HASIL CEK_Extending adamic

by Tif Extending Adamic

Submission date: 12-Aug-2023 10:35AM (UTC+0700)

Submission ID: 2144682861

File name: 15. Extending adamic adar for cold-start problem in link prediction based on network metrics.pdf (734.56K)

Word count: 7333

Character count: 38465

International Journal of Advances in Intelligent Informatics
Vol. 8, No. 3, November 2022, pp. 271-284

ISSN 2442-6571

Extending adamic adar for cold-start problem in link prediction based on network metrics

Herman Yuliansyah [a,b,1,*], Zulaiha Ali Othman [a,2], Azuraliza Abu Bakar [a,3]

[a] Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[b] Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Bantul, D.I. Yogyakarta, Indonesia
[1] herman.yuliansyah@tif.uad.ac.id; [2] zao@ukm.edu.my; [3] azuraliza@ukm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received August 1, 2022
Revised November 10, 2022
Accepted November 25, 2022
Available online November 30, 2022

Keywords
Link prediction
Triadic closure
Network metrics
Degree centrality
Closeness centrality
Clustering centrality

ABSTRACT

The cold-start problem is a condition for a new node to join a network with no available information or an isolated node. Most studies use topological network information with the Triadic Closure principles to predict links in future networks. However, the method based on the Triadic Closure principles cannot predict the future link due to no common neighbors between the predicted node pairs. Adamic Adar is one of the methods based on the Triadic Closure principles. This paper proposes three methods for extending Adamic Adar based on network metrics. The main objective is to utilize the network metrics to attract the isolated node or new node to make new relationships in the future network. The proposed method is called the extended Adamic Adar index based on Degree Centrality (DCAA), Closeness Centrality (CloCAA), and Clustering Coefficient (CluCAA). Experiments were conducted by sampling 10% of the dataset as testing data. The proposed method is examined using the four real-world networks by comparing the AUC score. Finally, the experiment results show that the DCAA and CloCAA can predict up to 99% of node pairs with a cold-start problem. DCAA and CloCAA outperform the benchmark, with an AUC score of up to 0,960. This finding shows that the extended Adamic Adar index can overcome prediction failures on node pairs with cold-start problems. In addition, prediction performance is also improved compared to the original Adamic Adar. The experiment results are promising for future research due to successfully improving the prediction performance and overcoming the cold-start problem.

1. Introduction

The cold-start problem is a common problem in recommendation systems (RS) [1]-[3]. User cold-start problems [4],[5] and product cold-start problems [6],[7] are two categories of cold-start problems in the RS. Both types of problems arise because no information is available and cause the recommender system to underperform in handling sparse data [8]. In addition, the recommender system cannot provide specific recommendations due to insufficient information [9]. For prediction, information is usually collected based on a particular period as training data to provide appropriate recommendations [10]. Therefore, the recommender system must recognize the user or product in a cold-start problem, so it needs special handling in the prediction process.

The cold-start problem also occurs in link prediction [11]-[13]. Link prediction is a method for identifying the future link based on existing information [14]-[17]. Leroy *et al.* [18] introduced link prediction with the cold-start problem in 2010. The cold-start problem is classified into partial and pure

cold-start [13]. A node with a low degree number has the potential to cause a partial cold start problem, and a node isolated has the potential to cause a pure cold start problem. A method needs to propose for overcoming the problem by attracting the node with pure cold-start to select the appropriate node and starting a new relationship. This specific node is the basis for other relations in the future, and the future link can be predicted using similarity-based methods. There are two classifications in link prediction: similarity-based methods [19],[20] that use topological information such as Triadic Closure, and learning-based methods that use non-topological information such as attributes of nodes or edges to predict the future link. The Triadic Closure principle is that a common friend between two people in a network has a greater probability that the two people will become friends in the future [21]-[23].

Several studies proposed solutions to deal with the cold-start problem. Leroy et al. [18] proposed a method based on the probabilistic bootstrap graph. Ge & Zhang [24] presented the pseudo-cold-start link prediction with the two-phase method for predicting social structure in case only available small subgraph and multiple heterogeneous sources of the social network. The first phase generates a feature selection scheme and proposes a regularization method in the second phase to control over-fitting risk. Yan *et al.* [25] investigated friend recommendations based on cross-platform social relationships and behavior information. Zhang *et al.* [26] accommodated the information distribution using additional transfer information of old users from auxiliary sources for the new user. Rohani *et al.* [27] proposed an algorithm incorporating social features and faculty and friend's mate's ratings in social networks for academics.

Furthermore, Han *et al.* [28] examined the users' attributes and proposed new users prediction using users' social features based on the Support Vector Machine. Wang *et al.* [29] also presented a possible connection between cold-start users and existing users based on topological information extraction. The proposed method used the latent-feature representation model and established the relationship based on topological and non-topological information. Zhu *et al.* [30] proposed a recommendation system combining auxiliary and heterogeneous information in multiple networks. The latest approaches were multi-relational networks [13], [31] and learning community-specific [32], [33]. Wu *et al.* [31] summarized the complex system into multi-relational networks and used the latent space network model for extracting low-dimensional sub-networks factors. The target and auxiliary sub-networks regression was also proposed for predicting the potential links of cold-start nodes in target sub-networks [13]. Meanwhile, Xu *et al.* [32], [33] presented two models for learning community similarity metrics using community detection and named the community-weighted ranking (CWR) and probability (CWP) models.

Auxiliary networks and community information are generally chosen to address the cold-start problem. Both approaches have derivative problems: personalization, privacy, and overlapping community. The lack of auxiliary networks also needs cross-platforms of other networks and transfers existing links from the auxiliary to target networks [28]. Besides, both approaches require complex computing. Therefore, this research proposes a simple method by extending the Adamic Adar index and utilizing the network metrics to attract low-degree or isolated nodes. Although, several previously proposed methods have also extended Adamic Adar. The proposed methods did not aim to overcome the cold-start problem. Nassar *et al.* [34] proposed pairwise link prediction to predict new triangles based on extending Jaccard similarity and Adamic Adar. Later, Liu *et al.* [35] proposed weighted similarity based on extending unweighted similarity: Common Neighbor, Adamic Adar, Salton similarity, Jaccard similarity, Resource Allocation, and Local Path.

