# Small-World-Like Structured MST-Based Clustering Algorithm

Sheila R. Lingaya, Bobby D. Gerardo, and Ruji P. Medina

***Abstract*—Graph-theoretic clustering is one method of clustering where dataset is represented with a connected undirected graph having the distance between these points as the weights of the links between them. One approach is the construction of the Minimum Spanning Tree of said graph where the connected subgraphs formed after the removal of an inconsistent edge are the clusters. However, such methods suffer with drawbacks including partitioning without sufficient evidence and robustness to outliers. Hence, this work aims to modify the Prim's MST-based clustering algorithm to produce a spanning tree of the dataset infusing the small-world network thereby invoking its properties (i.e. small mean shortest path length and high clustering coefficient) which manifest inherent or natural clustering characteristics.**

***Index Terms*—Graph-theoretic clustering, minimum spanning tree, small world networks, clustering coefficient.**

## I. Introduction

In data mining, clustering is one important technique [1] to reduce the data by means of categorizing or grouping similar data items together in order to achieve valuable information [2]. The principle is simply to achieve high intra-cluster similarity based on a measure derived from the data itself, and low inter-cluster similarity where elements in separate clusters are maximally apart from each other. However, clustering algorithms' main drawback is their performance being affected by the shape and size of the cluster to be detected [3] and grouping together noisy samples as a result of always trying to attribute each sample to a cluster is practically difficult because the number of cluster has to be provided in advance [4]. As stated in [5], there is no universal clustering method that can deal with all cluster problems because clusters in the real world can come in arbitrary shapes, varied densities and unbalanced sizes that is why graph-theoretic methods of clustering are also a focus of the cluster analysis research arena.

Graph-theoretic or graph-based clustering is where a dataset is represented with an undirected graph denoted as **G={V, E}** where **V** is the set of all data points of the dataset and **E** is the set of all links between said data points associated with a distance measure or weight. As such, a cluster is a connected graph. As expounded in [6], [7], graph-theoretic clustering algorithms either use Minimum Spanning Tree (MST) or limited neighborhood set approaches using the local properties of the graph as the basis or reference of the partitioning criterion. The construction of a MST approach in graph-theoretic clustering has the distance between data points as the weights of edges and the MST leads to establishing clusters [8] which are the connected components that the MST construction algorithm creates after a certain point. In [9], MST-based clustering is expound to basically consist of three primary steps, namely: construction of the MST, identification of inconsistent edge which is the primary difference between algorithms of this kind, and removal of the inconsistent edge to form the clusters. However, such methods suffer with problems such as not being able to detect outliers [8], eliminating inconsistent edges [10] and partitioning without sufficient evidence since only the information about the edge included in the MST is used to partition while the information about the other edges is lost that such methods are assumed to have all the same weaknesses as other distance-based methods [3].

As such, one approach to solve said problems is to enhance the MST construction algorithms like the Prim's which uses a distance function to specify the closeness of objects to establish the weight between the data points. It works by choosing an arbitrary data point and construct the tree by adding or connecting this arbitrary data point to the next or adjacent vertex of the minimum weight. When used in clustering as indicated in [8], the algorithm constructs a fully connected graph of the data set such that the edge weights are distances between the data points then constructs an MST followed by the search and removal of inconsistent edge to get the set of connected components – hence, the cluster. However, the constraint is how to determine the appropriateness of deleting the edge in order to proceed to the terminating condition, robustness to outliers [11], and computational complexity being expensive which obstructs the application to large scale data sets [3].

This work aims to modify the Prim's MST-based graph-theoretic approach to clustering to produce spanning tree of the dataset by invoking the properties of the small-world network structure through the utilization of the infused principle of local clustering coefficient in the MST construction for clustering.

This paper is organized as follows. Section II will present the theoretical foundation or background of the proposition. Section III will present the conceptual framework of the modified Prim's MST-based clustering algorithm invoking the properties of the small-world network of graph theory. Section IV concludes this paper focusing on future directions of work.

