

Link Prediction in Social Networks based on Spectral Clustering using k -medoids and Landmark

Asmaa M. Mahmoud
Al-Azhar University
Faculty of Science, Cairo, Egypt

Lamiaa M. El Bakrawy
Al-Azhar University
Faculty of Science, Cairo, Egypt

Neveen I. Ghali
Al-Azhar University
Faculty of Science, Cairo, Egypt

ABSTRACT

Social networks have become popular because they provide many services and advantages to its users like helping them to connect with new people, share content and share opinions with like-minded people. To conclude which new interactions between members are likely to occur in the near future? This problem defined as the link prediction problem. The traditional methods of link prediction are based on path features and graph features but few consider clustering information. So two link prediction methods based on spectral clustering using k -medoids and landmark are proposed. The first method uses k -medoids to cluster nodes of the graph based on eigenvectors that obtained from the normalized Laplacian matrix. While the second method selects a subset of data points as the landmarks and represents the original data points as the linear combinations of these landmarks. Experimental results demonstrate that our methods achieve high accuracy compared with LSk -means method (Link prediction using Spectral clustering by k -means).

General Terms

Link prediction, landmark based spectral clustering, k -medoids algorithm

Keywords

Social networks, link prediction, spectral clustering, landmark, k -medoids, sparse coding.

1. INTRODUCTION

Many real world systems can be described as networks with nodes (vertices) acting as individuals and links (edges) refer to the relationships between them [1][2][3]. Social network is dynamically important, because new edges and vertices are added to the graph over time. To understand the association between two specific nodes, link prediction is discussed where link prediction focus on predicting the existent edge between two nodes that will probably occur in the near future [4][5].

Computational intelligence (CI) is a method that involves computing that provides a system with an capability of learning and dealing with new situations [6]. CI is including neural networks, neuro fuzzy systems, evolutionary algorithms (EA), genetic algorithms (GA), swarm intelligence, classification and clustering [7][8][9][10]. Actually, clustering information plays a significant role in link prediction where it connects the sparse nodes and other

communities to form dense communities [11]. Many methods of link prediction are based on path features and graph features but few consider clustering information which is very important for link prediction.

Xu et al. in [12] studied the relation between the network structure and the precision of link prediction methods and showed that when the clustering structure of the network grows, the accuracy of link prediction methods improves. Sucheta et al. in [13] also showed that the inclusion of clustering information improved the accuracy of similarity based link prediction methods.

Recently, many methods for link prediction depend on similarity between vertices have been proposed since similar vertices likely share the same relations. Abbassi et al. in [14] promoted two main ideas for a recommender system depending on the link structure of the weblog graph. They examined these two ideas on directed and undirected graphs. One idea was to use the personalized PageRank vector to get the relevance of nodes on a graph and apply the local truncated random walk algorithms to approximate this vector. Another idea was to introduce a metric based on the spectral methods that can be used to rank the relevant nodes in a network. Hoseini et al. in [15] used the well-known DBLP dataset (Digital Bibliography and Library Project) and tried to find some skills that are connected to an author that they didn't know before. They clustered authors and skills using Spectral Graph Clustering algorithm, then simultaneously obtained author and skill clusters via Bipartite Graph (Bi-graph) Spectral Co-clustering approach, and then generated predictions based on the outputs of clustering and co-clustering steps.

Kunegis et al. in [16] showed a link prediction algorithm based on the extrapolation of a network's spectral evolution. This link prediction method generalizes several common graph kernels that they can express as spectral transformations. In contrast to these graph kernels, the spectral extrapolation algorithm didn't make assumptions about specific growth patterns beyond the spectral evolution model. Li et al. in [3] showed that the clustering results provided the essential information for link prediction, and these vertices common neighbors may play different roles depending on if they belong to the same cluster. They proposed a link prediction method based on previous assumption. Carlos et al. in [1] proposed a new measure (WIC) for link prediction between a pair of vertices considering the sets of their within-cluster (W) and inter-cluster (IC) common neighbors. Also, they proposed a set of measures, referred to as W forms, using only the set given by the within-cluster common neighbors instead of using the set of all common neighbors as usually considered in the basic local similarity measures.

Huang et al. in [17] explored the connection between link prediction and graph topology. The focus was exclusively on the predictive value of the clustering coefficient measure. The standard clustering coefficient measure was generalized to capture higher-order clustering tendencies. The proposed framework consisted of a cycle formation link probability model, a procedure for estimating model parameters based on the generalized clustering coefficients, and model-based link prediction generation. Kim et al. in [18] proposed LPCSP (Link Prediction inferred from Cluster Similarity and cluster Power), a novel link prediction method which exploited the generalized cluster information containing cluster relations and cluster evolution information. Wang et al. in [19] proposed a link prediction approach based on a new multiple independent Bernoulli matrix completion model to obtain an accurate multiple-independent-Bernoulli-distribution model of the topological feature space, this approach adjusted the sampling of the adjacency matrix of the network using the clustering information in the node feature space.

