Attractiveness Based Correlated Probabilistic Graphs Clustering with user Profile Information

M. Kalpana
M.E Final Year Student
Department of Computer Science and Engineering,
J.K.K.Nattraja College of Engineering and Technology,
Komarapalayam - 638 183, Tamilnadu, India

P. Ramya
Assistant Professor,
Department of Computer Science and Engineering,
J.K.K.Nattraja College of Engineering and Technology,
Komarapalayam - 638 183, Tamilnadu, India

***Abstract*- Recently, probabilistic graphs have attracted significant interests of the data mining community. It is observed that correlations may exist among adjacent edges in various probabilistic graphs. Graph clustering aims to divide data into clusters according to their similarities, and a number of algorithms have been proposed for clustering graphs. In this paper, the clustering process based on the weight value node and edge in the network is considered. Additionally, the proposed system combines the user profile of users with the efficient graph clustering technique. In other words, the user profile information for clustering process is analysed. This consideration of user profile information is improving the clustering accuracy and enhances the performance. The effectiveness and efficiency of the proposed algorithm is to be evaluated.The proposed system is well reducing the time complexity and the error rate of the graph clustering system as compared with leading clustering algorithms for correlated probabilistic garphs.**

Keywords – Clustering; correlated; probabilistic graphs algorithm; weighted; user profile information

I. INTRODUCTION

In recent years, graph mining has gained significant attention for a broad range of applications, such as social networks, protein-protein interaction networks, road networks, etc. The data from such applications typically displays an inherent property of uncertainty, and they can be rationally modeled as probabilistic graphs, in which each edge e_i is labeled with an existence probability to represent the uncertainty of the data. In a probabilistic graph, any two edges e_i and $e_j$are called conditionally independent if and only if $p(e_i,e_j)=p(e_i)p(e_j)$,and conditionally dependentif and only if $p(e_i,e_j)\neq p(e_i)p(e_j)$. For the standard probabilistic graph model, any two edges are conditionally independent of each other.

Particularly, according to statistical models in many real scenarios, the correlations among edges do not simply follow mutex or coexistence patterns, and more complicated dependency may exist.As one of the basic data mining techniques, clustering is widely used in various graph analysis applications, such as community detection and index construction, etc.

This paper focuses on clustering correlated probabilistic graphs which aims to partition the vertices into several disconnected clusters with high intra-cluster and low inter-cluster similarity.To cluster a correlated probabilistic graph G, a possible world graph $G_i$of$G$ can be modeled as a deterministic instantiation sampled from the correlated probabilistic graph according to the joint probability distribution. The edit distance $D(G_i,Q)$ from $G_i$to the cluster graph Q is defined as the number of edges that need to be added or removed to transform $G_i$into Q. The edit distance $D(G,Q)$can be obtained and viewed as a measurement to evaluate the deviation from a correlated probabilistic graph to the cluster graph. Hence, a smaller deviation implies a more precise result, and the objective turns to the goal of finding a cluster graph Qthat can minimize $D(G,Q)$.However, it is extremely time-consuming if the expected edit distance is calculated by considering all possible world graphs.

Based on the estimation model, the proposed clustering algorithm named Attractiveness-based Community Clustering (ACC) algorithmthat initializes a cluster graph, and then iteratively improve it by adding or removing vertices from some clusters if the estimated edit distance can be reduced. Note that in this process, instead of calculating $D(G,Q)$,there is need to judge whether $D(G,Q)$ is reduced by considering the change of a cluster graph.

In Protein-Protein Interaction (PPI) networks, the interaction between two proteins is generally established with a probability property due to the limitation of observation methods. Clustering applied to such correlated probabilistic protein-protein interaction network data is helpful in finding complexes to analyze the structure properties of the PPI Network.

Consider another example in social networks. The edge probability is used to quantify the reliability of a link. Obviously, there are correlations for the links in social networks. Accordingly, clustering applied to social networks while considering the potential probabilities and correlations could be more effective to detect user communities. Compared with clustering probabilistic graphs without any correlation among edges, clustering correlated probabilistic graphs has more constraints.