In summary, the contribution of this research includes:

1. To propose three novel methods by extending the Adamic Adar index based on network metrics by combining the local and global measures of the network: degree centrality, closeness centrality, clustering coefficient, and network density. Furthermore, this proposed method is called the extended Adamic Adar index based on Degree Centrality (DCAA), extended Adamic Adar index based on Closeness Centrality (CloCAA), and extended Adamic Adar index based on Clustering Coefficient (CluCAA).

2. To conduct experiments and measurements using four networks to examine whether the proposed method achieves the better performance in link prediction.

The rest of the sections are arranged as follows. Section 2 introduces extending Adamic Adar, and the experiment results are discussed in Section 3. The conclusion is presented in Section 4.

2. Method

2.1. Network Metrics

The network metrics are mathematical properties to quantify network topology information. The information topology is classified into local and global information. Local information is the information quantification of a node, and global information is related to the complete information of a network. The degree of centrality, closeness centrality, and clustering coefficient are local measures. The degree of centrality shows a node's importance based on the number of connected nodes [36]. The degree of the node has a hub information function and has a higher impact on influencing the other nodes. The degree centrality d(u) of node u is defined in Equation (1).

$$d(u) = \sum_v m_{uv} \tag{1}$$

where m_{uv} is equal to 1 or 0 to show the exiting links of node u and v.

The closeness centrality shows the node's impact on receiving and sending information to other nodes by calculating the path length between nodes [36]. The closeness centrality c(u) of a node u is defined in Equation (2).

$$c(u) = \sum_v d_{uv} \tag{2}$$

where d_{uv} is a value between 0 and 1 to show the number of the shortest path link from node u to v.

The clustering coefficient C(u) of a node u calculates the link possibility between two nodes based on the total possibility of all links. The clustering coefficient, called cliques, communities, or clusters, is defined in Equation (3).

$$C(u) = \binom{k_u}{2}^{-1} T(u) = \frac{2T(u)}{K_u(k_u-1)} \tag{3}$$

where $K_u(k_u - 1)$ is the maximum possible links in neighbors of u and T(u) is the number of distance triangles with node u.

Network density is a global measure of a network with a range value between 0 and 1 to show the number of links and closeness to a complete network. A dense network is a network with many connections, and a sparse network is a network with few links. The undirected and directed network densities D(G) are defined in Equations (4) and (5).

$$D(G) = \frac{2m}{n(n-1)/2} \tag{4}$$

$$D(G) = \frac{2m}{n(n-1)} \tag{5}$$

where m = |E| is the number of edges and n = possible number of edges.

2.2. Adamic Adar Index

Adamic/Adar (AA) was proposed by Adamic Pepper and Eytan Adar [37] to calculate scores as an index similarity between two web pages. The AA index depresses the common neighbors with the node degree and is defined in Equation (6).

$$ScoreAA(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log|\Gamma(z)|} \tag{6}$$

where $z\epsilon\Gamma(u) \cap \Gamma(v)$ represents list of common neighbors for node pairs of u and v, $\Gamma(z)$ represents the neighbor of each common neighbor's node.

An undirected network is used for the example of calculating predictive links with Adamic Adar, as shown in Fig.1. The undirected network consists nodes "A", "B", "C", "D", and "E", and consists of 5 edges: ("A", "B"), ("A", "D"), ("B", "C"), ("C", "D"), and ("C", "E"), as shown in Fig.1. Later, edges ("A", "C") is a future link. The AA score calculation for edge ("A", "C") is:

- The neighbors of A or $\Gamma(A)$are ["B", "D"]
- The neighbors of C or $\Gamma(C)$are ["B", "D", "E"]
- The common neighbors of A and C or $z\epsilon\Gamma(A) \cap \Gamma(C)$ are ["B", "D"]
- The neighbors of B or $\Gamma(B)$ are ["A", "C"] and $|\Gamma(E)| = 2$
- The neighbors of D or $\Gamma(D)$ are ["A", "C"] and $|\Gamma(D)| = 2$

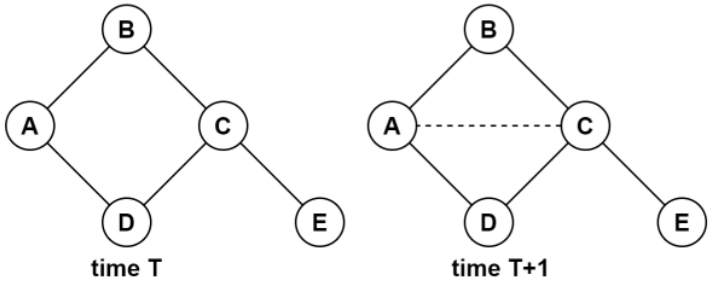


Fig. 1. An undirected network

The final AA score for the edge ("A", "C") is 6.6438. In more detail, the calculation of the AA score is as follows:

$$ScoreAA(A,C) = \sum_{z\epsilon\Gamma(A)\cap\Gamma(C)} \frac{1}{log|\Gamma(z)|}$$

$$ScoreAA(A,C) = \frac{1}{log|\Gamma(B)|} + \frac{1}{log|\Gamma(D)|}$$

$$ScoreAA(A,C) = \frac{1}{\log(2)} + \frac{1}{\log(2)}$$

$$ScoreAA(A,C) = \frac{1}{0{,}301} + \frac{1}{0{,}301}$$

$$ScoreAA(A,C) = 3.3219 + 3.3219$$

$$ScoreAA(A,C) = 6.6438$$

2.3. Extended Adamic Adar index based on the Network Metrics

Most link prediction methods are based on Triadic Closure principles and cannot predict the future link for the node with the cold-start condition. For instance, a simple analogy to this situation is that a new participant comes to an international conference, and this participant does not know anyone at the international conference. Then naturally, this new participant must get acquainted with someone popular and famous who influences many people at international conferences, such as keynote speakers, committees, or moderators. Thus, the new participant can get acquainted with friends of this popular

and famous person. Furthermore, network metrics are used to measure the edge weight and network density for the distance between nodes. The proposed method is defined in Equation (7).