In this paper, spectral clustering using k -medoids and landmark are applied for link prediction in social networks to improve prediction accuracy as will be shown experimentally later. In the proposed first method k -medoids is used to cluster nodes of a graph then using this information to detecting similarity between nodes of a graph and recommend friends according to their weights. In the proposed second method landmark based spectral clustering is implemented to design an efficient way for graph construction and Laplacian matrix eigen-decomposition.

The remainder of this paper is organized as follows: Introduction of k -means clustering algorithm, k -medoids clustering algorithm, preliminaries in graphs, spectral clustering and landmark based spectral clustering are respectively introduced in Section II. The proposed methods are presented in Section III. Results of the experiments and comparison are shown and discussed in Section IV. Finally, Section V concludes this paper.

2. AN OVERVIEW

2.1 k -means Clustering Algorithm

Clustering is to group objects into classes so that objects within a cluster are similar in comparison to one another, but aren't similar to objects in other clusters. the algorithm can be summarized as follows:

Number of clusters k is determined and the centroid of these clusters is assumed. Random objects are taken as the initial centroids then assign each object to the group that has the closest centroid. When all objects are assigned, recalculate the positions of the k centroids based on the mean value of the items in the cluster and repeat the final two steps until the centroids no change [20][21].

2.2 k -medoids Clustering Algorithm

k -medoids Algorithm which is also known as Partition Around Medoids (PAM) is developed by Kaufman and Rousseuw in 1987 [22][23][24]. k -medoids algorithm is comparatively robust than k -means particularly in the context of outliers and noise because the k -means algorithm is sensitive to outliers. Because a mean is sensitive to the outliers, i.e., if you look at an ages of students in one class, if you adding another very high age, the average age of the whole class shifts alot. So instead of taking the mean value of object in a cluster as a centroid the most centrally located object is used in the cluster as a medoids in k -medoids [22][24]. k -medoids

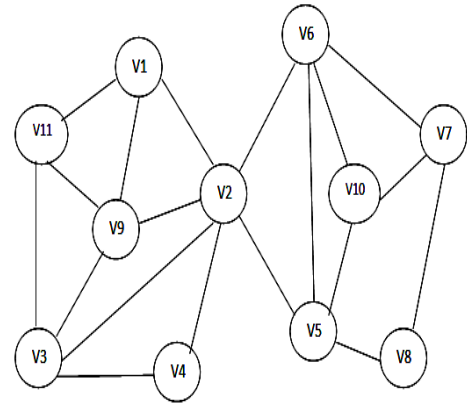


Figure 1. Network example

is a classical partitioning technique of clustering that clusters the data set of m objects into k number of clusters [24]. The user gives the number of clusters required. This algorithm depends on the minimizing the sum of dissimilarities (distance) between each object and its medoid.

The algorithm randomly chooses the k objects in dataset m as initial representative objects called medoids. A medoid can be defined as the object of a cluster, which has minimal dissimilarity to all the objects in the cluster. Then for all objects in the dataset, it assigns each object to the nearest cluster depending upon the objects distance to the cluster medoid. After every assignment of a data object to particular cluster, non-medoid object is selected random. Then calculate the total cost of swapping for non-medoid object if it is negative then swap initial medoid with the new one. Repeat final two steps until there is no change in the medoid [21][25].

2.3 Preliminaries In Graphs

A graph $G = (V, E)$ is a set of vertices V and edges E . In Figure 1 vertices denote friends and edges denote interactions between them.

In this paper, G will always be an undirected graph. The adjacency matrix A of an undirected graph G is a square matrix and symmetric where $A \geq 0$ represents a measure of the similarity between data points with indexes i and j . An element in the adjacency matrix has value equal to 1 if pairs of vertices are adjacent and 0 otherwise. i.e. the value of edge that connect nodes (v_1, v_{11}) is equal to 1, while the element (v_1, v_7) will have value 0. Notice that each element in the diagonal of matrix A is equal to zero to depict that a node is not connected to itself. In case of a large graph A can be characterized by high dimensionality [2].

2.4 Spectral Clustering

In recent years, spectral clustering has become one of the most popular modern clustering algorithms. It is simple to implement, can be solved efficiently by standard linear algebra software, and often outperforms traditional clustering algorithms such as the k -means algorithm [26].

Given a set of data points $x_1, x_2, \dots, x_n \in R^m$, spectral clustering first constructs an undirected graph $G = (V, E)$ represented by its adjacency matrix. Then the lablacian matrix is calculated by

equation

$$L = D^{-\frac{1}{2}} \times (D - A) \times D^{-\frac{1}{2}} \quad (1)$$

where D is the degree matrix and A is the adjacency matrix of graph G [2]. The goal of spectral clustering is to partition the data into k disjoint classes such that each x_i belongs to one and only one class. The general approach to spectral clustering is to use a standard clustering methods (like k -means) on relevant eigenvectors of a Laplacian matrix L [27].

