IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

# Spectral Clustering based Approach on Large Water Distribution Network

**S.V.S. Santhi[1], Poosapati Padmaja[2]**

Research Scholar, Department of Information Technology, GITAM University, Visakhapatnam, India.[1]

Associate Professor, Department of Information Technology, GITAM University, Visakhapatnam, India.[2]

**Abstract:** Minimum spanning tree(MST) based clustering approach is a powerful tool for performing cluster analysis on large graphs. The algorithm produces k-clusters with the MST. To identify the crucial components, MST based clustering approach plays a major role. Real world applications like Water Distribution Networks are complex networks that require innovative technological solutions for efficient management of the system. Spectral clustering algorithm is more effective than traditional clustering algorithm. In this paper we propose a clustering approach on Water Distribution Network. We suggest to find the MST for the large graph and apply spectral clustering algorithm to divide the MST into K-clusters. Spectral clustering algorithm is capable of detecting clusters with irregular boundaries. Experimental results show that the proposed approach performed well and found to fit to the expected results for both synthetic data and real world data.

**Keywords**: Large graph, Minimum Spanning Tree, Clustering, Spectral Clustering, Water distribution network.

## I. INTRODUCTION

Graphs are structures formed by set of vertices and set of edges that are connections between pairs of vertices. A graph is a tuple G=(V,E,∑, L) where V is a set of vertices, $E \subseteq V \times V$ is a set of edges, ∑ is a set of labels and L: $V \cup E \rightarrow \sum$ is a function assigning labels to vertices and edges. The size of the graph is |E(G)| i.e., the number of edges it contains. A graph becomes 'large' when the amount of data becomes 'big'. Applications of large infrastructure networks include urban roads, rail networks, power grid, gas pipe line networks, water distribution networks etc., In this paper Water Distribution System is taken as a real world application for large graph mining.

With the continuous development of social economy, the demand for the water supply is increased drastically. To meet the requirements and also solve the economic problems in laying pipelines , minimum spanning tree is needed to be implemented.

A Minimum Spanning Tree is an acyclic sub graph of a graph G, which contains all vertices from G. Let G=(V,E) be an undirected connected graph. A sub graph T=(V',E') of G is a spanning tree of G iff T is a tree where there are no cycles. It covers all the vertices V and contains |V|-1 edges. A single graph can have many different spanning trees. A minimum cost spanning tree is a spanning tree which has a minimum total cost. Addition of even one single edge results in the spanning tree losing its property of acyclicity and removal of one single edge results in its loosing the property of connectivity. The length of a tree is equal to the sum of the length of the arcs on the tree. If each edge has a distinct weight then there will be only one unique minimum spanning tree. By applying MST the cost of the network can be reduced. Few of the existing algorithms for finding MST are Kruskal's and Prime's algorithms.

A water distribution system is a collection of hydraulic control elements connected together to convey water from sources to consumers. Water supply system is in the order of intake structures, water treatment structure, water supply pumping stations, and water supply pipe network. In laying city network, street interchanges must be considered because the pipelines must be along the street to lay. According to graph theory, WDS can be viewed as a graph. The intersections of the street is the vertices or nodes, and the route laying between the intersections are known as edges. For large systems it contains hundreds to thousands of nodes and links, hence difficult to control the structure of the system and interactions of its components. WDS can be classified into three categories. *Layout*, for analysing the system connectivity and topology. *Design*, to specify the sizes of the system for the given layout. *Operation*, is for the designed system.

The task of discovering natural grouping of input patterns or clustering is an important aspect of machine learning and pattern analysis. Spectral clustering algorithm which clusters the data using eigenvectors of similarity/affinity matrix is derived from a data set. For modelling large amount of data, clustering techniques are more frequently adopted by various research committees.

Clustering is an important research field in data mining. The purpose of clustering is to divide a dataset into natural groups. Spectral clustering is a clustering method based on algebraic graph theory. It became popular due to its solid theoretical foundation, as well as the good performance of clustering. Spectral clustering does not make any assumptions on the global structure of the data. It can converge to global optimum and performs well for the sample space of arbitrary shape. The idea of spectral clustering is based on spectral graph theory. It can be

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

proved that the classified information of vertices is contained in Eigen values and Eigenvectors of graph Laplacian matrix. And we can get good clustering results if we make full use of the classified information during clustering process. Spectral clustering algorithms provide a new idea to solve the problem of clustering and can effectively deal with many practical problems. Therefore this paper has great scientific value and application potential.

Spectral clustering has been successfully applied to many areas such as data analysis, speech separation, video indexing, character recognition, image processing etc., Spectral clustering is a powerful approach for clustering, which has been widely used in many fields. Especially in the graph and network areas. There are mainly three reasons, why spectral clustering attracts so many researchers. Firstly, it has a solid theoretical foundation-algebraic graph theory. Secondly, for complex cluster structure, it can get a global solution. Thirdly, it can solve the problem within a polynomial time.