$$Score(u,v) = \begin{cases} \frac{network\ metric(u)+network\ metric(v)}{network\ density^2}, & \text{if path length u and v} = 0 \\ \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\log|\Gamma(z)|}, & \text{otherwise} \end{cases} \quad (7)$$

Nodes with cold-start problems are indicated by conditions where the two nodes are directly connected through neighbors. In other words, the path length between the two nodes equals 0 or has no neighbors. Network metrics examined in this research are extended Adamic Adar based on Degree Centrality (DCAA), extended Adamic based on Closeness Centrality (CloCAA), and extended Adamic Adar based on Clustering Coefficient (CluCAA). The difference from the original Adamic Adar is that this proposed method adds cold-start problem detection based on the path length of the predicted node pairs. If node pairs are known to have cold-start problems, then predictions are made with extended Adamic Adar based on network metrics. The proposed methods are defined in Equations (8), (9), and (10).

$$ScoreDCAA(u,v) = \begin{cases} \frac{d(u)+d(v)}{D(G)^2}, & \text{if path length u and v} = 0 \\ \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\log|\Gamma(z)|}, & \text{otherwise} \end{cases} \quad (8)$$

$$ScoreCloCAA(u,v) = \begin{cases} \frac{c(u)+c(v)}{D(G)^2}, & \text{if path length u and v} = 0 \\ \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\log|\Gamma(z)|}, & \text{otherwise} \end{cases} \quad (9)$$

$$ScoreCluCAA(u,v) = \begin{cases} \frac{C(u)+C(v)}{D(G)^2}, & \text{if path length u and v} = 0 \\ \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\log|\Gamma(z)|}, & \text{otherwise} \end{cases} \quad (10)$$

The Adamic Adar index is chosen for nodes with no cold-start conditions to calculate predictions for future links due to the results of the Liben-Nowell and Kleinberg experiments that show that Adamic Adar is at least as good as other common neighbor predictors [38]. Besides, Adamic Adar considers the common neighbor and the degree of the common neighbor. Fig. 2 shows the proposed extended Adamic Adar based on network metrics: network density, degree of centrality, closeness centrality, and clustering coefficient are extracted from the examined graph. The path length calculates the predicted node pairs based on Eq. (7), and the extended Adamic Adar applies in the path length is zero. The similarity score is calculated based on the Adamic Adar for the path length is more than zero.

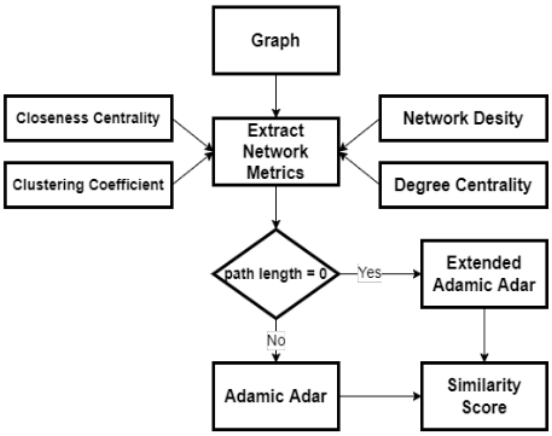


Fig. 2. The proposed extended Adamic Adar

The pseudo-code in Algorithm 1 shows the experiments conducted to implement the proposed method. The experiments are conducted using NetworkX 2.4 and Python 3.6. The algorithm's input is four real-world network datasets, three proposed methods, and nine existing local similarity-based methods as the baseline. Several outputs of the algorithm are the similarity score of each examined method and measurement results, namely, the AUC Score and ROC Curve show in Fig.3.

Algorithm 1: Pseudo-code Link Prediction using Similarity-based Methods
INPUT:
Edges list of edges u and v
OUTPUT:
Score similarity score
ALGORITHM:

```
u = node u
v = node v
d(u,v) = path length u and v
D(G) = Network density using Eq 5
Function AA
    AA = Adamic adar using Eq 6
Function CLoCCA
    If d(u,v) == 0
        c(u) = closeness centrality node u using Eq 2
        c(v) = closeness centrality node v using Eq 2
        CLoCCA = c(u)+c(v) / D(G)^2
    else
        CLoCCA = call function AA
Function CLuCCA
    If d(u,v) == 0
If Score > 0
        C(u) = Clustering coefficient node u using Eq 3
        C(v) = Clustering coefficient node v using Eq 3
        CLuCCA = C(u)+C(v) / D(G)^2
    else
        CLuCCA =call function AA
Function DCCA
    If d(u,v) == 0
        d(u) = degree centrality node u using Eq 1
        d(v) = degree centrality node u using Eq 1
        DCAA = d(u)+d(v) / D(G)^2
    else
        DCAA = call function AA
```

Fig. 3. AUC Score and ROC Curve

2.4. Experiment Design

The experiment is conducted in three stages, i.e., graph generation, score computation, and result measurements, as shown in Fig. 2. In the first stage, the graph generation is conducted by creating the graph dataset from the list of edges in the files dataset. Furthermore, split the edge into edges train and edges test using scikit-learn [39], with the test size of each dataset is 10%. Later the graph train is created from the list of edges train connects and a list of all nodes from the graph dataset. The confusion matrix needs at least two classes to measure the prediction results: connected and not connected. The edges test not connect obtained using scikit-learn based on the graph train list of not connected edges. The number of test sizes in the second train test split depends on the number of edges test connect. Lastly, the edges test is called connect, and the edges test is called not connect as the edge sample is

merged into the edge sample. The edge sample is the set of test data to compute the similarity score of the proposed and benchmark methods. Experiment design show in Fig. 4.

