13	Conferência presente	Domínio	Soma	x		x	x
14	Programas em comum	Domínio	Intersecção	x		x	
15	Subáreas em comum	Domínio	Intersecção	x		x	
16	Distância geográfica	Domínio	-	x		x	x
17	CN (Common Neighbors) presente	Estrutural	Local	x		x	x
18	CN passado e presente	Estrutural	Local	x		x	
19	JC (Jaccard Coefficient)	Estrutural	Local	x		x	
20	AA (Adamic-Adar)	Estrutural	Local	x		x	
21	PA (Conexão Preferencial)	Estrutural	Local	x		x	
22	SA (Salton)	Estrutural	Local	x		x	
23	SO (Sorensen)	Estrutural	Local	x		x	
24	HPÌ (Hub Promoted Index)	Estrutural	Local	x		x	
25	HDI (Hup Depressed Index)	Estrutural	Local	x		x	
26	Leicht-Holme-Newman (LHN)	Estrutural	Local	x		x	
27	RA (Resource Allocation)	Estrutural	Local	x		x	
28	SP (Shortest Path)	Estrutural	Global	x		x	
29	Katz $\beta = 0,05$	Estrutural	Global	x		x	x
30	Katz $\beta = 0,005$	Estrutural	Global	x		x	
31	Katz $\beta = 0,0005$	Estrutural	Global	x		x	
32	CN_SH	Estrutural	Local		x	x	
33	RA_SH	Estrutural	Local		x	x	
34	WIC	Estrutural	Local		x	x	
35	CCPA	Estrutural	Local		x	x	
36	CC	Estrutural	Local		x	x	

5.2 A inclusão de atributos que levam em consideração a comunidade pode contribuir para a predição?

Sobre a segunda pergunta "a inclusão de atributos que levam em consideração a comunidade pode contribuir para a predição?", foi realizada uma comparação entre as abordagens com agrupamento, para cada variação de modelos e para cada conjunto de atributos. Os gráficos presentes na Fig. 7 foram criados para as métricas de precisão, sensibilidade e medida-F.

Com a inclusão dos atributos que consideram comunidade, foi observando que os conjuntos *original* e *todos* os resultados obtidos com os modelos *Random Forest* e *Random Forest + Voting* apresentaram uma precisão e medida-F superiores. Esse aumento de precisão também foi observado por Soundarajan and Hopcroft (2012) que propuseram os atributos CN_SH e RA_SH.

6 Conclusão

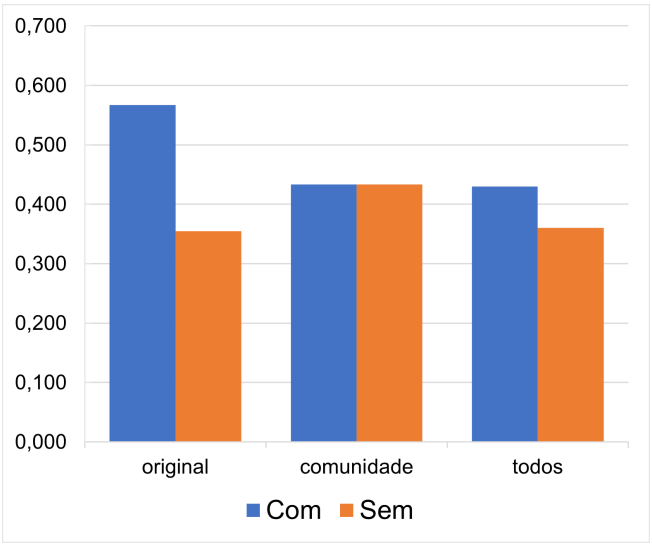
Este trabalho avaliou a utilização de agrupamento como uma das etapas no processo de predição de relacionamentos. Adicionalmente, atributos relacionados a *comunidades* foram calculados e adicionados ao conjunto de características utilizado pelos classificadores.

A partir dos resultados obtidos, foi possível verificar que o uso da estratégia de agrupamento e adição de atributos do conjunto *comunidade* pode melhorar a predição, especificamente na precisão e medida-F. Contudo para aplicar a estratégia de agrupamento é necessário encontrar um número apropriado de grupos (*clusters*). Além do agrupamento, esse resultado foi alcançado com à combinação dos algoritmos *Random Forest* por meio do *Stacking* ou *Voting*. Isto é, estes modelos conseguiram detectar corretamente mais casos positivos.