Spectral clustering uses the eigen values and eigenvectors of a matrix associated to the network, it is computationally very efficient, and it works for any choice of weights. We carry out the clustering on a weighted graph with the weights measuring pair wise distances. Clustering nodes in a graph is a useful general technique in data mining of large networks. In recent years, much attention has been paid to spectral clustering algorithm. Clustering is the process to reveal some structure implied in patterns into sensible clusters, that is assigning data objects into natural groups so that the data objects within the same cluster have high similarity while the data objects belonging to different clusters have low similarity. In order to efficiently find out groups in heterogeneous data, a large number of algorithms have been developed to perform cluster analysis. Among them, there is category known as clustering algorithm based on minimum spanning tree(MST), which are capable of detecting clusters of various shapes.

The construction of similarity matrix is an important aspect in the performance of spectral clustering algorithm. Spectral clustering has recently became one of the most popular clustering algorithms. Compared with traditional clustering techniques, spectral clustering exhibits many advantages and is applicable to different on the construction of similarity matrix. Ng-Jordan-Weiss (NJW) method is one of the most widely used spectral clustering algorithms. For a K clustering problem, this method partitions data using the largest K eigenvectors of the normalized affinity matrix derived from the dataset. It has been demonstrated that the spectral relaxation solution of K-way grouping is located on the subspace of the largest K-eigenvectors.

Worldwide growing water demand has been forcing utilities to successfully manage this costs. An efficient urban water management is needed to get a balance between consumer satisfaction and infrastructural assets indurent to WDN. Spectral clustering is usually adopted for network analysis tasks. Example: Community (or) sub network discovery. A graph based analysis is proposed to improve leakage management in water distribution networks.

The graph is then analysed in the eigenspace of its Normalized Laplacian matrix and specifically into the eigensubspace with much higher computational requirements, to be applied also to large problems. The results obtained in the eigenspace are eventually mapped back into the physical space where the capacity of leakage localization may be further improved through the fusion with leak severity estimation.

Urban water distribution networks suffer mainly due to the age of their pipeline infrastructure, frequent leaks and failures leading to service disruptions, large amounts of non revenue water, higher energy and rehabilitation costs (Puust, 2010). A more smart management of urban water distribution networks (WDN) is therefore needed to achieve higher levels of efficiency. This paper investigates the benefits provided by a new clustering methods based on eigen values analysis compared to other classical partitioning strategies (K-means, K-medoids, etc.,).

## II. RELATED WORK

The studies on constructing an exact MST starts with Boruka's algorithm (1926), similar algorithm invented by G.Choquet (1938), K. Florek (1951), M. Sollin (1965) respectively. One of the most popular prim's algorithm, was proposed by Jornik (1930), Prim (1957), and Dijkstra (1959) which selects a vertex as a tree and then repeatedly adds the shortest edge that connects a new vertex to the tree, until all the vertices are included. Kruskal's algorithm (1956) is another widely used exact MST algorithm, in which all the edges are sorted by their weights in an increasing order. It starts with each vertex being a tree, and iteratively combines the tree by adding edges in the sorted order excluding those leading to a cycle, until all the trees are combined into one tree.

J.C. Gower and G.J.S. Ross(1969) proposed, two hierarchical clustering algorithms which first employed to Minimum spanning trees and single linkage cluster analysis. Guan-Wei Wang et.al(2014), proposed most popular MST-based clustering algorithm, based on identifying inconsistent edges. Y. He and L. Chen(2005), proposed a threshold criterion, auto-detection and its use in MST-based clustering. P. Foggia et.al(2007), proposed a graph-based clustering method and its applications using Fuzzy C-means clustering algorithm to obtain two clusters of the edges of an MST. The edges with small weights in a cluster will be preserved while those belonging to the other cluster will be removed from the MST. Y.J. Li(2007), proposed a clustering algorithm based on maximal theta-distant sub trees.

Shvartser L., et.al (1993),explained about the Forecasting hourly water demands by pattern recognition approach. Zhou S.L., et.al (2002), explained about the Forecasting operational demand for an urban water supply zone. .


DOI 10.17148/IJARCCE.2016.5212

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

Preis, A. et.al (2010), proposed ,on-line hydraulic modelling of a Water Distribution System to identify the demand zones (i.e., clusters of water consumers) within the complex topology of the urban water supply system. Puust, R., et.al (2010), explained about the methods for leakage management in pipe networks. Herrera M., et.al (2010), proposed Predictive models for forecasting hourly urban water demand. Herrera, A.M., (2011), explained about Improving water network management by efficient division into supply clusters. Candelieri, A., et.al (2012), proposed , Clustering-based Services for Supporting Water Distribution Networks Management by implementing partition of the network into independent sub-sectors to perform district identification and leak localization on pipelines according to flow and pressure values continuously measured at crucial points of the network. Candelieri, A., et.al (2012), proposed, the application of data analytics approaches on flow and pressure data, continuously measured at crucial points of the network, for improving efficiency of leak localization and to reduce time and costs for physical check and consequent rehabilitation activities. Gutierrez-Perez, J., et.al (2012), proposed an approach as a support to the vulnerability analysis of Water Supply Networks (WSNs). The method is based on graph measurements such as the relative importance (ranking) and the degree of the vertices of a graph. Herrera M., et.al (2012), proposed an approach for Combining multiple perspectives on clustering: Node-pipe case in hydraulic sectorization.