II.LITERATURESURVEY

A. Algorithms for Clustering Deterministic Graphs

Graph clustering has been extensively studied in data mining research and a number of clusteringalgorithms have been developed. Ahmed [1] provided a survey of graph clustering methods. They discussed the different categories of clustering algorithms and recent efforts to design clustering methods for various kinds of graph data. They presented a taxonomy of clustering techniques and applications of clustering algorithms. That paper shows the efficiency factors to be measured for clustering deterministic graphs.

As one of the most widely used graph clustering algorithms, spectral clustering has received increased interest of researchers. Spectral clustering relies on the eigen- structure of a graph Laplacian matrix to partition vertices into disjoint clusters, with points in the same cluster having high similarity and points in different clusters having low similarity. The rationality of the spectral clustering method was analyzed by Bach[2]. They derived new cost functions for spectral clustering based on measures of error between a given partition and a solution of the spectral relaxation. Furthermore, a number of optimizations for spectral clustering were proposed.

However, most of the existing algorithms are applied in clustering deterministic graphs. Particularly, as correlations exist among edges, it is inappropriate to directly apply these algorithms to clustering correlated probabilistic graphs.

B.*Querying and Mining of Probabilistic Graphs*

Recently, querying and mining of probabilistic graphs have attracted growing attention by researchers. Many classical data mining problems have been redefined in probabilistic graphs, such as the reachability query, shortest path query, *K*-NN query, etc. Jin [5] studied the Distanceconstraint Reachability query and presented a sampling algorithm to answer the NP-hard problem. Michalis[7] introduced an efficient algorithm for *K*-NN queries in probabilistic graphs based on the random walk method.

As an important preliminary work, George [6] advanced the state of the art by exploring the problem of clustering probabilistic graphs. They proposed efficient algorithms to find a cluster graph. Nevertheless, these algorithms do not consider the correlations among edges, and thus are not applicable for clustering correlated probabilistic graphs.

C. *Querying and Mining the Probabilistic Data with Correlations*

Recently, correlations among uncertain data have received increased interest. Sen [8] proposed a framework to represent the correlations among probabilistic tuples. An efficient strategy was developed for query evaluation over such probabilistic databases by casting the query processing problem as inference problem in an appropriately constructed probabilistic graphical model. They investigated the nearest neighbor query on uncertain data with local correlations..There also exist studies on evaluating correlated probabilistic graphs. Hua[4] defined the problem of probabilistic path queries in correlated probabilistic networks. They devised three effective heuristic evaluation functions to in advance estimate the conditional probability of each edge. Yuan[9] proposed a method for subgraph similarity search over correlated probabilistic graphs based on possible world semantics.

III.PROBLEM DEFINITION

The model of a correlated probabilistic graph as $G=\{V,E,P,F\}$is defined, where V is the set of vertices, E is the set of edges, P is the existence probability, and F is the joint probability distribution of edges. It is assumed that the joint probabilities only exist among edges that share the same vertex. The output graph is modeled as a cluster graph which is composed of several disconnected clusters and each vertex in the graph only belongs to one cluster.

In this model, the joint probability distribution F is of the form $F(e_1,\bar{e}_2,\ldots,e_k)=p$, where $e_i$denotes existence,$e_i$denotes nonexistence for edge e, and p is the value of the joint probability.A possible world graph serves as an efficient model in dealing with probabilistic graphs. For a correlated probabilistic graph $G=\{V,E,P,F\}$, a possible world graph $G_i=\{V',E'\}$is an instantiation sampled from G, where $V'=V$ and E'thatbelongs to$E$. Additionally, $X_i(e_j)$is referred as the existence state of edge $e_j$inG_i,i.e., if the edge $e_j$exists inG_i,$X(e_j)=e_j$;otherwise,$X_i(e_j)=\bar{e}_j$.Similarly, $X_Q(e_j)$ indicates the existence state of an edge e_j in the cluster graph Q.When calculating the sampling probability of a possible world graph G_i, an edge order *(EO)* is necessary for conditional probability calculation.

Next, the definition of the edit distance is extended from a probabilistic graph to a cluster graph proposed in [8] to accommodate the correlations. Given a correlated probabilistic graph $G=\{V,E,P,F\}$ and an output cluster graph Q, according to the possible world semantics, the expected edit distance is defined from G to Q as $D(G,Q) = E_{Gi\subseteq G}[D(G_i,Q)]$.