The normalized Laplacian matrix L is positive semi-definite and has n non-negative real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \lambda_k$. Moreover, if the value of eigenvalue equal to 0 this means that there is a connected components in a graph. Thus, spectral clustering is more elastic than k -means, in capturing the non-connected components of a graph [2].

2.5 Landmark Based Spectral Clustering

Landmark based Spectral Clustering (LSC) depends on the recent progress of sparse coding [28].

Sparse coding is a matrix factorization technique [29] [30] which tries to rebuild input vectors using a linear combination of basis vectors with sparse coefficients [31]. Let $X = [x_1, \dots, x_n] \in R^{m \times n}$ be the data matrix (each column is an input vector), the matrix factorization method find two matrices $U \in R^{m \times p}$ and $Z \in R^{p \times n}$, where p is the dimension of sparse feature, whose product can best approximate $X : X \approx UZ$. Each column of U can be regarded as a basis vector and each column of Z is the p -dimensional representation of the original inputs with respect to the new basis.

To measure the approximation they use Frobenius norm of a matrix $\| \cdot \|$. Thus, the matrix factorization can be defined as the optimization problem as follows :

$$\min_{U, Z} \| X - UZ \|^2. \quad (2)$$

By fixing U , the optimization problem becomes a constraint linear regression problem. So they use Nadaraya-Watson kernel regression to solve the optimization problem to compute the representation matrix Z .

The approximation \hat{x}_i for any data point x_i defined as

$$\hat{x}_i = \sum_{j=1}^p z_{ji} u_j \quad (3)$$

where u_j is j -th column vector of U (landmark matrix) and z_{ji} is ji -th element of Z . So z_{ji} computed as

$$z_{ji} = \frac{K_h(x_i, u_j)}{\sum_{j' \in U(i)} K_h(x_i, u_{j'})} \quad j \in U(i) \quad (4)$$

where $U(i) \in R^{m \times r}$ denote a sub-matrix of U consisted of r nearest landmarks of x_i and $K_h(\cdot)$ is a kernel function with a bandwidths h . The Gaussian kernel $K_h(x_i, u_j) = \exp(-\|x_i - u_j\|^2 / 2h^2)$ is used here.

In landmark based spectral clustering they compute the graph matrix as $W = \hat{Z}^T \hat{Z}$ which can have a very efficient eigen-decomposition. Where $\hat{Z} = D^{-\frac{1}{2}} Z$, D is the rowsum of Z and

Z^T is the transpose of matrix Z . Also to achieve efficient eigen-decomposition they assume the SVD (Singular Value Decomposition) of \hat{Z} is $\hat{Z} = A \sum B^T$ where $\sum = \text{diag}(\sigma_1, \dots, \sigma_p)$ and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$ are the singular values of \hat{Z} . Thus $W = \hat{Z}^T \hat{Z} = B \sum A^T A \sum B^T = B \sum \sum B^T$, and $\hat{Z} \hat{Z}^T = A \sum \sum A^T$ since A and B are the left/right singular vectors and are both unitary matrix. Notice that B is obviously the eigenvectors of W and can be computed by $B = \hat{Z}^T A \sum^{-1}$. Then apply standard clustering method such as k -means to the matrix B to get the clusters [28][29].

3. THE PROPOSED METHOD

In this section, two proposed methods (LSk_medoids, LLSk_medoids) for link prediction that use k -medoids clustering technique are presented. The main goal of using k -medoids comes from its robustness as it is not affected by the presence of outliers or noise or extremes unlike clustering techniques based on k -means and also makes the prediction results better against k -means.

3.1 Link prediction using Spectral clustering by k -medoids (LSk_medoids)

In LSk_medoids, similarities between nodes in an undirected graph are found by taking into consideration the clustering information. Therefore, two friends can be recommended to each other, according to their values (weights) in the similarity matrix. The proposed method that shown in Figure 2 is illustrated as following. Firstly, compute the laplacian matrix L which is a combination of the adjacency and the degree matrix. The normalized Laplacian matrix of graph G is computed by Equation (1), then computes the first k eigenvectors c_1, \dots, c_k with the corresponding $\lambda_1, \dots, \lambda_k$ eigenvalues of L based on equation $L \times C = \lambda \times C$. Where $C = [c_1, \dots, c_k] \in R^{n \times k}$ is formed by stacking the eigenvectors in columns.

Secondly, form the matrix S with n rows (no of vertices) and k columns (eigenvectors). then treating each row of S as a point and cluster it into k clusters.