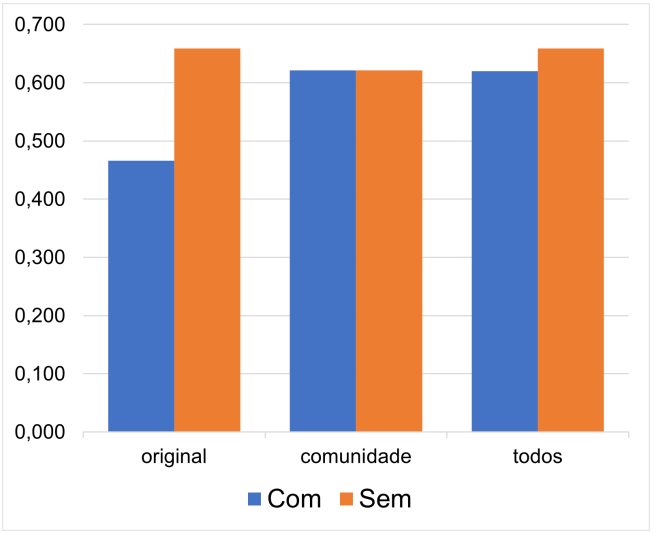
Outro ponto a ser destacado foi o melhor resultado obtido pela medida-F com o conjunto *seleção*, formado por seis atributos (Tabela 1). Deste modo, foi possível obter um valor maior com uma redução de dimensão do conjunto.

Tabela 2: Resultados obtidos em cada conjunto, com variações nos algoritmos e com variação no uso do agrupamento.

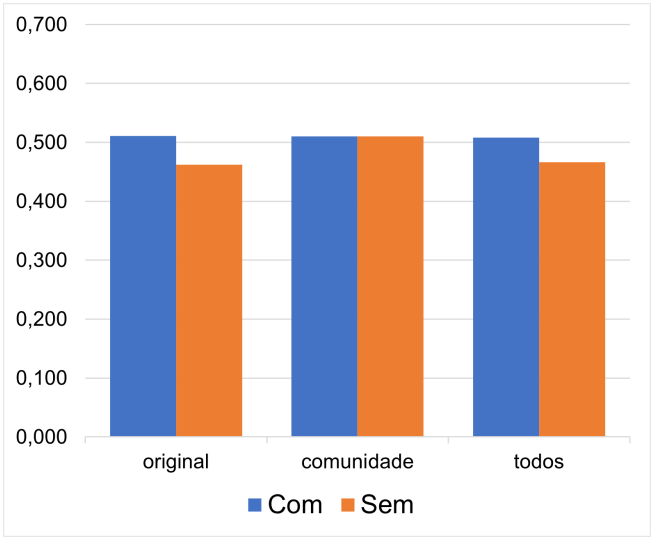
Linha	Atributos	Agrupamento?	Modelos	Acurácia Balanceada	Acurácia	Medida-F	Precisão	Sensibilidade	AUC
1	original	Com	Random Forest	0,790	0,899	0,387	0,272	0,670	0,870
2	original	Com	Random Forest + Stacking	0,724	0,958	0,511	0,567	0,466	0,884
3	original	Com	Random Forest + Voting	0,799	0,927	0,463	0,357	0,657	0,882
4	original	Sem	Random Forest	0,800	0,927	0,462	0,355	0,659	0,882
5	comunidade	Sem	Random Forest	0,790	0,943	0,510	0,433	0,621	0,818
6	comunidade	Com	Random Forest	0,790	0,943	0,510	0,433	0,621	0,817
7	comunidade	Com	Random Forest + Stacking	0,667	0,950	0,402	0,463	0,355	0,818
8	comunidade	Com	Random Forest + Voting	0,790	0,943	0,510	0,433	0,621	0,818
9	todos	Com	Random Forest	0,803	0,923	0,454	0,343	0,671	0,879
10	todos	Com	Random Forest + Stacking	0,714	0,956	0,490	0,544	0,447	0,885
11	todos	Com	Random Forest + Voting	0,789	0,943	0,508	0,430	0,620	0,885
12	todos	Sem	Random Forest	0,800	0,928	0,466	0,360	0,659	0,883
13	seleção	Com	Random Forest	0,758	0,950	0,511	0,480	0,546	0,848
14	seleção	Com	Random Forest + Stacking	0,728	0,956	0,507	0,541	0,477	0,879
15	seleção	Com	Random Forest + Voting	0,787	0,945	0,513	0,441	0,613	0,878