Given a correlated probabilistic graph G, this paper proposes two types of algorithms to find a cluster graph Q. The algorithm named ACC assumes that the number of the output clusters is not fixed, while the other one named CPGS focuses on a fixed number of clusters. The effectiveness of an output cluster graph Q is evaluated by $D^k(G, Q)$.It is NP-hard to cluster a deterministic graph via edit distance, known as the Cluster-Editproblem. Obviously, clustering correlated probabilistic graphs is an NP-hard problem as it is a generalization of the Cluster-Editproblem. Therefore, the approximate algorithmis designed in this paper.

IV. EXISTING SYSTEM

The major contributions of existing system [3] are the development of two clustering algorithms. They are PEEDR (Partially Expected Edit Distance Reduction) and CPGS (Correlated Probabilistic Graphs Spectral). PEEDRis rather efficient for clustering correlated probabilistic graphs with several pruning methods for this algorithm. CPGS is for clustering correlated probabilistic graphs based on the spectral clustering algorithm, which can produce better cluster results, although it is less efficient than PEEDR.

A.PEEDRClustering Algorithm

The PEEDR algorithm initializes a cluster with one vertex. Then for each vertex that is adjacent to the cluster, it is removed into the cluster if it reduces the expected edit distance from G to the current cluster graph. The above step is iteratively applied until the cluster cannot be expanded. Next a vertex has to be choosenfrom the unclustered vertices and repeat the above procedure to generate another cluster. The procedure is repeated until all vertices of G are grouped into clusters. Consequently, the final cluster graph is obatined.

The algorithm first sorts all the vertices in descending order of their degrees. It next initializes a virtual cluster C', where $V_{C'}= \{v|v\in V\}$,which keeps all the unclustered vertices. The algorithm finds the vertex with the maximum degree in $V_{C'}$, denoted asV'. As an optimization of proposedalgorithm, builds a Distance-Probability-Threshold Clique DPTC centered at V', denoted as C. The algorithm will check each vertex $V_j \in V_{C'}$ that is adjacent to Q, and put $V_j$into C if it can reduce the objective function $D(G,Q)$. The checking is a key step of the algorithm, for which it invokes a function *isReduceEdit*and several optimization techniques. Then update the set $V_{C'} = V_{C'} \backslash V_{Ci}$ and create the next cluster from the vertices in the virtual cluster C'.The procedure is repeated until all the vertices in C' are grouped into disconnected clusters.

B. CPGSClustering Algorithm

The clustering process of the PEEDR algorithm starts from a local graph and establishes the cluster graph gradually. As vertices will never be separated once grouped into a cluster, it is essentially a greedy algorithm. The PEEDR algorithm may not meet the need for high precision. Besides,
there exists no prior information about the number of final clusters. In some applications, graph clustering aims to partition vertices into a certain number of clusters.Spectral clustering refers to a class of techniques which rely on the eigen-structure of a graph Laplacian matrix to partition vertices into disjoint clusters with high intra-cluster and low inter-cluster similarity In the CPGS algorithm, cluster a graph by establishing DPTCs and representing these DPTCs as objects to be clustered. This can improve the time efficiency as the number of objects to be clustered is much smaller than the number of vertices.
However, directly applying the spectral clustering algorithm in correlated probabilistic graphs will incur high time complexity. In this existing work, CPGS (Correlated Probabilistic Graphs Spectral) is to cluster correlated probabilistic graphs. Given a correlated probabilistic graph G and a cluster number K, it reduces the number of objects by establishing DPTCs first, and represent these DPTCs as the objects to be clustered. Second, the similarity is defined between pairwise adjacent DPTCs to find the K-NN of each DPTC. Third, a Laplacian matrix can be obtained according to the K-NN results, and the eigenvectors of the Laplacian matrix is to be calculated. Last, each DPTC Q will be represented by a point $P_{Ki}((U_{1i}, U_{2i}, \ldots, U_{Ki})$ in a K-dimensional space, and these points are iteratively clustered with a K-means algorithm, such that the final cluster graph is obtained.