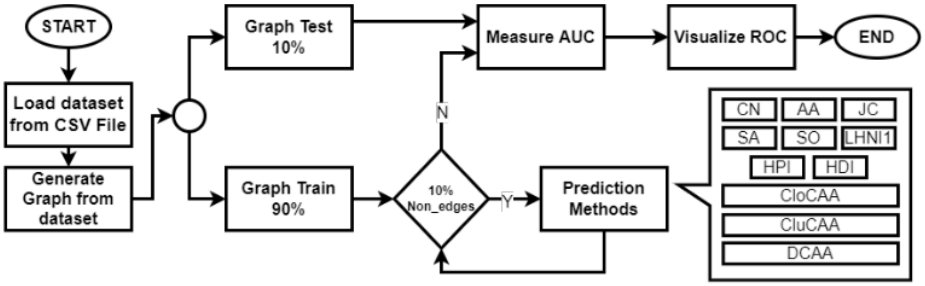


Fig. 4. Experiment Design

The second stage is similarity score computation to determine the score of each node pairs prediction. Every proposed and benchmark method is computed to get a prediction score and label actual and prediction. A pair prediction node is labeled actual true or false based on the edge sample label, whereas a pair prediction node is labeled actual true if the prediction score is more than 0 and vice versa. The last stage is the measurements. The AUC score and ROC curve are measurement methods for the experiments.

3. Results and Discussion

3.1. Benchmark Methods

The proposed method is compared with local similarity-based or node neighborhood methods for link prediction. The local similarity-based methods are chosen as benchmark methods because most researchers consider the similarity-based method as an appropriate research approach, as it has a relatively low computational level and does not require conducting complicated network analysis stages. Most researchers also use local similarity-based methods as benchmark methods to compare the proposed methods. The benchmark methods are defined in Table 1.

Table 1. Benchmark Methods

#	Baseline methods	Formula	Ref
1.	Common Neighbors (CN)	$ScoreCN(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert$	[40]
2.	Adamic/Adar (AA)	$ScoreAA(u,v) = \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\log \lvert\Gamma(z)\rvert}$	[37]
3.	Resource Allocation (RA)	$ScoreRA(u,v) = \sum_{z\epsilon\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$	[41]
4.	Jaccard Coefficient (JC)	$ScoreJC(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\lvert\Gamma(u) \cup \Gamma(v)\rvert}$	[42]
5.	Salton Cosine Similarity (SA)	$ScoreSA(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\rvert * \lvert\Gamma(v)\rvert}}$	[43]
6.	Sørensen Index (SO)	$ScoreSO(u,v) = \frac{2 * \lvert\Gamma(u) \cap \Gamma(v)\rvert}{\Gamma(u) + \Gamma(v)}$	[44]
7.	Leicht-Holme-Nerman-1 Index (LHNI1)	$ScoreLHN1(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\lvert\Gamma(u)\rvert * \lvert\Gamma(v)\rvert}$	[45]
8.	Hub Promoted Index (HPI)	$ScoreHPI(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\min(\Gamma(u), \Gamma(v))}$	[46] [47]
9.	Hub Depressed Index (HDI)	$ScoreHDI(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\max(\Gamma(u), \Gamma(v))}$	[46] [47]

3.2. Datasets

The experiment was conducted using four real-world networks dataset from several network categories to compare the proposed cold-start link prediction. The datasets are also downloaded on http://snap.stanford.edu/data/index.html [48], http://networkrepository.com [49], and http://konect.uni-koblenz.de/networks [50]. Dataset information is shown as follows:

- Power [51] is a network that contains the Western States Power Grid topology of the United States.
- Firm-hi-tech [49] is a network of hi-tech firms.
- Wiki-vote [52] is Wikipedia voting data representing Wikipedia users, and a directed edge represents the user voted.
- Adolescent [53] is a network from a survey that asks the student to select their five best female and male friends.

The basic topological feature of dataset consists of several information. The Net, |V|, |E|, K_n, NE, ⟨k⟩, C, and ρ are network categories, nodes, edges, complete graph, unobserved node pairs, average degree, average cluster coefficient, and network density, respectively. The datasets are composed of five network categories: infrastructure networks and social networks, as shown in Table 2.

Table 2. Datasets information

Dataset	Net	|V|	|E|	K_n	NE	⟨k⟩	C	ρ
Power	Inf	4941	6594	12204270	12197676	2.669	0.080	0.0005
Firm-hi-tech	Soc	33	91	528	437	5.515	0.453	0.1723
Wiki-vote	Soc	889	2914	394716	391802	6.556	0.153	0.0074
Adolescent	Soc	2539	10455	3221991	3211536	8.236	0.147	0.0032

One characteristic of the cold-start problem is the path length of the predicted node pair is equal to zero. The path length data is obtained from the testing data, and each real-world network dataset appears to have node pairs with a cold-start problem. Table 3 shows the path length of the testing data for experiments.

Table 3. Training and testing data for experiments

Dataset	Training	Testing			Path Length Testing			
		Con	Not Con	Total	0	2	>2	Σ
Power	5934	660	660	1320	311	119	890	1320
Firm-hi-tech	81	10	10	20	1	12	7	20
Wiki-vote	2622	292	292	584	41	193	350	584
Adolescent	9409	1046	1046	2092	13	585	1494	2092

- The experiment results of the proposed and benchmark methods are reported in Tables 4-6. Common neighbors-based methods are a collection of benchmark methods based on common neighbors, including CN, AA, RA, JC, SA, SO, LHNI1, HPI and HDI. A closeness-degree-based method is proposed based on closeness centrality and degree centrality. The cluster coefficient-based method is a proposed method based on the clustering coefficient.