Clustering data points in matrix S using k -medoids is as follow:

- (1) Randomly select k objects as medoids from matrix S .
- (2) Associate each data object in the matrix S to most similar medoid.
- (3) Randomly select non-medoid object O and calculate the total cost of swapping for non-medoid object O , if it negative then initial medoids is replaced by new medoids.
- (4) Finally repeat steps 2 and 3 until there is no change in the medoids.

Thirdly, our method uses Equation (5) to calculate the similarity between nodes that belong in the same cluster Moreover, it uses Equation (6) to calculate the similarity between nodes that belong to different clusters.

$$\text{SimSC}(i, j) = \frac{|d(i, \text{idx}(i)) - d(j, \text{idx}(j))|}{d(i, \text{idx}(i)) + d(j, \text{idx}(j))} \quad (5)$$

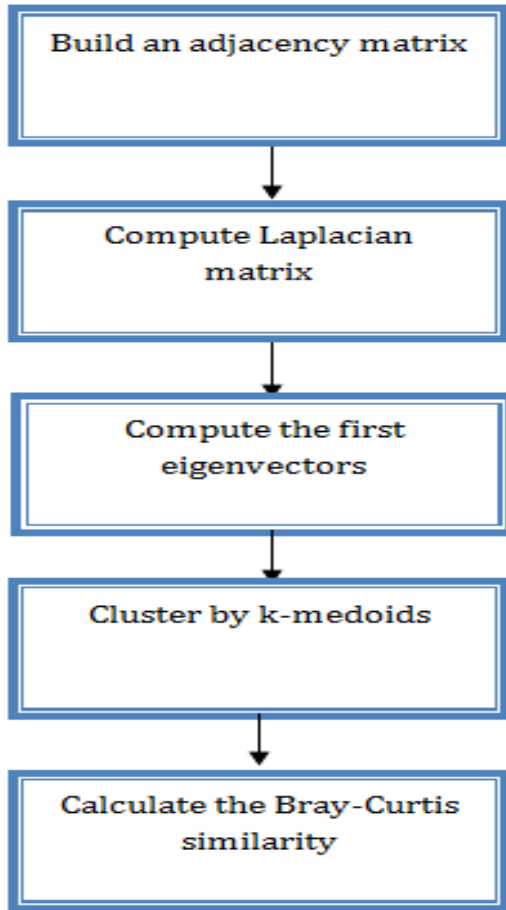


Figure 2. The first proposed method

$$SimDC(i, j) = \frac{|d(i, idx(j)) - d(j, idx(i))|}{d(i, idx(j)) + d(j, idx(i))} \quad (6)$$

Where d is the distances of each node from the centroid of each cluster, idx is a Vector which assigns each node in a specific cluster.

Calculating similarity between nodes based on triangle inequality which states that any side of a triangle is always shorter than the sum of the other two sides. The distance (dissimilarity) between any pair of nodes i and j is shorter than the sum of the distances via a third point, that is from i to $idx(i)$ and from j to $idx(j)$ as a follow $|d(i, idx(j)) - d(j, idx(i))| \leq dist(i, j) \leq d(i, idx(i)) + d(j, idx(j))$. To determine the similarity between nodes the Bray-Curtis similarity measure is adapted as in Equation (5) and Equation (6).

For a test node the calculated similarities with other nodes are ranked and predict the top ranked ones as his possible friends.

The $LSk_medoids$ method summarized as:

- Form the adjacency matrix A of a graph G .
- Define D to be the diagonal degree matrix where

$$D_{ii} = \sum_{j=1}^n A_{ij} \quad (7)$$

and construct the normalized Laplacian matrix according to Equation (1).

- Find c_1, \dots, c_k the first eigenvector of L and form the matrix S that contain c_1, \dots, c_k eigenvectors as columns and nodes $v_i \in R$ with $i = 1 : n$ correspond to the i -row of S .
- Treating each row of S as a point and cluster them into k clusters using k -medoids.
- Compute the similarity for node v_i with each node v_j based on Equation (5) and Equation (6) according to these points belongs in the same cluster or to different cluster.

In this method spectral clustering needs to build adjacency matrix and compute the eigen-decomposition of the corresponding Laplacian matrix but these steps are computational expensive, where for a data set consisting of n data points the above two steps will have time complexities of $O(n^2)$ and $O(n^3)$. So another method that depend on landmark based spectral clustering is proposed to overcome the computational cost of the eigen-decomposition step.

3.2 Link prediction using Landmark based Spectral clustering by k -medoids (LLSk_medoids)

In this method the problem that face us in first method that is data size and the computational cost of the eigen-decomposition of the graph Laplacian is solved using landmark based on spectral clustering that selects p data points as the landmarks and the remaining points as the linear combinations of these landmarks.

LLSk_medoids that shown in Figure 3 is briefed in the following five steps.