Studies on spectral clustering began in 1973. Donath WE and Hoffman AJ(1973), introduced lower bounds for partitioning the graph based on the eigenvectors of adjacency matrix. Fiedler M(1973), proved that the bipartition of a graph using algebraic connectivity of the graphs is closely related to the second eigenvector of Laplacian matrix. Hagen L, Kahng AB (1992) found new spectral methods for radio cut partitioning and clustering using eigenvectors of similarity matrix. Shi J, Malik J (2000) proposed Normalized cuts and image segmentation, which considers the external connections between clusters and the internal connections within a cluster, so it can produce balanced clustering results. Ding CHQ, He X, Zha H et al (2001) proposed A min-max cut algorithm for graph partitioning and data clustering. Ng AY, Jordan MI, Weiss Y (2002) proposed classic NJW algorithm on spectral clustering. These algorithms are based on matrix spectral theory to classify data points, so they are called spectral clustering. Since 2000, spectral clustering has gradually become a research hotspot of data mining. At present, spectral clustering has been successfully applied to many fields, such as computer vision , integrated circuit design , load balancing, biological information, and text classification , etc. Spectral clustering algorithms provide a new idea to solve the problem of clustering and can effectively deal with many practical problems, so their research has great scientific value and application potential.

In this paper, we propose an algorithm for finding MST for a large water distribution network and also finding clusters for the MST. For MST of WDS , we have considered the features of Kruskal's algorithm for finding minimum spanning tree and for finding clusters for MST, we have considered the features of NJW, spectral clustering algorithm and applied them for real world data of Water distribution Network .

The rest of the paper is organized as follows. In section 3, the proposed approach is presented and explained using synthetic data . In section 4, Experimental results and analysis on real world Water Distribution Network is discussed. Finally, we conclude the work in section 5.

## III. METHODOLOGY

The proposed approach consists of mainly two phases. In phase-I, constructing MST for the large graph. In phase-II, minimum spanning tree clustering using spectral clustering for k- clusters. Block diagram of the proposed method is shown in Fig. 1.

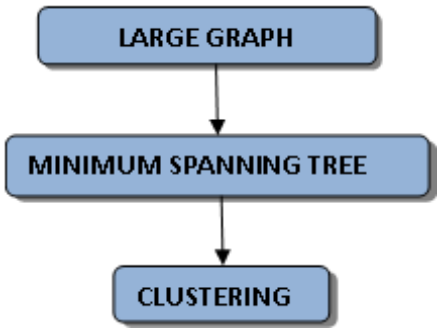
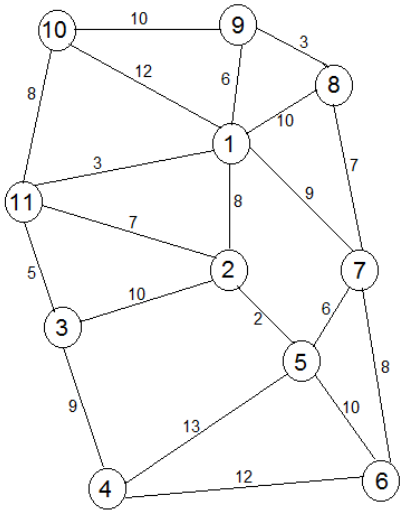



Fig. 1. Block Diagram of the proposed approach

### A. *Phase-I : Finding the MST for large graph*

MST for the graph in Fig. 2(a) can be constructed by using Kruskal's algorithm. Using this the edges are added in the sorted order excluding those leading to a cycle, until all the edges are combined into one tree. The sorted edges weights of the graph are 2,3,3,5,6,6,7,7,8,8,8,9, 9,10, 10, 10, 10,12,12,13. The MST formed with edge weights 2,3,3,5,6,6,7,8,8,9 excluding those edge weights 7,8,9, 10, 10,10,10,12,12,13 which are leading to a cycle as shown in Fig. 2(b).



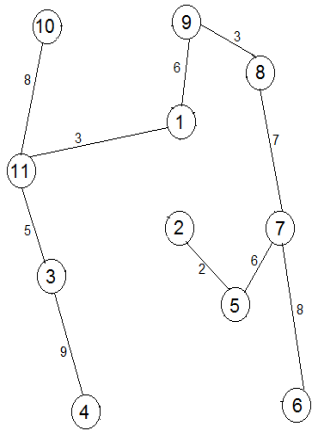

(a)


DOI 10.17148/IJARCCE.2016.5212

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*




(b)

Fig. 2. (a). Input Graph , (b). MST for the input graph

B. *Phase-II : Finding k-clusters for the MST*

MST for the input graph( Large graph) is partitioned into sub graphs or clusters using NJW spectral clustering algorithm, which is based on eigen values and eigenvectors of a matrix associated to the network. Adjacency matrix/similarity matrix denoted as A, and Laplacian matrix, denoted as L are commonly used representations for graph.