3.3. Number of Prediction Ratio Results

Furthermore, the ratio is the ratio between the number of path lengths predicted with the path length of the testing data. The Centrality and Degree-based Predict collection methods in this comparative data can better predict as many as 99% of node pairs. This result shows that centrality and degree-based prediction outperform the baseline and cluster coefficient-based methods in predicting node pairs with a cold-start problem. More in-depth analysis is conducted to find out the cause of some of the ratios is not 100% achieved by looking at the detailed data. The analysis results show that two nodes are isolated from relationships with other nodes in a node pair. So that the impact on no attraction

between the two nodes. This isolated node pair also occurs based on the clustering coefficient in the proposed method. In addition, the cluster coefficient-based method shows that some nodes fail to be an influencer to nodes with cold-start conditions. Even though this attraction node is connected to other nodes, the ability to solve cold-start problems is not as good as closeness and degree-based methods.

The results of the prediction ratio in Table 4 mean that the extension to Adamic Adar can reduce cold-start problems. This result is because the proposed method can predict node pairs with cold-start problems to 46% and 99% for the cluster coefficient-based method and the closeness and degree-based methods, respectively. The cluster coefficient-based method does not get as good results as the closeness and degree-based method because the cluster coefficient still relies on community information from the network. Furthermore, measurements with AUC were conducted to examine the performance of the proposed method.

Table 4. The number of prediction ratios based on path length pair node prediction

Dataset	Common Neighbors-based Methods				Ratio CN	Closeness and Degree-based Methods				Ratio C&D	Cluster Coefficient-based Method				Ratio CC
	0	2	>2	Σ		0	2	>2	Σ		0	2	>2	Σ	
Power	0	119	0	119	0%	305	119	0	424	98%	44	119	0	163	14%
Firm-hi-tech	0	12	0	12	0%	1	12	0	23	100%	1	12	0	13	100%
Wiki-vote	0	193	0	193	0%	41	193	0	234	100%	26	193	0	219	63%
Adolescent	0	585	0	585	0%	13	585	0	598	100%	6	585	0	591	46%
Average Ratio					0%					99%					46%

3.4. Some Common Mistakes

The experiment used the AUC score to evaluate the proposed cold-start link prediction [54]. AUC score is calculated by randomly selecting one of the links tested and comparing it to a link that does not exist randomly [55]. AUC is also used to measure link prediction performance [56]. The ideal AUC score is 1 and is defined in Equation (11).

$$AUC = \frac{(n_1+0.5*n_2)}{n(n-1)/2} \quad (11)$$

where n is the number of independent comparisons, n_1 is the number of missing links with a higher score than non-existent links, and n_2 is the number of missing links that have equal scores with non-existent links [57]. Besides the AUC, ROC is a visual analysis based on the calculation of the AUC and is used to get information about different loss matrices of two error types and to get the classifier behavior under different loss matrices [58]. AUC measurement results show that the proposed method outperforms all real-world network benchmark datasets, as shown in Table 4. The AUC score is compared to the information dataset, as shown in Table 1, and the path length of the testing data, as shown in Table 2. The AUC score indicates that the proposed method is more suitable for networks with the number of isolated nodes and low values at the average degree, clustering coefficient, and network density, besides being able to solve the cold-start problem. See experiment results for Power, Firm-hi-tech, Wiki-votes, and Adolescent datasets, as shown in Tables 5 and Table 6.

Table 5. AUC score measurement results

Dataset	CN	JA	AA	RA	SA	SO
Power	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902
Firm-hi-tech	0.8750	0.8500	0.8950	0.8950	0.8550	0.8500
Wiki-vote	0.7815	0.7752	0.7826	0.7822	0.7753	0.7752
Adolescent	0.7625	0.7624	0.7628	0.7626	0.7625	0.7624
Avg AUC	0.7483	0.7402	0.7525	0.7528	0.7416	0.7402

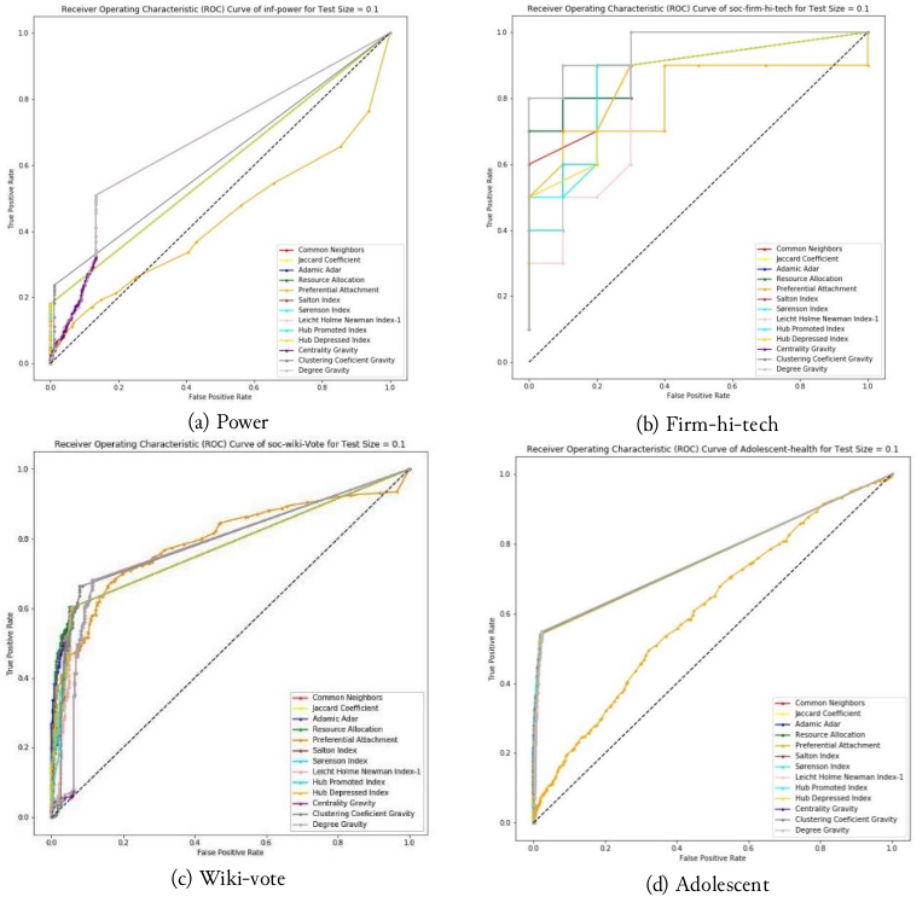
Table 6. AUC score measurement results (continue)