Manuscript received December 22, 2018; revised June 16, 2019.

S. R. Lingaya is with the Technological Institute of the Philippines, Quezon City, Philippines (e-mail: srlingaya@tau.edu.ph, bobby.gerardo@gmail.com, ruji.medina@tip.edu.ph).

doi: 10.18178/ijmlc.2019.9.4.829

## II. Theoretical Framework

### A. Prim's Minimum Spanning Tree

Prim's Minimum Spanning Tree is praised for being usable on distrusted machines as well as on shared memory machines and that it can be made to run on linear time. It is also noted more significantly for being faster in the limit when dealing with really dense graphs with more edges than vertices although it has a slow step in its loop compared to other algorithms.

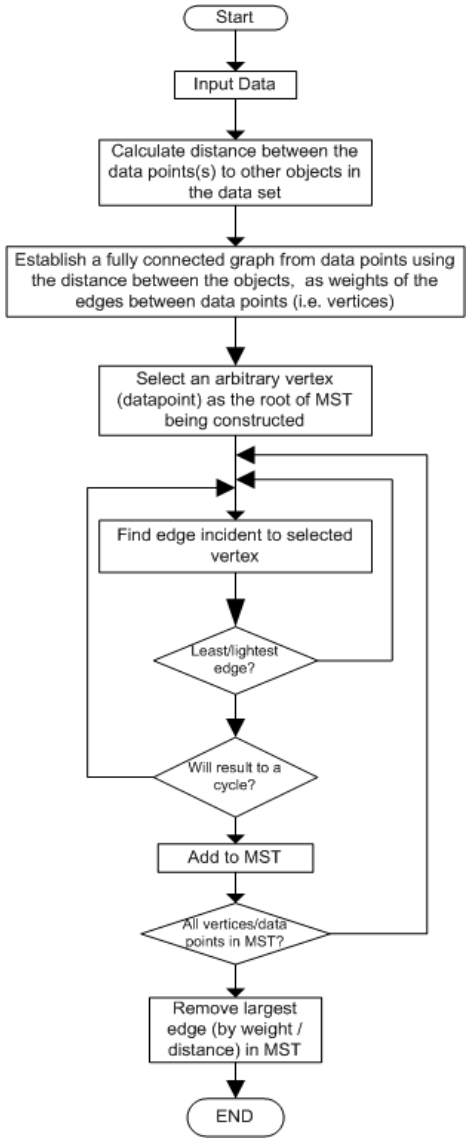


Fig. 1. Prim's MST-based clustering.

In [12], Prim's was used to construct MST under the assumption that the vertices are approximately distributed according to a spatial homogenous Poisson process and the number of clusters can be accurately estimated by thresholding the sequence of edge lengths added to the MST by the authors' Prim's trajectory. On the other hand, the Fibonacci heap is also used with Prim's MST for more efficiency [3], [9] while parallelizing the MST-based clustering algorithm.

On the other hand, [13] used Prim's as the technique of generating a MST as a result of growing a subtree while [14] used it to emphasize taking into consideration the intra-cluster quality.

Fig. 1 shows the flowchart of the Prim's MST construction algorithm applied in clustering as described by [8] where it can be posited that graph-based clustering based on Prim's MST is closely motivated by its close conformity to the "proximity" principle with its utilization of a distance function specifying the closeness of data points – hence, establishing the weight between them as it works by choosing an arbitrary data point and constructs the tree by adding or connecting to the next vertex of the minimum weight.

The contribution of this study is how the Prim's MST-based clustering algorithm can be modified by infusing the small-world network thereby invoking its properties (i.e. small mean shortest path length and high clustering coefficient) which manifest inherent or natural clustering characteristics. That is, with [8] suggesting that it might be possible to use some other kind of networks in clustering or use other network models.