Adjacency matrix: n x n symmetric matrix

$A_{ij}$ = $W_{ij}$: weight of edge(i,j)

= 0 : if no edge between i,j

$$A=\begin{bmatrix}
0&0&0&0&0&0&0&0&6&0&3\\
0&0&0&0&2&0&0&0&0&0&0\\
0&0&0&9&0&0&0&0&0&0&5\\
0&0&9&0&0&0&0&0&0&0&0\\
0&2&0&0&0&0&6&0&0&0&0\\
0&0&0&0&0&0&8&0&0&0&0\\
0&0&0&0&6&8&0&7&0&0&0\\
0&0&0&0&0&0&7&0&3&0&0\\
6&0&0&0&0&0&0&3&0&0&0\\
0&0&0&0&0&0&0&0&0&0&8\\
3&0&5&0&0&0&0&0&0&8&0
\end{bmatrix}$$

$$D=\begin{bmatrix}
9&0&0&0&0&0&0&0&0&0&0\\
0&2&0&0&0&0&0&0&0&0&0\\
0&0&14&0&0&0&0&0&0&0&0\\
0&0&0&9&0&0&0&0&0&0&0\\
0&0&0&0&8&0&0&0&0&0&0\\
0&0&0&0&0&8&0&0&0&0&0\\
0&0&0&0&0&0&21&0&0&0&0\\
0&0&0&0&0&0&0&10&0&0&0\\
0&0&0&0&0&0&0&0&9&0&0\\
0&0&0&0&0&0&0&0&0&8&0\\
0&0&0&0&0&0&0&0&0&0&16
\end{bmatrix}$$

Laplacian matrix, L = D-A

where D is the diagonal matrix of degrees

$L_{ij}$ = $d_i$ : if i = j

= $-w_{ij}$ : if (i,j) is an edge

= 0 : if no edge between i,j

$$L=\begin{bmatrix}
9&0&0&0&0&0&0&0&-6&0&-3\\
0&2&0&0&-2&0&0&0&0&0&0\\
0&0&14&-9&0&0&0&0&0&0&-5\\
0&0&-9&9&0&0&0&0&0&0&0\\
0&-2&0&0&8&0&-6&0&0&0&0\\
0&0&0&0&0&8&-8&0&0&0&0\\
0&0&0&0&-6&-8&21&-7&0&0&0\\
0&0&0&0&0&0&-7&10&-3&0&0\\
-6&0&0&0&0&0&0&-3&9&0&0\\
0&0&0&0&0&0&0&0&0&8&-8\\
-3&0&-5&0&0&0&0&0&0&-8&16
\end{bmatrix}$$

The adjacency matrix of weighted graph uses real numbers to reflect the different relations between vertices. Most spectral clustering algorithms are based on the spectrum of Laplacian matrix to split graphs. There are two kinds of Laplacian matrices: Un-normalized Laplacian matrix(L) and normalized Laplacian matrix (denoted as $L_{norm}$ ). Laplacian matrix L=D-A, where D is a diagonal matrix, the diagonal values are equal to the absolute row sum of A, and the non diagonal elements are 0.

Normalized Laplacian matrix, $L_{norm} = D^{-1/2}AD^{-1/2}$, Where D is the diagonal matrix with non zero entries.

$L_{norm}$ = 1, if i=j,

= $-W_{ij}/\sqrt{d_i d_j}$, if i!=j and (i,j)ε E

= 0 otherwise

$$Lnorm=\begin{bmatrix}
1&0&0&0&0&0&0&0&-0.074&0&-0.021\\
0&1&0&0&-0.125&0&0&0&0&0&0\\
0&0&1&-0.714&0&0&0&0&0&0&-0.0223\\
0&0&-0.714&1&0&0&0&0&0&0&0\\
0&-0.125&0&0&1&0&-0.036&0&0&0&0\\
0&0&0&0&0&1&-0.048&0&0&0&0\\
0&0&0&0&-0.036&-0.048&1&-0.033&0&0&0\\
0&0&0&0&0&0&-0.033&1&-0.033&0&0\\
-0.074&0&0&0&0&0&0&-0.033&1&0&0\\
0&0&0&0&0&0&0&0&0&1&-0.0625\\
-0.021&0&-0.0223&0&0&0&0&0&0&-0.0625&1
\end{bmatrix}$$