Dataset	LHNI1	HPI	HDI	CloCAA	CluCAA	DCAA
Power	0.5902	0.5902	0.5902	**0.6753**	0.6119	0.6749
Firm-hi-tech	0.8000	0.8650	0.8550	**0.9600**	**0.9600**	**0.9600**
Wiki-vote	0.7687	0.7742	0.7756	0.7756	**0.7930**	0.7758
Adolescent	0.7616	0.7625	0.7623	**0.7642**	0.7632	**0.7642**
Avg AUC	0.7292	0.7448	0.7406	**0.7938**	0.7820	**0.7938**

Based on the AUC score results and comparing it with the prediction ratio results, the proposed method can outperform the benchmark method's performance achievements, in addition to solving the cold-start problem of the predicted node pairs. Although none of the proposed methods has consistently superior results for all datasets. This finding is essential for future research because the proposed method is a new link prediction method

3.5. Receiver Operating Characteristics (ROC) Curve

The experiment results also draw the ROC curve's performance to evaluate the proposed methods and other benchmarks further, as shown in Fig. 5. The ROC curve is presented to visualize the AUC results in the previous Table 4 and Table 5. The ROC curve shows that the proposed methods can solve sparse networks that contain nodes with cold-start problems.



(a) Power (b) Firm-hi-tech

(c) Wiki-vote (d) Adolescent

Fig. 5. Visualizations of ROC Curve

The addition of filtering to detect node pairs with the cold-start problem by extending the original Adamic Adar represents a new difference in obtaining essential results from the findings. Therefore, extended Adamic Adar is presented as a novelty and originality in the link prediction research area. The results of this research are promising for future research due to the proposed method improves the prediction performance and solves the cold-start problem.

4. Conclusion

This research has proposed three novel methods by extending the Adamic Adar index based on network metrics. This proposed method is called the extended Adamic based on Degree Centrality (DCAA), Closeness Centrality (CloCAA), and Clustering Coefficient (CluCAA). The AUC value achieved by the proposed method is up to 0.9600. Furthermore, the proposed method based on closeness and degree can predict node pairs with cold-start problems up to a ratio of 99%. The experiment results demonstrate that DGAA and CloCAA outperform the benchmark methods and can predict node pairs with a cold-start problem better than the original Adamic Adar. However, the drawback of the proposed method is that the prediction formula is more complex than the original Adamic Adar because the proposed method has a condition check beforehand to find out the predicted node pairs in the cold-start problem. If the node pair is in a cold-start problem, the predictor uses its extension function, and vice versa. The predictor uses the original Adamic Adar. The proposed methods (DCAA, CloCAA, CluCAA) are more suitable for networks with high isolated nodes and low values at the average degree, clustering coefficient, and network density. In future research, the proposed method can be combined with machine learning and ensemble learning approaches by examining more varied datasets from several domains such as social networks, terrorist networks, co-authorship networks, and others.

Acknowledgment

The authors thanks to Universiti Kebangsaan Malaysia and Universitas Ahmad Dahlan, Indonesia for support the research.

Declarations

Author contribution. Herman Yuliansyah: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – the original draft, Review & Editing, and Visualization. Zulaiha Ali Othman: Methodology, Validation, Formal analysis, Writing - Review & Editing, Visualization, Supervision, and Funding acquisition. Azuraliza Abu Bakar: Methodology, Validation, Writing - Review & Editing, Visualization, and supervision..

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] F. Tahmasebi, M. Meghdadi, S. Ahmadian, and K. Valiallahi, "A hybrid recommendation system based on profile expansion technique to alleviate cold start problem," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 2339–2354, Jan. 2021, doi: 10.1007/s11042-020-09768-8.

[2] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data," *Expert Syst. Appl.*, vol. 149, p. 113248, Jul. 2020, doi: 10.1016/j.eswa.2020.113248.

[3] Y. Zhu *et al.*, "Addressing the Item Cold-Start Problem by Attribute-Driven Active Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 631–644, Apr. 2020, doi: 10.1109/TKDE.2019.2891530.

[4] J. Misztal-Radecka, B. Indurkhya, and A. Smywiński-Pohl, "Meta-User2Vec model for addressing the user and item cold-start problem in recommender systems," *User Model. User-adapt. Interact.*, vol. 31, no. 2, pp. 261–286, Apr. 2021, doi: 10.1007/s11257-020-09282-4.

[5] D. Cai, S. Qian, Q. Fang, J. Hu, and C. Xu, "User Cold-start Recommendation via Inductive Heterogeneous

Graph Neural Network," *ACM Trans. Inf. Syst.*, Sep. 2022, doi: 10.1145/3560487.