Step 1. Create p landmark points using k -means.

Step 2. Calculate the representation matrix $Z \in R^{p \times n}$ between observations and landmark points according to Equation (4). The goal of our approach is designing an efficient way to construct our graph and Laplacian matrix eigen-decomposition. So, the affinity matrix is designed as follows: $W = \hat{Z}^T \hat{Z}$ Where $\hat{Z} \in R^{p \times n}$ and $p \ll n$. Thus, the graph can be build in $O(np)$ and compute eigenvectors of the graph Laplacian in $O(p^3 + p^2n)$ which is a significant reduction from $O(n^3)$.

Step 3. Calculate $A = [a_1, \dots, a_k]$ the first k eigenvectors of $\hat{Z} \hat{Z}^T$.

Step 4. Calculate $B = [b_1, \dots, b_k] \in R^{n \times p}$ according to $B = \hat{Z}^T A \Sigma^{-1}$.

Step 5. Each row of B is a data point and apply k -medoids clustering to the matrix B to get the clusters.

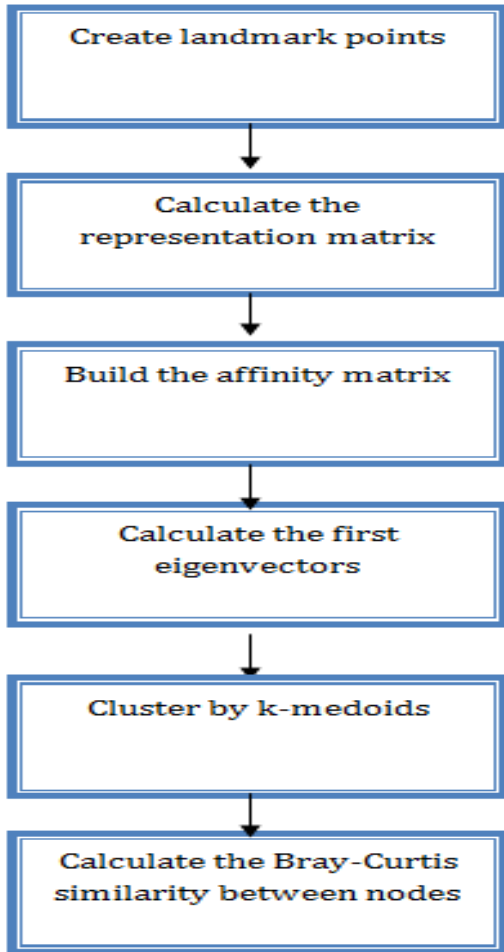


Figure 3. The second proposed method

Step 6. Based on triangle inequality calculate the similarity between nodes that belong in the same cluster using Equation (5) and the similarity between nodes that belong to different clusters using Equation (6), higher similarity means higher probability that corresponding nodes are connected.

4. EXPERIMENTAL RESULTS AND COMPARISON

The proposed methods are implemented in MATLAB R2013a, and has been tested on nine real datasets. An brief description of the data sets is listed below:

Karate is a social network of interactions between members of a karate club by Wayne Zachary [33]; UST is a network of the US (united states) air transportation system [34]; PB is a directed network of US political blogs, here its links are treated as undirected and self-connections are omitted [35]; Dolphin is a network of frequent associations between 62 dolphins in a community living off Doubtful Sound [36]; Football is a network of American football games between colleges during regular season Fall 2000 [37]; Taro is a network that represent the relation of gift-giving (taro exchange) among 22 households

in a Papuan village [38]; David Copperfield is the undirected network of common noun and adjective adjacencies for the novel "David Copperfield", here an edge connects two words that occur in adjacent positions [39]; Korea is a communication network within a small enterprise, Two employees were linked in the communication network if they rated their contact as three or more [40]; Hens is a directed network of of 32 White Leghorn hens observed in 1946. A node represents a hen and an edge represents dominance of the left hen over the right hen [41], here its links are treated as undirected and self-connections are omitted. The detailed information about these networks is described in Table 1 where N is the number of nodes and M denotes the number of links.

In our experiments the observed links E is randomly divided into two parts: the training set E^r (90% of the observed links E) is treated as known information, while the test set E^e (10% of the observed links E), is treated as unknown information for prediction. Obviously, $E = E^r \cup E^e$ and $E^r \cap E^e = \phi$.