(a) Precisão

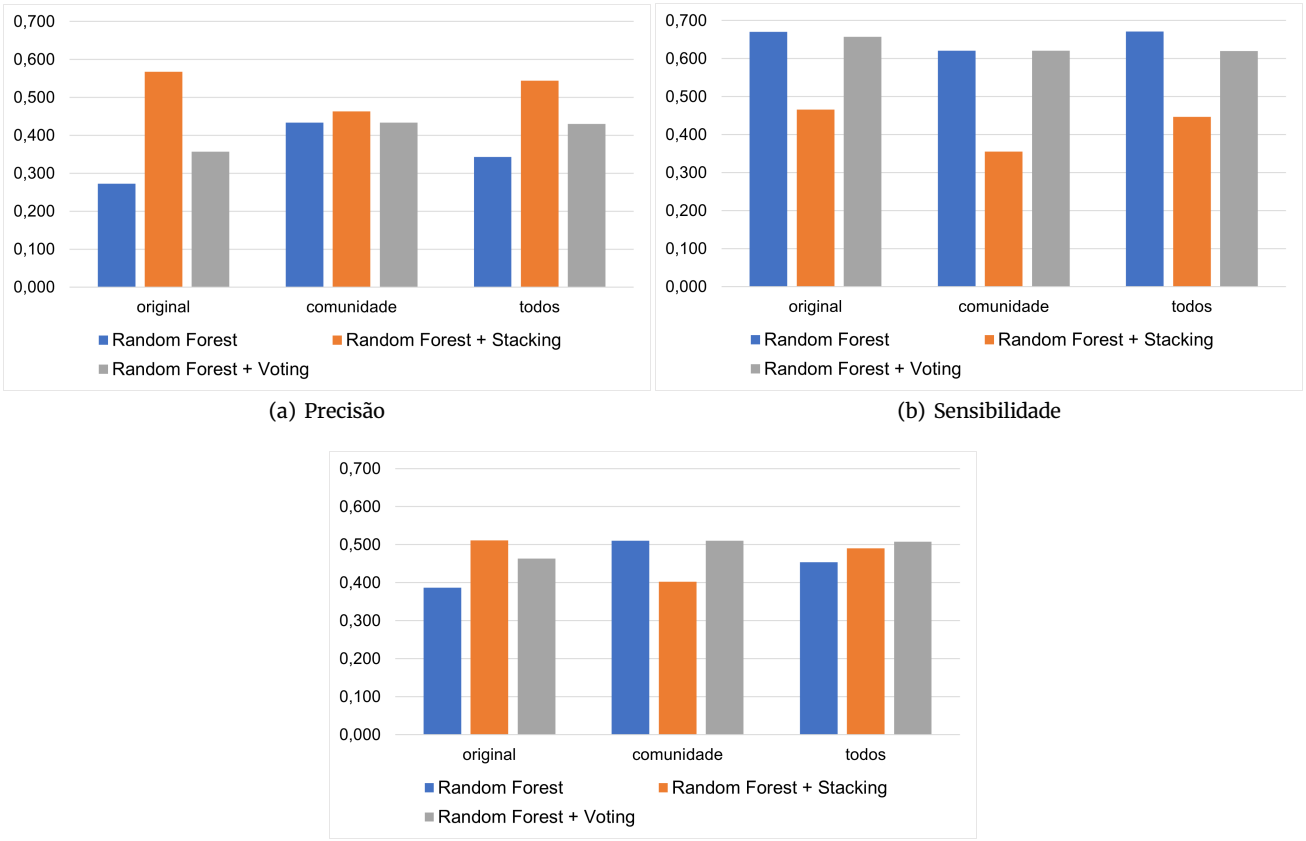


(b) Sensibilidade



(c) Medida-F

Figura 6: Gráficos para comparativo em relação ao agrupamento.



(a) Precisão

(b) Sensibilidade

(c) Medida-F

Figura 7: Gráficos para comparativo em relação aos atributos de comunidade.

Como trabalhos futuros, para um aprofundamento sobre o assunto, sugere-se a aplicação de outros seletores de atributos para explorar novas combinações de atributos e a aplicação de estratégias específicas para o tratamento do desbalanceamento do conjunto de dados.

Referências

Aghabozorgi, F. and Khayyambashi, M. R. (2018). A new similarity measure for link prediction based on local structures in social networks, **501**: 12–23. https://doi.org/10.1016/j.physa.2018.02.010.