[6] K. Pliakos, S.-H. Joo, J. Y. Park, F. Cornillie, C. Vens, and W. Van den Noortgate, "Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems," *Comput. Educ.*, vol. 137, pp. 91–103, Aug. 2019, doi: 10.1016/j.compedu.2019.04.009.

[7] M. Vahidi Farashah, A. Etebarian, R. Azmi, and R. Ebrahimzadeh Dastjerdi, "A hybrid recommender system based-on link prediction for movie baskets analysis," *J. Big Data*, vol. 8, no. 1, p. 32, Dec. 2021, doi: 10.1186/s40537-021-00422-0.

[8] J. Feng, Z. Xia, X. Feng, and J. Peng, "RBPR: A hybrid model for the new user cold start problem in recommender systems," *Knowledge-Based Syst.*, vol. 214, p. 106732, Feb. 2021, doi: 10.1016/j.knosys.2020.106732.

[9] J. Moon, J. Kim, P. Kang, and E. Hwang, "Solving the Cold-Start Problem in Short-Term Load Forecasting Using Tree-Based Methods," *Energies*, vol. 13, no. 4, p. 886, Feb. 2020, doi: 10.3390/en13040886.

[10] Z. Li *et al.*, "HML4Rec: Hierarchical meta-learning for cold-start recommendation in flash sale e-commerce," *Knowledge-Based Syst.*, vol. 255, p. 109674, Nov. 2022, doi: 10.1016/j.knosys.2022.109674.

[11] M. Tang and W. Wang, "Cold-start Link Prediction Integrating Community Information via Multi-Nonnegative Matrix Factorization," *Chaos, Solitons & Fractals*, vol. 162, p. 112421, Sep. 2022, doi: 10.1016/j.chaos.2022.112421.

[12] M. Tang, W. Yu, X. Li, X. Chen, W. Wang, and Z. Liu, "Cold-start Link Prediction via Weighted Symmetric Nonnegative Matrix Factorization with Graph Regularization," *Comput. Syst. Sci. Eng.*, vol. 43, no. 3, pp. 1069–1084, 2022, doi: 10.32604/csse.2022.028841.

[13] S. Wu, Q. Zhang, C. Xue, and X. Liao, "Cold-start Link Prediction in Multi-relational Networks based on Network Dependence Analysis," *Phys. A Stat. Mech. its Appl.*, vol. 515, pp. 558–565, 2019, doi: 10.1016/j.physa.2018.09.082.

[14] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "Taxonomy of link prediction for social network analysis: a review," *IEEE Access*, vol. 8, pp. 183470–183487, 2020, doi: 10.1109/ACCESS.2020.3029122.

[15] K. Li, L. Tu, and L. Chai, "Ensemble-model-based Link Prediction of Complex Networks," *Comput. Networks*, vol. 166, p. 106978, 2020, doi: 10.1016/j.comnet.2019.106978.

[16] J. Liu *et al.*, "Collaborative linear manifold learning for link prediction in heterogeneous networks," *Inf. Sci. (Ny).*, vol. 511, pp. 297–308, Feb. 2020, doi: 10.1016/j.ins.2019.09.054.

[17] G. Chen, C. Xu, J. Wang, J. Feng, and J. Feng, "Graph Regularization Weighted Nonnegative Matrix Factorization for Link Prediction in Weighted Complex Network," *Neurocomputing*, vol. 369, pp. 50–60, 2019, doi: 10.1016/j.neucom.2019.08.068.

[18] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 393–402. doi: 10.1145/1835804.1835855.

[19] P. Wang, B. W. Xu, Y. R. Wu, and X. Y. Zhou, "Link prediction in social networks: the state-of-the-art," *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, 2015, doi: 10.1007/s11432-014-5237-y.

[20] S. Haghani and M. R. Keyvanpour, "A Systemic Analysis of Link Prediction in Social Network," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1–35, 2017, doi: 10.1007/s10462-017-9590-2.

[21] H. Huang, Y. Dong, J. Tang, H. Yang, N. V. Chawla, and X. Fu, "Will Triadic Closure Strengthen Ties in Social Networks?," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 3, pp. 1–25, Jun. 2018, doi: 10.1145/3154399.

[22] A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä, "Cumulative effects of triadic closure and homophily in social networks," *Sci. Adv.*, vol. 6, no. 19, May 2020, doi: 10.1126/sciadv.aax7310.

[23] A. Aleta, M. Tuninetti, D. Paolotti, Y. Moreno, and M. Starnini, "Link prediction in multiplex networks via triadic closure," *Phys. Rev. Res.*, vol. 2, no. 4, p. 042029, Nov. 2020, doi: 10.1103/PhysRevResearch.2.042029.

[24] L. Ge and A. Zhang, "Pseudo Cold Start Link Prediction with Multiple Sources in Social Networks," in *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, 2012, pp. 768–779. doi: 10.1137/1.9781611972825.66.

[25] M. Yan, J. Sang, T. Mei, and C. Xu, "Friend Transfer: Cold-start Friend Recommendation with Cross-platform Transfer Learning of Social Knowledge," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6, doi: 10.1109/ICME.2013.6607510.

[26] J. Zhang, X. Kong, and P. S. Yu, "Predicting Social Links for New Users Across Aligned Heterogeneous Social Networks," in *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 1289–1294. doi: 10.1109/ICDM.2013.134.

[27] V. A. Rohani, Z. M. Kasirun, S. Kumar, and S. Shamshirband, "An Affective Recommender Algorithm for Cold-start Problem in Academic Social Networks," *Math. Probl. Eng.*, vol. 2014, 2014, doi: 10.1155/2014/123726.

[28] X. Han, L. Wang, S. N. Han, C. Chen, N. Crespi, and R. Farahbakhsh, "Link Prediction for New Users in Social Networks," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 1250–1255. doi: 10.1109/ICC.2015.7248494.