Also, [6] emphasized that MST construction can be integrated with neighborhood search to enhance MST-based clustering in modern large databases. This concept is noted to explore the local property of data points where for two points to be connected by an edge in an MST, [15] emphasized that at least one is the nearest neighbor of the other – thus, the efficiency of constructing an MST is determined by the number of comparisons of the distances between two data points. That is, it's not necessary to search the whole dataset but a small local portion to find nearest neighbors or to construct an MST in a complex graph, it is also not necessary to sort all edges but to find the edges of least weights.

The infusion is focused on the problems of MST-based clustering algorithms particular to partitioning without sufficient evidence and robustness to outliers with the local efficiency assessing connectedness of the edges among neighbors of a given data point offering a network's local robustness to a node's removal [16].

### B. Cellular Evolutionary Algorithm

To infuse the small-world network properties; there is a need to establish first the neighborhood of the vertices or data points, a prerequisite of the computation of the local clustering coefficient of the data points since small-world networks have the property of having a high clustering coefficient and small mean shortest path length.

A number of works approached graph-theoretic clustering by utilizing the concept of k-nearest neighborhood for its advantage of zero cost of learning process and procedure is simple for local approximation and is usable to learn complex concepts. In [17], a shared nearest neighbor (SNN) graph from a single RNA-seq data set and applied MST-based clustering algorithm to cluster the graph nodes. However, the algorithm's inability to deal with outliers was identified as limitation as it can produce bad clustering result for the datasets having a large number of outliers. The kNN is also being focused as for instance, [7] explored reverse kNN as the neighborhood density estimation model while [18] established the kNN for each data point, then a mini MST was constructed upon each data points and its k-nearest neighbors. Finally, a small number of outliers were identified relatively within each mini MST using a proposed outlier score.

But the high clustering coefficient and small mean shortest

path of small-world networks emanate in real world systems processes like the cellular evolutionary algorithms (cEA). Hence, the neighborhood to which the adjacency of vertices is defined is based on cEA's principle of being a kind of evolutionary algorithm in which individuals cannot mate arbitrarily but everyone interacts with its closer neighbors.

This principle is invoked to provide a connected graph of the dataset in which each vertex is an individual communicating with its nearest neighbors. As such, one principle is the von Neumann Neighborhood of cEA where for range **r**, neighborhood is defined as

$$N^{v}_{(x_0,y_0)} = \{(x,y) : | x - x_0 | + | y - y_0 | \leq r\} \quad (1)$$

Hence, the number of neighbors (cells) in 2 dimensional is defined as

$$2r(r+1)+1 \quad (2)$$

The von Neumann Neighborhood is classically defined in a two-dimensional square lattice composing of central cell and its four adjacent cells explored in north, east, west, and south at a Manhattan distance of 1. It is one of the two most commonly used neighborhood types for two-dimensional cellular automata that an extension of this is taking the set of points at a Manhattan distance of $r > 1$ resulting into a diamond-shaped region. This work emulates the process treating the data points as the cells.

### C. Small-World Network Properties

The unique processing or information transfer capabilities of small-world networks paved way for the curiosity of whether these properties are across all natural networks or restricted to a number of networks [19]. The discovery of small world networks, as further expound in [19], has revolutionized research in network science when [20] specifically described networks to be highly clustered like regular lattices yet manifests small path lengths.

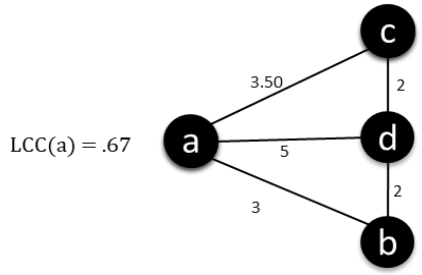


Fig. 2. Local clustering coefficient of node a.

The neighborhood established for an arbitrary vertex (or the vertices) for a given **r** can be used to establish the minimum or ideal distance between two points – hence, adjacency. This is since the edges of an MST comes from local neighborhood principle, the points which are far apart are not connected by an edge in the MST [21] as what is being constructed is MST based on small world network with high clustering coefficient and mean-shortest path length is small facilitated by the von Neumann Neighborhood.