DOI 10.17148/IJARCCE.2016.5212

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

Eigen values of the $L_{norm}$ , are

1.714 1.131 1.087 1.061 1.051 1 0.949 0.939 0.913 0.869 0.286

Eigenvectors of the $L_{norm}$ , are

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.001 | -0.001 | 0.015 | -0.604 | 0.271 | 0.057 | 0.015 | 0.343 | -0.604 | -0.271 | -0.057 |
| 0 | 0 | -0.665 | -0.07 | -0.202 | 0.112 | 0.665 | 0 | 0.07 | -0.202 | 0.112 |
| -0.707 | -0.707 | 0 | 0.001 | 0.001 | 0.002 | 0 | 0 | 0.001 | -0.001 | -0.002 |
| 0.707 | -0.707 | 0 | -0.01 | -0.007 | -0.019 | 0 | 0 | 0.01 | -0.007 | -0.019 |
| 0 | 0 | 0.698 | 0.049 | 0.083 | -0.055 | 0.698 | 0 | 0.049 | -0.083 | 0.055 |
| 0 | 0 | 0.088 | -0.07 | -0.543 | 0.23 | 0.088 | 0.528 | -0.07 | 0.543 | -0.23 |
| 0 | 0 | -0.24 | 0.127 | 0.583 | -0.295 | 0.24 | 0 | -0.127 | 0.583 | -0.295 |
| 0 | 0 | 0.067 | -0.285 | -0.21 | 0.274 | 0.067 | -0.768 | -0.285 | 0.21 | -0.274 |
| 0 | 0 | -0.025 | 0.623 | -0.255 | -0.216 | 0.025 | 0 | -0.623 | -0.255 | -0.216 |
| -0.002 | -0.002 | 0.001 | -0.218 | -0.284 | -0.604 | 0.001 | -0.115 | -0.218 | 0.284 | 0.604 |
| 0.022 | -0.022 | -0.003 | 0.302 | 0.234 | 0.594 | 0.003 | 0 | -0.302 | 0.234 | 0.594 |
| -0.28 | -0.231 | 0.399 | -0.4 | -0.28 | -0.231 | 0.399 | 0.04 | -0.065 | -0.638 | -0.065 |
| -0.281 | -0.019 | -0.085 | 0.36 | -0.281 | -0.019 | -0.085 | -0.36 | 0.533 | 0 | 0.533 |
| -0.351 | -0.223 | -0.414 | -0.394 | -0.351 | -0.223 | -0.414 | 0.394 | 0.007 | 0 | 0.007 |
| 0.281 | 0.19 | 0.387 | 0.483 | -0.281 | -0.19 | -0.387 | 0.483 | -0.039 | 0 | 0.039 |
| 0.265 | -0.353 | -0.025 | -0.04 | -0.265 | 0.353 | 0.025 | -0.04 | -0.55 | 0 | 0.55 |
| 0.265 | -0.477 | -0.065 | 0.12 | -0.265 | 0.477 | 0.065 | 0.12 | 0.428 | 0 | -0.428 |
| -0.351 | 0.595 | 0.074 | -0.104 | -0.351 | 0.595 | 0.074 | 0.104 | -0.082 | 0 | -0.082 |
| -0.162 | -0.16 | 0.379 | -0.163 | -0.162 | -0.16 | 0.379 | 0.163 | 0.08 | 0.737 | 0.08 |
| 0.28 | 0.261 | -0.565 | 0.184 | -0.28 | -0.261 | 0.565 | 0.184 | -0.02 | 0 | 0.02 |
| -0.265 | -0.131 | -0.112 | 0.42 | -0.265 | -0.131 | -0.112 | -0.42 | -0.446 | 0.225 | -0.446 |
| 0.449 | 0.209 | 0.163 | -0.465 | -0.449 | -0.209 | -0.163 | -0.465 | 0.109 | 0 | -0.109 |

From the above eigenvectors, selecting the largest k(=3), eigenvectors v1, v2, v3 are shown below

X=

| | | |
|---|---|---|
| -0.001 | 0.015 | -0.604 |
| 0 | -0.665 | -0.07 |
| -0.707 | 0 | 0.001 |
| 0.707 | 0 | -0.01 |
| 0 | 0.698 | 0.049 |
| 0 | 0.088 | -0.07 |
| 0 | -0.24 | 0.127 |
| 0 | 0.067 | -0.285 |
| 0 | -0.025 | 0.623 |
| -0.002 | 0.001 | -0.218 |
| 0.022 | -0.003 | 0.302 |

Renormalizing each row of X, to form Y matrix where, $Y_{ij} = X_{ij} / ( \Sigma X_{ij}^2)^{1/2}$

Y=

| | | |
|---|---|---|
| -0.00166 | 0.024827 | -0.99969 |
| 0 | -0.99451 | -0.10468 |
| -1 | 0 | 0.001414 |
| 0.9999 | 0 | -0.01414 |
| 0 | 0.997545 | 0.070028 |
| 0 | 0.782601 | -0.62252 |
| 0 | -0.88388 | 0.467719 |
| 0 | 0.228849 | -0.97346 |
| 0 | -0.0401 | 0.999196 |
| -0.00917 | 0.004587 | -0.99995 |
| 0.072652 | -0.00991 | 0.997308 |

Treating each row of Y as a point in $R^k$, cluster them into k - clusters applying k-means clustering algorithm, the clusters formed for different k values (i.e. for k=2,3,4,5) are shown below in the Table. I.