To evaluate the algorithm's performance of our methods on the nine real networks accuracy measured by AUC metric (area under the curve) is applied. The AUC value can be explicated as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. Through n independent comparisons, if there are n' times the missing link having a higher score and n'' times they have the same score [2][32], then define AUC by Equation (8) as:

$$AUC = \frac{n' + 0.5n''}{n} \quad (8)$$

Table 1. The basic topological characteristics of nine real-world networks

Networks	N	M
Karate	34	78
UST	332	2126
PB	1224	19090
Dolphin	62	159
Football	115	616
Taro	22	32
David Copperfield	112	425
Korea	39	69
Hens	32	469

Table 2. The parameters used in our methods

Parameters	p	r	number of clusters
Karate	7	5	5
UST	110	9	45
PB	10	8	6
Dolphin	38	28	12
Football	71	19	13
Taro	11	8	4
David Copperfield	62	14	28
Korea	3	3	12
Hens	19	19	10

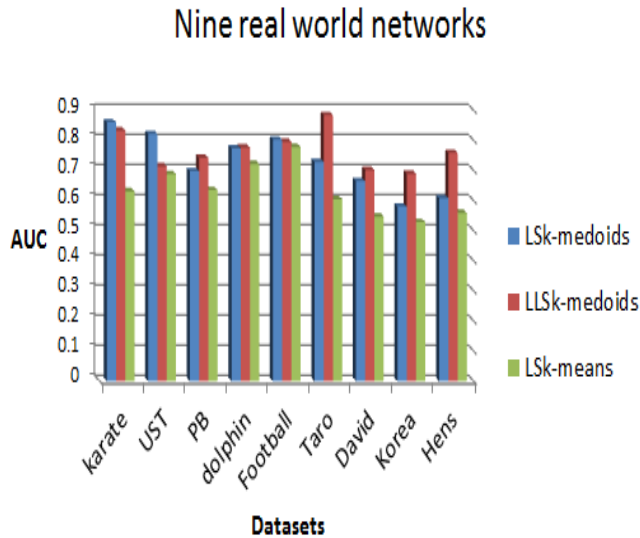


Figure 4. Prediction accuracy on real-world networks

Table 3. The prediction accuracy measured by AUC on four real-world networks

Dataset	LSk-medoids	LLSk-medoids	LSk-means
Karate	0.8606	0.8332	0.6312
UST	0.8210	0.7186	0.6886
PB	0.6975	0.7425	0.6348
Dolphins	0.7735	0.7767	0.7220
Football	0.8022	0.7954	0.7778
Taro	0.7282	0.8840	0.6047
David Copperfield	0.6651	0.7018	0.5469
Korea	0.5800	0.6892	0.5286
Hens	0.6080	0.7586	0.5595

In the proposed methods of this paper there are three parameters: The number of landmark points p , the number of nearest landmarks r which belong to the second method and the number of clusters that reported in Table 2.

The results simulated in Table 3 and Figure 4 demonstrate the prediction accuracy measured by AUC on nine real world networks, where the our proposed methods achieve high accuracy compared with LSk-means.

The first parameter in LLSk-medoids method is the number of landmark points p . Figure 5 shows how the prediction accuracy changes as p varying from 10 to 110 on PB dataset. It can be seen that the accuracy results decrease when the number of landmarks increase and LLSk-medoids achieves better prediction accuracy at $p = 10$. Another parameter in LLSk-medoids method is the number of nearest landmarks r . Figure 6 shows how the prediction accuracy varies with this parameter. It achieves good performance at $r = 8$ on PB dataset.

5. CONCLUSION

The aim of this work is to achieve high accuracy in link prediction in social network. Specifically a link prediction using spectral clustering by k -medoids and link prediction using landmark based