The concept of small world network is that most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a minimal number of steps.

In graph theory, local clustering coefficient of a vertex in a graph quantifies how close neighbors are being a clique basically computed as the number of triangles connected to vertex over the number of triples centered on the given vertex.

As shown in Fig. 2, the local clustering coefficient of node **a** is two pairs of neighbors over three neighbors. Hence, the basic definition of local clustering coefficient is

$$LCC = \frac{\text{no.\_of\_connected\_neighbors}}{\text{no.\_of\_neighbors}} \quad (3)$$

In other words, it is the ratio of the number of pairs of neighbors that are connected and the number of possible couples of neighbors of the vertex or data points – hence, the probability that duos of neighbors of a vertex are connected defined by an immediate connection. The value is

$$0 \leq LCC \leq 0 \quad (4)$$

## III. Modified Prim's MST-Based Clustering Algorithm

Recent observation is that there should be a good cluster around high degree vertices in real-world networks with a power law distribution [22]. As such, the steps in this part of the algorithm include the Prim's MST taking into account the local clustering coefficient and distance between points is formally invoked.

### A. Minimum Spanning Tree Construction

The MST construction by Prim's Algorithm primarily considers and uses the distance between points or from the reference vertex towards its adjacent data points. In this work, the LCC of the adjacent vertices are priority reference that when all adjacent vertices have equal local efficiency, the algorithm will refer to the distance.

The coefficient is defined by its adjacent vertices (i.e. all other points besides itself) and the number of neighbors (i.e. adjacent vertices of this point) being adjacent to each other.

The traditional Prim's MST construction algorithm utilizes a priority queue to facilitate the efficient finding of candidate edge to the reference vertex which respects the disjoint of the data points in the MST and the set of vertices in the data set (i.e. distance between two data points) using the information **(u, v, d(u, v))**. Thus, to reflect the properties of the small-world network bearing on the LCC, the four tuple information **(u, v, LCC(v), d(u, v))** is processed where **u** and **v** are the two data points whose distance is represented by **d(u, v)**. The **LCC(v)** noted by the algorithm is that of the vertex or data point **v.**

Given **G={V, E}** representing the graph of data points with **V** as the set of vertices and **E** as the set of pairs of distinct vertices (i.e. edge); a set of the edges of the MST being constructed is established. Hence, $E_{MST}$ is a subset of E with $V_{MST}$ being the set of vertices of the MST – thus, a subset of the V set of vertices of the graph.

Fig. 3 depicts a MST constructed from a weighted

undirected graph with a priority reference on the LCCs of points and the distance between the points as secondary basis. It can be observed that points with high LCCs are inside the neighborhood.

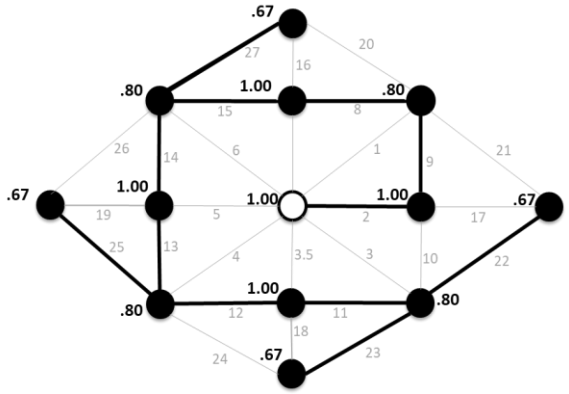


Fig. 3. MST constructed by Prim's in an undirected graph infusing the local clustering coefficient.

According to [23], most of graph-based clustering algorithms consist of two important steps, that is, (1) constructing the similarity matrix and (2) selecting a suitable partition approach which is called "cut". As such, the candidate edge to be added in the MST is the edge of least distance incident to a node that should respect the cut ($V_{MST}$, V- $V_{MST}$) partition of the V, to avoid a cycle in the MST being built. Note that in a simple undirected graph **d(u,v) = d(v,u).** This process of the MST construction infused with the reference to the LCC primarily is defined in the following.