Ahmad, I., Akhtar, M. U., Noor, S. and Shahnaz, A. (2020). Missing link prediction using common neighbor and centrality based parameterized algorithm, **10**(1): 364. https://doi.org/10.1038/s41598-019-57304-y.

Bastami, E., Mahabadi, A. and Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks, **44**: 176–186. https://doi.org/10.1016/j.swevo.2018.03.001.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10): P10008. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T. and Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric, **48**(8): 2470–2486. https://doi.org/10.1007/s10489-017-1086-x.

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F. and Anuar, N. B. (2020). Applications of link prediction in social networks: A review, **166**: 102716. https://doi.org/10.1016/j.jnca.2020.102716.

Digiampietri, L. A. and Maruyama, W. T. (2014). Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação, *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, p. 6. Disponível em https://sol.sbc.org.br/index.php/brasnam/article/view/6821.

Digiampietri, L. A., Santiago, C. R. N. and Alves, C. M. (2013). Predição de coautorias em redes sociais acadêmicas: Um estudo exploratório em ciência da computação, *Anais do II Brazilian Workshop on Social Network Analysis and Mining*, p. 12. Disponível em https://sol.sbc.org.br/index.php/brasnam/article/view/6839.

Fu, C., Zhao, M., Fan, L., Chen, X., Chen, J., Wu, Z., Xia, Y. and Xuan, Q. (2018). Link weight prediction using supervised learning methods and its application to yelp layered network, **30**(8): 1507–1518. https://doi.org/10.1109/TKDE.2018.2801854.

Hasan, M. A. and Zaki, M. J. (2011). *A Survey of Link Prediction in Social Networks*, Springer US, Boston, MA, pp. 243–275. https://doi.org/10.1007/978-1-4419-8462-3_9.

Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2006). Link prediction using supervised learning.

Kaya, B. (2020). Hotel recommendation system by bipartite networks and link prediction, **46**(1): 53–63. https://doi.org/10.1177%2F0165551518824577.

Kumar, A., Singh, S. S., Singh, K. and Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey, **553**: 124289. https://doi.org/10.1016/j.physa.2020.124289.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey, *Physica A: Statistical Mechanics and its Applications* **390**(6): 1150–1170. http://dx.doi.org/10.1016/j.physa.2010.11.027.

Li, K., Tu, L. and Chai, L. (2020). Ensemble-model-based link prediction of complex networks, **166**: 106978. https://doi.org/10.1016/j.comnet.2019.106978.

Li, S., Song, X., Lu, H., Zeng, L., Shi, M. and Liu, F. (2020). Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm, **139**: 112839. https://doi.org/10.1016/j.eswa.2019.112839.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks, *J. Am. Soc. Inf. Sci. Technol.* **58**(7): 1019–1031. https://doi.org/10.1002/asi.20591.

Martínez, V., Berzal, F. and Cubero, J.-C. (2016). A survey of link prediction in complex networks, *ACM Comput. Surv.* **49**(4). https://doi.org/10.1145/3012704.

Maruyama, W. T. and Digiampietri, L. A. (2016). Co-authorship prediction in academic social network, *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, Sociedade Brasileira de Computação - SBC, pp. 61–72. https://doi.org/10.5753/brasnam.2016.6445.

Ozcan, A. and Oguducu, S. G. (2016). Temporal link prediction using time series of quasi-local node similarity measures, *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 381–386. https://doi.org/10.1109/ICMLA.2016.0068.

Prado, F. F. and Digiampietri, L. A. (2020). A systematic review of automated feature engineering solutions in machine learning problems, *XVI Brazilian Symposium on Information Systems*, ACM, pp. 1–7. https://doi.org/10.1145/3411564.3411610.

Soundarajan, S. and Hopcroft, J. (2012). Using community information to improve the precision of link prediction methods, *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, ACM Press, p. 607. https://doi.org/10.1145/2187980.2188150.

Srilatha, P. and Manjula, R. (2017). Structural similarity based link prediction in social networks using firefly algorithm, *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, IEEE, pp. 560–564. https://doi.org/10.1109/SmartTechCon.2017.8358434.

Valverde-Rebaza, J. C. and de Andrade Lopes, A. (2012). Link prediction in complex networks based on cluster information, *in* L. N. Barros, M. Finger, A. T. Pozo, G. A. Gimenénez-Lugo and M. Castilho (eds), *Advances in Artificial Intelligence - SBIA 2012*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 92–101. https://doi.org/10.1007/978-3-642-34459-6_10.