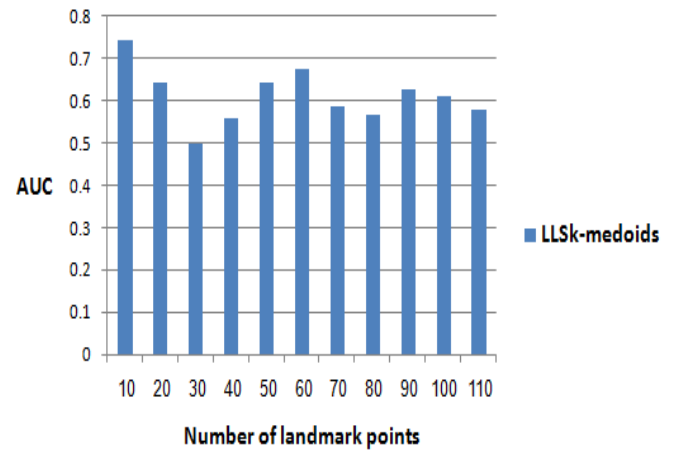


Figure 5. Clustering accuracy and number of landmark points on PB data set

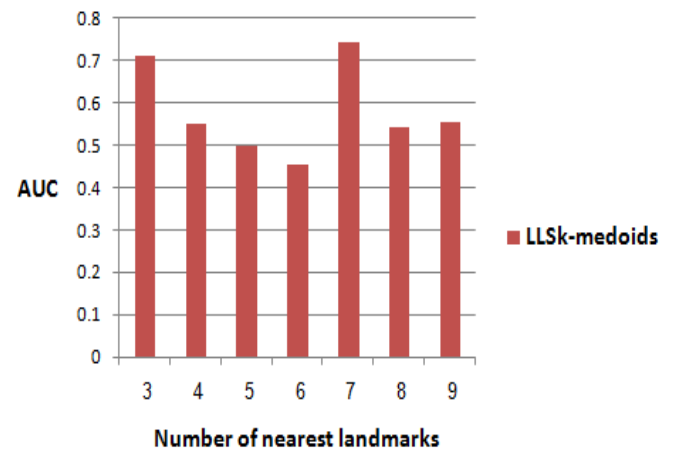


Figure 6. Clustering accuracy and number of nearest landmarks on PB data set

spectral clustering by k -medoids methods are presented. The experiments are applied to 9 different datasets. The results obtained are compared to other benchmark method (LSk-means) and shows an enhancement of accuracy of link prediction.

6. REFERENCES

- [1] J. Valverde-Rebaza and A. de Andrade Lopes, "Link Prediction in Complex Networks Based on Cluster Information", SBIA'12 Proceedings of the 21st Brazilian conference on Advances in Artificial Intelligence, Vol 7589, pp 92-101, 2012.
- [2] P.Symeonidis, N. Iakovidou, N. Mantas and Y. Manolopoulos: "From biological to social networks Link prediction based on multi-way spectral clustering", data Knowl. Eng, Vol 87, pp 226-242, May 2013.
- [3] Li F, He J, Huang G, Zhang Y, Shi Y (2014), "A clustering-based link prediction method in social networks", In: 14th

- international conference on computational science. ICCS 2014, Vol 29, pp 432-442 ,2014.
- [4] M. Al Hasan, V. Chaoji, S. Salem, and M. J. Zaki, "Link Prediction using Supervised Learning", In Proceedings of the 4th Workshop of Link analysis, Counterterrorism and Security (with SIAM). Bethesda, MD, USA, pp 1-10, 2006 .
- [5] M. Fire, Lena Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, SocialCom, "Link Prediction in Social Networks using Computationally Efficient Topological Features", IEEE Third International Conference on Social Computing (Social-Com), Vol 2, pp 73-80, 2011.
- [6] B.C.Craenen ,A.E. Eiben, "Computational Intelligence", ch6 in ARTIFICIAL INTELLIGENCE in Encyclopedia of Life Support Systems (EOLSS), Vol 1, pp 105-118, 2009.
- [7] M.C. Nicoletti¹, L.C. Jain², and R.C. Giordano, "Computational Intelligence Techniques for Bioprocess Modelling", Supervision and Control Volume 218 of the series Studies in Computational Intelligence, pp 1-23, 2009.
- [8] Parodi, "Annals of Actuarial Science", Vol 6, part 2, pp 344-380, 2012.
- [9] Lamiaa M. El Bakrawy, Abeer S. Desuky, "A hybrid classification algorithm and its application on four real-world data sets", International Journal of Computer Science and Information Security (IJCSIS), Vol 13, No. 10,pp 93-97, 2015
- [10] Khurram I. Qazi a, H.K. Lam a, Bo Xiao a, Gaoxiang Ouyang b and Xunhe Yin c, " Classification of epilepsy using computational intelligence techniques", CAAI Transactions on Intelligence Technology, Vol 1, Issue 2, pp 137149, 2016.
- [11] Han Yuan, Yunlong Ma, Feng Zhangaand Min Liu and Weiming Shen, "A Distributed Link Prediction Algorithm Based on Clustering in Dynamic Social Networks", IEEE International Conference on Systems, Man, and Cybernetics 2015, pp 1341-1345, 2015.
- [12] X. Chen and D. Cai, "Large Scale Spectral Clustering via Landmark-Based sparse Representation", IEEE Transactions on Cybernetics, Vol 45, Issue 8, pp 1669-1680, 2015 .
- [13] Soundarajan S.,Hopcroft J, "Using community information to improve the precision of link prediction methods", in Proceedings of the 21st International Conference Companion on World Wide Web (companion volume), pp 607-608, 2012.
- [14] Z. Abbassi and V. S. Mirrokni, "A Recommender System Based on Local Random Walks and Spectral Methods", Advances in Web Mining and Web Usage Analysis, Vol 5439 of the series Lecture Notes in Computer Science, pp 139-153, 2007.
- [15] E. Hoseini, S. Hashemi and A. Hamzeh, "Link Prediction in Social Network Using Co-clustering based Approach", 2012 26th International Conference on Advanced Information Networking and Applications Workshops, Vol 189 ,pp 795-800, 2012.
- [16] Kunegis, J. Fay, D.; Bauckhage, C., "Network Growth and the Spectral Evolution Model", CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, pp 739-748, 2010
- [17] Z. Huang, "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient", In Workshop on Link Analysis, the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 20-23, 2006.
- [18] J. Kim, M. Choy, D. Kim, U. Kang, "Link prediction based on generalized cluster information", WWW'14 companion Proceedings of the 23rd International Conference on World Wide Web (companion volume), pp 317-318, 2014.
- [19] X.Wang, D. He, D. Chen and J. Xu, "Clustering-Based Collaborative Filtering for Link Prediction" , AAI Publications, Twenty-Ninth AAI Conference on Artificial Intelligence, pp 332-338, 2015
- [20] Dr.T. Velmurugan, "Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points", International Journal of Computer Applications in Technology, Vol 3(5), pp 1758-1764, 2012.
- [21] T.S.Madhulatha, "Comparison between K-Means and K-Medoids Clustering Algorithms", Advances in Computing and Information Technology, Vol 198, pp 472-481, 2011.
- [22] S.Chu, J.F Roddick, Yi . Chen, J.S. Pan, "Efficient search approaches for k-medoids-based algorithms" TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002.
- [23] H.S Park, C.H Jun, "A simple and fast algorithm for K-medoids clustering", Expert Systems with Applications, Vol 36 , pp 3336-3341, 2009.
- [24] R. Pratap, K .S.Vani, J .R. Devi and Dr.K Nageswara Rao, "An Efficient Density based Improved K- Medoids Clustering algorithm", International Journal of Advanced Computer Science and Applications (IJACSA), Vol 2 Issue 6, pp 49-54, 2011.
- [25] S. Saket, S. Pandya, "Implementation of Extended K-Medoids Algorithm to Increase Efficiency and Scalability using Large Datasets", International Journal of Computer Applications (0975 - 8887), Vol 6, No 5, pp 19-23, 2016.
- [26] Ulrike von Luxburg," A Tutorial on Spectral Clustering",Statistics and Computing December 2007, Vol 17, Issue 4, pp 395-416, 2007.
- [27] D. Yan, L. Huang, M. Jordan, "Fast approximate spectral clustering", Proceedings 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp 907-916, (Paris, France), 2009.