**Algorithm 1. Modified Prim's MST-based Clustering with Small World Like Structure**

procedure MST Clustering (V: set of data points; N: set of points in neighborhood of vertex)
  select arbitrary vertex v from V
  establish neighborhood N of v
    such that the edge weights between data points are distances
    compute local clustering coefficient for every data point in the neighborhood of v
    construct a minimum spanning tree T of N
      find all adjacent data points with lowest LCC
      remove as the inconsistent edge from T
        such tat distance to data point of lowest LCC is the greatest from any data point in the T
  define connected components as clusters

Traditionally, Prim grows the MST one edge at a time. On the other hand, Fig. 4 depicts that when adding a vertex in the $V_{MST}$ , the modified MST-based algorithm considers the LCC of the adjacent vertices of the node last added and that of which highest will be prioritized to be added in the MST. If there is an equal LCC, the algorithm then turns to the distances from the reference vertex and its adjacent vertices. Until such time that the MST contains **N-1** edges for **N** number of vertices or data points.

### *B. Removing Inconsistent Edge*

According to [10], inconsistent edges of MST are removed by defining an error ration based on weights of the edges sorted in non-increasing order. Basically, these edges are removed until error ratio is greater than a certain value. The authors generalized that graph-based clustering is hierarchical with clusters being formed recursively by using agglomerative or divisive modes. Furthermore, a graph is constructed which builds similarity between the data points. Then a criterion is proposed to remove the inconsistent edges of the graph to collect the densely packed or connected components which are nothing but the required clusters.

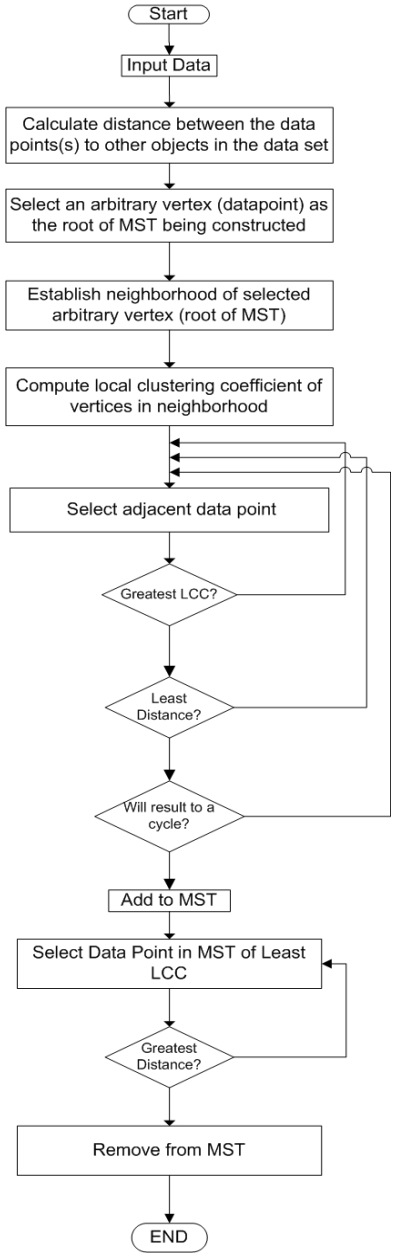


Fig. 4. Modified prim's MST-based clustering algorithm infused with small-world network properties.

On the other hand, the work of [24] has been referred to as the foundation of MST-based clustering which expound that an edge whose weight is significantly larger than the average of nearby weights on both sides of this edge should be deleted. Significantly here means how many standard deviation separate this edge's weight is from the average weights on each side or that the faculty or ratio between this weight and their respective averages can be calculated. Hence, the factor

of inconsistency which is the ratio between edge weight and average of other nearby edge weights where factor of 2 means separation is apparent.

In an MST-based clustering, algorithm needs to decide for the appropriateness of deleting a candidate inconsistent edge.