C. Disadvantages

Most of the networks in the real world are weighted networks. Thus this system doesn't give well accuracy rate for such a network clustering. Clustering accuracy of the system is reduced in this system. Time complexity of the system is increased in this system. Since this system doesn't anlayse the weight value for the node and edge in the network.

V.PROPOSED SYSTEM

In the existing system, the two algorithms, namely Partially Expected Edit Distance Reductionclustering algorithm (PEEDR) andCorrelated Probabilistic Graph Spectral clustering algorithm (CPGS) has number of issues such as increased time complexity and running time. In addition, the clustering accuracy of the system is less. In order to overcome these problems in this system, the novel weighted clustering algorithms which is called as attractiveness-based community clustering algorithm (ACC) is proposed. The user profile information is to be considered while the clusters are created.

In the proposed system, the weighted graph clustering is proposed. In this system, the clustering process is performed based on the weight value node and edge in the social network. Consider suppose each person or a community has a density value, and each pair of persons or communities has an attractiveness value. The social network is a graph, each person is a node, and edges are the relationship between people. Given a graph $G(V, E,$ P, F, $W_V, S_E)$ which consists of V is the set of vertices, E is the set of edges, P is the existence probability, and F is the joint probability distribution of edges, the weight of node set W_V, and the weight of edge set S_E, The clusters of G as communities are interested to find. The weight of node implies the core degree of the person in the network, and the weight of edge means the attractiveness between the two nodes. The result of graph clustering should partition a graph into several sub-graph(clusters), each part has a weight value, what's more, there are attractiveness values between clusters which similar with the edge weights. The candidate communities should have weights higher than the attractiveness with other clusters.

To define the weighted graph clustering process based on the weight value node and edge in the social network. To develop attractiveness-based community clustering algorithm and analysing user profile information for clustering process. This consideration of user profile information is improving the clustering accuracy and enhances the performance. In the case of large social network, the clustering process taken the long time for clustering in the existing system. To solve this problem, the inclusion of user profiles is used for well reduce the time complexity and improve the performance of the system.

A. ACC Algorithm

It is an amalgamation algorithm, the merge between clusters could be considered while the attractiveness of clusters (as the edge weight) is bigger than the densities of clusters (as the node weight). ACC algorithm is designed to make some breakthrough on the time complexity of community detection for large social networks. The algorithm does not require specifying the number of clusters, because the number is usually not known in advance and is difficult to estimate in actual applications.

The weight of node implies the core degree of the person in the network, and the weight of edge means the attractiveness between the two nodes. The result of graph clustering should partition a graph into several sub-graph(clusters), each part has a weight value, what's more, there are attractiveness values between clusters which similar with the edge weights. The candidate communities should have weights higher than the attractiveness with other clusters.It determines the threshold for probability and distance. ThenDistance ProbabilityThreshold Clique Constructionis applied.

ACC algorithm can be divided into two main steps, iterating between the two steps to get clusters. The first part consists of merging the pair of clusters which has the largest attractiveness. The second part is as follows. The cluster density and cluster attractiveness matrix is calculated and is updated. Executing the update of cluster density and attractiveness matrix, and the cluster merger process iteratively, until the structure of clusters does not change, or there is only one cluster left.

B. Analysing User Profile Information

The list of users, contains information about education, interests of a user, etc. According to the information of this user profile, perform the graph clustering effectively. From this consideration of user profile, the number of common neighbours for the users is obtained. This is important factor for clustering process in order to reduce the long time in large social network data. Here, the similarity between the user profiles is performed. Based on this similarity value of the data it is clustered in this proposed system.

C. Advantages

Time complexity of the system is well reduced also running time of this system is decreased. Time complexity of community detection for large social networks is reduced.Clustering accuracy rate of this weighted graph clustering technique is enhanced compared to the existing system.Most of the networks in the real world are weighted networks. Therefore, the proposed system is highly applicable in real world applications.

VI. EXPERIMENTS

A. Experimental Setup

In this section, the performance of the proposed algorithms is studied. The algorithms are implemented in Microsoft Visual Studio C++ on a PC with a 4 dual core CPU and 8GB memory.