TABLE I: DETAILS OF CLUSTERS FOR DIFFERENT VALUES OF K

| Values of k | Clusters formed |
|---|---|
| 2 | C1={ 1,3,5,6,8,10 }<br>C2= { 2,4,7,9,11 } |
| 3 | C1={ 1,5,6,8,10 }<br>C2= { 3 }<br>C3= { 2,4,7,9,11 } |
| 4 | C1={ 4,11 }<br>C2= { 5 }<br>C3= { 1,3,6,8 }<br>C3= {2,7,9,11 } |

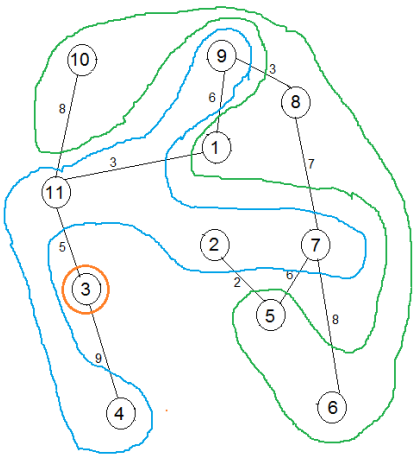



Fig. 3. Spectral Clustering for the MST in Fig. 2(b).


DOI 10.17148/IJARCCE.2016.5212

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

From the above results shown in Table. 1, considering the results for k=3, which are more appropriate and efficient. The spectral clustering results for the input graph are shown in Fig. 3.

*Algorithm*:

Input: Undirected weighted Graph G, Number of clusters K Output: Clusters $C_1$, $C_2$, .....$C_k$.

Step 1: Find MST for the input graph using Kruskal's algorithm.

Step 2: Compute the similarity/Adjacency matrix A for MST.

Step 3: Compute the diagonal matrix D.

Step 4: Compute the Laplacian matrix L = D - A.

Step 5: Compute the Normalized Laplacian matrix $L_{norm} = D^{-1/2}LD^{-1/2}$

Step 6: Compute eigen values and eigenvectors of $L_{norm}$

Step 7: Find $V_1,V_2,.....V_k$, the largest eigenvectors of $L_{norm}$ and from the matrix $X = [ V_1,V_2,...V_k] \varepsilon R^{nxk}$ by stacking the eigenvectors in columns.

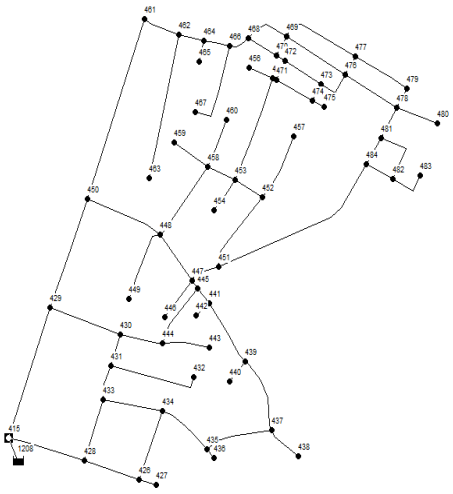
Step 8: Form the matrix Y, from X by renormalizing each of X's rows to have unit length

$$Y_{ij} = X_{ij} / (\Sigma_j X_{ij}^2)^{1/2}$$

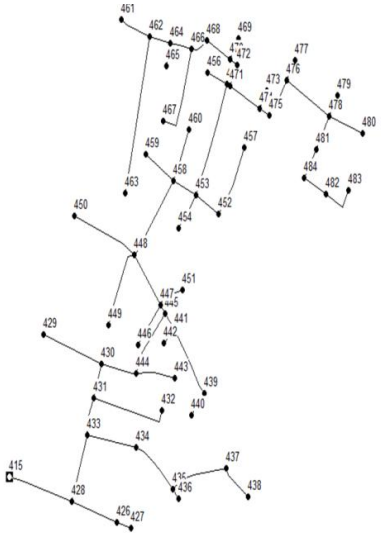
Step 9: Treating each row of Y as a point in $R^k$, cluster them into k clusters using k-means and compute the clusters $C_1$, $C_2$, .....$C_k$

## IV. EXPERIMENTAL RESULTS AND ANALYSIS ON LARGE WDS NETWORK

In this section, we present the experimental results on real world water distribution network of Balgaon, Parvathipuram in Andhra Pradesh as shown in Fig. 4(a). The distribution system is subjected to 60 junctions and 72 links which is having a reservoir of capacity 227 Kilo Litres. In this network the junction parameters are node id, elevation, head, demand, base demand, pressure and link parameters are link id, length, diameter, roughness, flow, velocity, unit head loss, friction factor. In the WDS network the junction represents the nodes and links represent the edges of a graph. In WDS, the node id is considered as junction or node parameter and length as link or edge parameter.



(a)



(b)

Fig. 4. (a). Input graph (WDS), (b) MST for the input graph.

The WDS, input graph in Fig. 4(a) is taken as a large graph and the MST for the input graph is obtained by using Kruskal's algorithm. Using Kruskal's algorithm, the edges of the input graph are sorted and connected to form a tree by avoiding cycle, which is nothing but the construction of MST as shown in Fig. 4(b). The length of each edge is considered as the weight of an edge.