[29] Z. Wang, J. Liang, R. Li, and Y. Qian, "An Approach to Cold-start Link Prediction: Establishing Connections between Non-topological and Topological Information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2857–2870, 2016, doi: 10.1109/TKDE.2016.2597823.

[30] J. Zhu *et al.*, "CHRS: Cold Start Recommendation Across Multiple Heterogeneous Information Networks," *IEEE Access*, vol. 5, no. 8, pp. 15283–15299, 2017, doi: 10.1109/ACCESS.2017.2726339.

[31] S. Wu, Q. Zhang, and M. Wu, "Cold-start Link Prediction in Multi-relational Networks," *Phys. Lett. A*, vol. 381, no. 39, pp. 3405–3408, 2017, doi: 10.1016/j.physleta.2017.08.046.

[32] L. Xu, X. Wei, J. Cao, and P. S. Yu, "On Learning Community-specific Similarity Metrics for Cold-start Link Prediction," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489683.

[33] L. Xu, X. Wei, J. Cao, and P. S. Yu, "On Learning Mixed Community-Specific Similarity Metrics for Cold-start Link Prediction," in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 2017, pp. 861–862. doi: 10.1145/3041021.3054269.

[34] H. Nassar, A. R. Benson, and D. F. Gleich, "Neighborhood and PageRank methods for pairwise link prediction," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, p. 63, Dec. 2020, doi: 10.1007/s13278-020-00671-6.

[35] S. Liu, X. Ji, C. Liu, and Y. Bai, "Similarity indices based on link weight assignment for link prediction of unweighted complex networks," *Int. J. Mod. Phys. B*, vol. 31, no. 02, p. 1650254, Jan. 2017, doi: 10.1142/S0217979216502544.

[36] M. Z. Al-Taie and S. Kadry, *Python for Graph and Network Analysis*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-53004-8.

[37] L. A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Soc. Networks*, vol. 25, no. 3, pp. 211–230, Jul. 2003, doi: 10.1016/S0378-8733(03)00009-1.

[38] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007, doi: 10.1145/956863.956972.

[39] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, available at: Google Scholar.

[40] M. E. J. Newman, "Clustering and Preferential Attachment in Growing Networks," *Phys. Rev. E*, vol. 64, no. 2, p. 025102, Jul. 2001, doi: 10.1103/PhysRevE.64.025102.

[41] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting Missing Links via Local Information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009, doi: 10.1140/epjb/e2009-00335-8.

[42] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901, doi: 10.5169/seals-266450.

[43] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, pp. 528 , 1983,

available at: Google Scholar

[44] T. J. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *I kommission hos E. Munksgaard*, vol. 5, pp. 34, 1948, available at: Books Google

[45] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex Similarity in Networks," *Phys. Rev. E*, vol. 73, no. 2, p. 026120, Feb. 2006, doi: 10.1103/PhysRevE.73.026120.

[46] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks," *Science (80-.).*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002, doi: 10.1126/science.1073374.

[47] L. L. Linyuan and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A Stat. Mech. its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011, doi: 10.1016/j.physa.2010.11.027.

[48] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection." Jun. 2014. Available at: https://snap.stanford.edu/data/

[49] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," 2015. [Online]. Available: http://networkrepository.com

[50] J. Kunegis, "Konect: the koblenz network collection," in *International Conferenceon World Wide Web*, 2014, pp. 1343–1350, doi: 10.1145/2487788.2488173.

[51] D. J. Watts and S. H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998, doi: 10.1038/30918.

[52] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed Networks in Social Media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1361–1370, doi: 10.1145/1753326.1753532.

[53] J. Moody, "Peer influence groups: identifying dense clusters in large networks," *Soc. Networks*, vol. 23, no. 4, pp. 261–283, Oct. 2001, doi: 10.1016/S0378-8733(01)00042-9.

[54] B. Pandey, P. K. Bhanodia, A. Khamparia, and D. K. Pandey, "A Comprehensive Survey of Edge Prediction in Social Networks: Techniques, Parameters and Challenges," *Expert Syst. Appl.*, vol. 124, pp. 164–181, 2019, doi: 10.1016/j.eswa.2019.01.040.

[55] X. Xu *et al.*, "Distributed Temporal Link Prediction Algorithm based on Label Propagation," *Futur. Gener. Comput. Syst.*, vol. 93, pp. 627–636, 2019, doi: 10.1016/j.future.2018.10.056.

[56] Z. Wang, J. Liang, and R. Li, "A Fusion Probability Matrix Factorization Framework for Link Prediction," *Knowledge-Based Syst.*, vol. 159, no. June, pp. 72–85, 2018, doi: 10.1016/j.knosys.2018.06.005.

[57] P. K. Sharma, S. Rathore, and J. H. Park, "Multilevel learning based modeling for link prediction and users' consumption preference in Online Social Networks," *Futur. Gener. Comput. Syst.*, vol. 93, pp. 952–961, 2019, doi: 10.1016/j.future.2017.08.031.

[58] E. Alpaydin, Introduction to machine learning, 3rd ed. London: The MIT Press, pp. 640, August, 2014, available at :https://mitpress.mit.edu/9780262028189/introduction-to-machine-learning/

HASIL CEK_Extending adamic

ORIGINALITY REPORT

7%	4%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to SDM Universitas Gadjah Mada Student Paper	3%
2	Minghu Tang, Wei Yu, Xiaoming Li, Xue Chen, Wenjun Wang, Zhen Liu. "Cold-Start Link Prediction via Weighted Symmetric Nonnegative Matrix Factorization with Graph Regularization", Computer Systems Science and Engineering, 2022 Publication	2%
3	Olorunsola Stephen Olufunso, Abraham Eseoghene Evwiekpaefe, Martins Ekata Irhebhude. "Gender recognition based fingerprints using dynamic horizontal voting ensemble deep learning", International Journal of Advances in Intelligent Informatics, 2022 Publication	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		