In the YouTube social network, each vertex represents a user and an edge between two vertices represents there exists a connection between them. This dataset comprises 1,134,890 vertices and 5,987,624 edges. The edge existence probability is randomly generated to indicate the link reliability between users.This dataset do not contain the correlation probabilities among adjacent edges. To generate these probabilities, the several definitions are present. For the *m* adjacent edges, the correlation rate is defined among these edges sharing the same vertex, so that the correlation only exists among edges, and the other edges are independent of each other.The efficiency and effectiveness of different parameters on the proposed algorithms are studied.

B. Comparison with Existing Methods

The proposed method is compared with existing graph clustering methods in this subsection. Fig. 6.1 reports the accuracy rate of Attractiveness-based Community Clustering (ACC),PEEDR and CPGS (leading existing clustering algorithms).Accuracy rate or classification accuracy is calculated as

$$\text{Accuracy Rate} = \frac{\text{TP} + \text{T N}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (1)$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.True negative (TN) has occurred when both the prediction outcome and the actual value are n in the number of input data. False negative (FN) is when the prediction outcome is n while the actual value is p. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP). If the outcome from a prediction is p and the actual value is n then it is said to be a false positive (FP).The accuracy rate decreases as the vertex number increases. ACC algorithm generates high accuracy rate than PEEDR and CPGS algorithms.

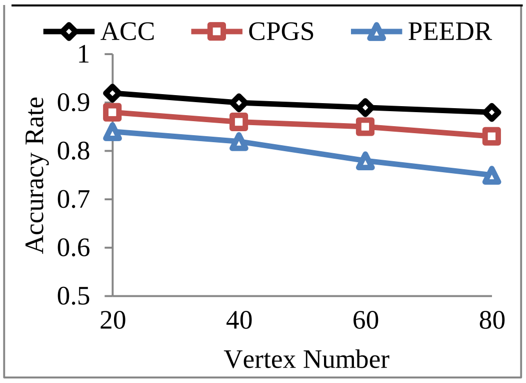

Fig. 6.1 Performance Evaluation

VII. CONCLUSION

In social network, to address the problem of clustering correlated probabilistic graphs using PEEDR and CPGS clustering algorithm based on the properties of Joint Probability is difficult and time complexity is high. To reduce the time complexity and obtain the better accuracy rate of clustering, Attractiveness-based Community Clustering (ACC) algorithm is developed. Cluster density is determined by Attractiveness Matrix and the clustering accuracy is evaluated. To perform the clustering process based on the weight value node and edge in the network, User Profiling Information is analysed based on proposed algorithm ACC. The time complexity of clustering process is reduced and clustering accuracy is in increasing range.

REFERENCES

[1] C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*, New York, USA: Springer, 2010.

[2] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.,* vol. 7, pp. 1963-2001, Oct. 2006.

[3] Y. Gu, C. Gao, G. Cong and G. Yu, "Effective and Efficient Clustering Methods for Correlated Probabilistic Graphs", *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1117-1130, May 2014.

[4] M. Hua and J. Pei, "Probabilistic path queries in road networks: Traffic uncertainty aware path selection," in *Proc. 13th Int. EDBT*, New York, 2010, pp. 347-358.

[5] R. Jin, L. Liu, B. Ding, and H. Wang, "Distance-constraint reachability computation in uncertain graphs," *PVLDB*, vol. 4, no. 9, pp. 551-562, Jun. 2011.

[6] G. Kollios, M. Potamias, and E. Terzi, "Clustering large probabilistic graphs," *IEEE Trans. Knowl. Data Eng.*,vol. 25, no. 2, pp. 325-336, Feb. 2013.

[7] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios, "K-nearest neighbors in uncertain graphs," *PVLDB*,vol. 3, no. 1, pp. 997-1008, Sept. 2010.

[8] P. Sen and A. Deshpande, "Representing and querying correlated tuples in probabilistic databases," in *Proc. ICDE*, Istanbul, Turkey, 2007, pp. 596-605.

[9] Y. Yuan, G. Wang, L. Chen, and H. Wang, "Efficient subgraph similarity search on large probabilistic graph databases,"*PVLDB*, vol. 5, no. 9, pp. 800-811, May 2012.