MST for the input graph( Large graph), is partitioned into sub graphs or clusters using NJW spectral clustering algorithm, which is based on eigen values and eigenvectors of a matrix associated to the network. Adjacency matrix/similarity matrix denoted as A, represented with edge weights of MST. Laplacian matrix L=D-A, where D is a diagonal matrix, the diagonal values are equal to the absolute row sum of A, and the non diagonal elements are 0. Normalizing the laplacian matrix to construct $L_{norm}$ matrix. Then finding the eigen values and eigenvectors for the $L_{norm.}$ and resultant matrix is represented as X by considering larges k eigenvectors. Renormalizing the matrix X to get the matrix Y. Now finding the clusters by applying k-means algorithm on the matrix Y. The clusters formed for different k values are shown below.

For the value k=6, the clusters formed are

C1={ 461, 463, 464 }
C2={ 415, 426, 427, 432, 433, 435, 438 }
C3={ 447, 449, 450, 452, 454, 455, 458, 462, 467, 468, 469, 472, 474, 476, 479, 480, 481, 482 }
C4={ 429, 431, 440, 441, 444, 446, 448, 451, 453, 457, 459, 460, 465, 466, 470 }
C5={ 456, 471, 473, 475, 477, 478, 483, 484 }
C6={ 428, 430, 434, 436, 437, 439, 442, 443, 445 }

For the value k=7, the clusters formed are

C1= { 462, 465, 466, 470 }
C2= {415, 426, 427, 432, 433, 435, 438 }
C3= {449, 450, 454, 455, 458 }
C4= {446, 451, 457 }
C5= {461, 463, 464 }


DOI 10.17148/IJARCCE.2016.5212

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

C6= {428, 429, 431, 434, 436, 437, 440, 441, 444, 456, 467, 468, 469, 471, 472, 473, 475, 477, 478, 483, 484 }
C7= { 430, 439, 442, 443, 445, 447, 448, 452, 453, 459, 460, 474, 476, 479, 480, 481, 482}

For the value k=8, the clusters formed are

C1={440, 441, 444, 456, 471, 473, 475, 477, 478 }
C2={455, 474, 476, 479, 480, 481 }
C3={428, 429, 430, 431, 434, 436, 437, 439, 442, 443, 445, 447, 448, 452, 453, 459, 460 }
C4={446, 449, 450, 451, 454, 457, 458, 483, 484 }
C5={465, 466, 470 }
C6={415, 426, 427, 432, 433, 435,438,482 }
C7={462, 467, 468, 469, 472 }
C8={461, 463, 464 }

From the above results , considering the results for k=6, which are more appropriate and efficient. The spectral clustering results for the input graph are shown in Fig. 5.

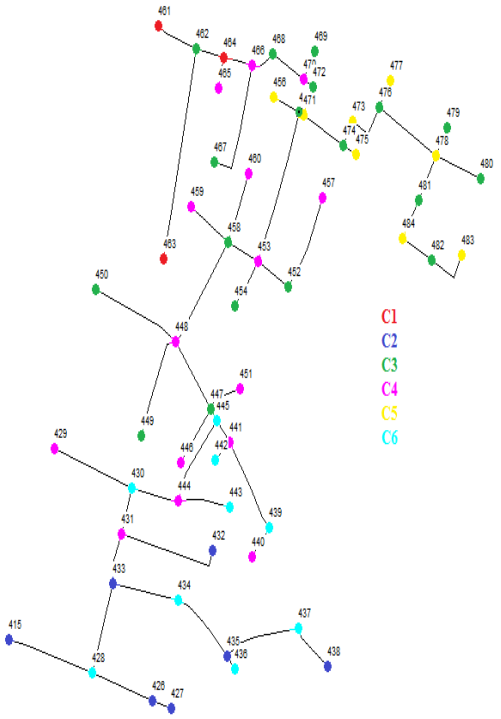



Fig. 5. Spectral Clustering for the MST in Fig. 4(b).

Applying spectral clustering on MST of WDS network, helps in identifying the location of the problem directly in the network. This helps in identifying the location and the problem immediately with minimum time and there is no need to check the entire WDS network.

Applying k-spanning tree clustering approach for MST, remove five large weighted edges i.e. large length pipes to form six clusters with non overlapping vertices. The clustering results are shown in Fig. 6.

The k-spanning tree clustering approach results for the real world WDS network are not efficient because the clusters C3, C4, C5 are only with nodes and there are no links/pipes connected to the nodes. Hence some pipes are missing. In this approach only length parameter of pipe is considered and the other parameters are ignored.

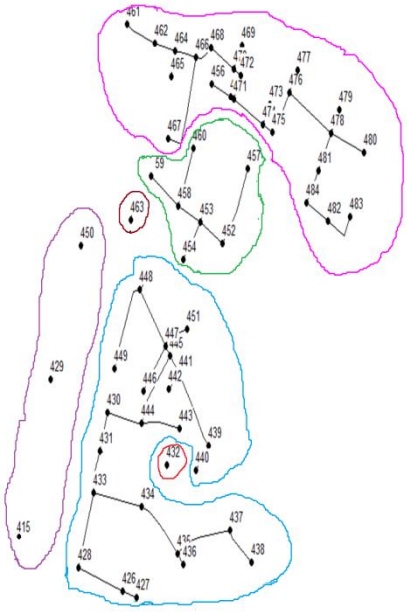


Fig. 6. k-spanning tree clustering for the MST in Fig. 4(b).