- [28] Zejun Gan, Chaofeng Sha and Junyu Niu, "Fast Spectral Clustering with Landmark-based Subspace Iteration", 2013 IEEE 13th International Conference on Data Mining Workshops, pp 773-779, 2013.
- [29] Feng X., Zhao J., Xu K, "Link prediction in complex networks: a clustering perspective", *The European Physical Journal B*, Vol 85, pp 1-9, 2012.
- [30] D. Cai, H Bao, X. He, "Sparse Concept Coding for Visual Analysis", *Proceeding CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.
- [31] H. Lee, R. Raina, A. Battle and A. Ng, "Efficient sparse coding algorithms", *NIPS*, Vol 22, pp 801-808, 2007.
- [32] L. L. Zhou T, "Link prediction in complex networks: a survey", *Physica A: Statistical Mechanics and its Applications*, Vol 390, pp 1150-1170, 2011.
- [33] Zachary W, "An information flow model for conflict and fission in small groups", *Journal of Anthropological Research*, Vol 33, pp 452-473, 1977,
<http://networkdata.ics.uci.edu/data/karate/karate.paj>
- [34] Batageli V, Mrvar A, Pajek datasets (2006),
<http://vlado.fmf.uni-lj.si/pub/networks/data/mix/usair97.net>.
- [35] Ackland R, "Mapping the US political blogosphere", *Presentation to Blog Talk Downunder*, 2005. <http://www-personal.umich.edu/mejn/netdata/>.
- [36] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations" *Behavioral Ecology and Sociobiology*, VOL 54, pp 396-405 (2003).
<http://networkdata.ics.uci.edu/data/dolphins/dolphins.paj>.
- [37] Girvan and M. E. J. Newman, "Community structure in social and biological networks", *Proc. Natl. Acad. Sci. USA*, VOL 99, pp 7821-7826, 2002.
<http://networkdata.ics.uci.edu/data/football/football.paj>.
- [38] Hage P. and Harary F, "Structural models in anthropology", Cambridge: Cambridge University Press, 1983.
<http://vlado.fmf.uni-lj.si/pub/networks/Data/ucinet/taro.dat>.
- [39] Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, Vol 74(3), 2006. http://konect.uni-koblenz.de/networks/adjnoun_adjacency
- [40] W. de Nooy, A. Mrvar, and V. Batagelj, "Exploratory Social Network Analysis with Pajek", Cambridge: Cambridge University Press, 2004. <http://vlado.fmf.uni-lj.si/pub/networks/data./esna/korea.htm>.
- [41] AM Guhl. "Social behavior of the domestic fowl", Manhattan, Kansas: Kansas State College, Agricultural Experiment