TABLE II: DETAILS OF NODE IDS IN EACH CLUSTER FOR K-SPANNING TREE CLUSTERING

| Cluster | Number of nodes | Node ids |
|---|---|---|
| C1 | 25 | 455,456,461,462,464,465,466, 467,468,469,470,471,472,473, 474,475,476,477,478,479,480, 481,482,483,484, |
| C2 | 7 | 425,453,454,457,458,459,460 |
| C3 | 1 | 463 |
| C4 | 3 | 415,429,450 |
| C5 | 1 | 432 |
| C6 | 23 | 426,427,428,430,431,433,434, 435,436,437,438,439,440,441, 442,443,444,445,446,447,448, 449,451 |

From the experimental results and analysis, the proposed approach shows an efficient results by considering eigen values and eigenvectors of a matrix associated to the network. For complex networks like WDS, spectral clustering is computationally very efficient.

## V. CONCLUSION

This study explores the solutions to reduce the topological and behavioural complexities and also for identifying crucial and critical components of WDS by implementing clustering methods to MST. By using this method we can design an efferent and economical WDS for any city or town. This involves finding the MST for the large graph and dividing the MST into clusters. By dividing into clusters it becomes easy to identify the crucial and critical components. This improves the efficient maintenance of WDS. The main objective of this paper is to maintain WDS efficiently with minimum time and cost. Experimental results on synthetic data and real world WDS shows the proposed approach and the efficiency for dividing the WDS into MST based clusters.


DOI 10.17148/IJARCCE.2016.5212

## REFERENCES

[1] von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395–416.
[2] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In NIPS, volume 2, pages 849–856, 2001.
[3] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.
[4] X. Wang, "A divide-and-conquer approach for minimum spanning tree-based clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp.945-958,2009.
[5] Candelieri, A., Messina, E., 2012. Sectorization and analytical leaks localizations in the H2OLeak project: Clustering-based services for supporting water distribution networks management. Environmental Engineering and Management Journal 11(5), 953-962.
[6] Candelieri, A., Archetti, F., Messina, E., 2013. Improving leakage management in urban water networks.
[7] Chen, WY, Song, Y., Bai, H., Lin CJ, Chang, E.Y., 2011. Parallel spectral clustering in distributed systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(3), 568-586.
[8] Chung, F., 1997. Spectral graph theory. Washington: Conference Board of the Mathematical Sciences.
[9] Fiedler, M., 1973. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal 23, 298–305.
[10] Hagen, L. and Kahng, A., 1992. New spectral methods for ratio cut partitioning and clustering. IEEE Transacations on Computer-Aided Design
[11] [Belkin2001] Belkin, M. & Niyogi, P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering Advances in Neural Information Processing Systems, 2001, 14, 585-591.
[12] [Belkin2002] Belkin, M. & Niyogi, P. Laplacian Eigen maps for Dimensionality Reduction and Data Representation Neural Computation, 2002, 15, 1373-1396.
[13] S. Yu and J. Shi. Multiclass spectral clustering. ICCV 2003.
[14] M. Meila, J. Shi. A random walks view of spectral segmentation, *AI and Statistics*, 2001.
[15] T. Xiang and S. Gong. Spectral clustering with eigenvector selection. *Pattern Recogn.*, 41(3), 2008.
[16] Rie Johnson and Tong Zhang On the Effectiveness of Laplacian Normalization for Graph Semi-supervised Learning, Journal of Machine Learning Research 8 (2007) 1489-1517.
[17] [Brandes et al., 2003] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In 11th Europe. Symp. Algorithms, volume 2832, pages 568–579. Springer, 2003.
[18] [Smyth and White, 2005] P. Smyth and S. White. A spectral clustering approach to finding communities in graphs. In Proc. of the 5th SIAM Int. Conf. on Data Mining (SDM), pages 76–84, 2005.
[19] [Yan et al., 2009] D. Yan, L. Huang, and M.I. Jordan. Fast approximate spectral clustering. In Proc. of the 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), pages 907–916, 2009.

## BIOGRAPHIES

**S.V.S. Santhi** is working as an Assistant Professor. She received her Masters degree in Information Technology from Andhra University in 2008 and is currently Pursuing her Part time PhD degree in Information Technology from GITAM University, Visakhapatnam, Andhra Pradesh, India. Her current research areas include data mining and graph mining.

**Padmaja Poosapati** is an Associate Professor in Department of Information Technology at GITAM University, Visakhapatnam, Andhra Pradesh, India. She received her Masters degree in Computer Science and Engineering from Andhra University in 1999 and PhD degree in Computer Science and Engineering from Andhra University, India in 2010. Her current research interests include Clustering and Classification in data mining and graph mining.


DOI 10.17148/IJARCCE.2016.5212