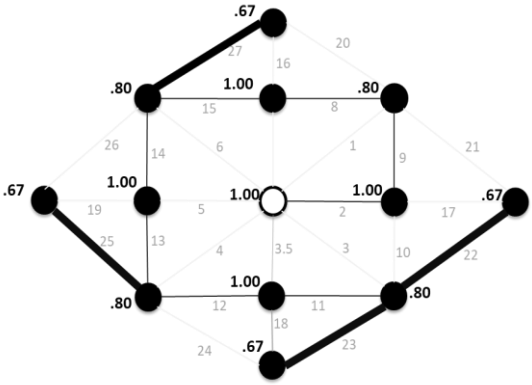


Fig. 5. Candidate edges of the MST infused with LCC.

Fig. 5 shows that four nodes are with LCC=.67. These are the candidate inconsistent edges which are incident to the points with the lowest LCCs in the neighborhood – hence, basis for appropriateness of edge inconsistency.

In this work, the inconsistent edge is that incident to the vertex having the lowest LCC. In Fig. 6, the inconsistent edge is removed – thus, the resulting MST with $|E_{MST}| = 11$ for $N$=12 vertices.

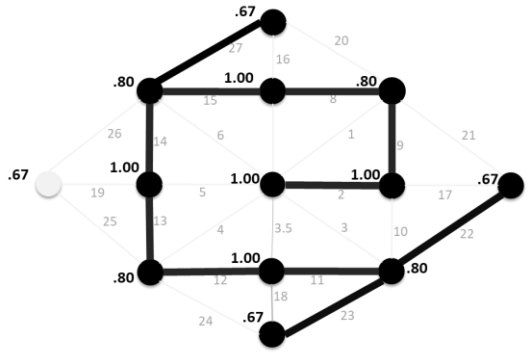


Fig. 6. Resulting MST of the modified prim's MST.

In case of equal LCCs, the greatest distance between these vertices of low LCCs and any vertex in the MST will be removed. The algorithm referred to the distance to these points – hence, the edge inconsistency is defined by low local clustering coefficient of end data point and greater distance. The resulting connected subgraph herein is the resulting cluster.

## IV. Conclusion

The main contribution of this study is how the Prim's MST-based clustering algorithm was infused with the properties of a small-world network as a modification by invoking the local clustering coefficient as a reference of the MST construction algorithm. The small-world network properties (i.e. small mean shortest path length and high clustering coefficient) were pursued as they manifest natural clustering characteristics. The neighborhood defined by the von Neumann Neighborhood established the adjacency of the data points that is necessary to compute the LCC defined as the number of connected neighbors over the total number of neighbors of a data point.

For future works, the validation of the clustering result will be focused on utilizing any of the three methods namely, internal, external or relative cluster validity indices in order to establish the appropriate performance measurement of graph-theoretic clustering algorithms. Also, the algorithm can be extended to focus on outlier detection and the algorithm will also be explored in real-world application.

## References

[1] S. Das, *Cluster Analysis for Overlapping Clusters Using Genetic Algorithm*, pp. 6–11, 2016.

[2] B. Kenidra, M. Benmohammed, A. Beghriche, and Z. Benmounah, "A partitional approach for genomic-data clustering combined with K-means algorithm," in *Proc. 2016 IEEE Intl Conf. Comput. Sci. Eng. IEEE Intl Conf. Embed. Ubiquitous Comput. 15th Intl Symp. Distrib. Comput. Appl. Bus. Eng.*, 2016, pp. 114–121.

[3] C. Zhong, D. Miao, and P. Fränti, "Minimum spanning tree based split-and-merge: A hierarchical clustering method," *Inf. Sci. (Ny).*, vol. 181, no. 16, pp. 3397–3410, 2011.

[4] P. Foggia *et al.*, *A Graph-Based Clustering Method and Its Applications*, pp. 277–287, 2007.

[5] R. Xu and D. Wunsh II, "Survey of clustering algorithms for MANET," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[6] X. Wang, X. L. Wang, and J. Zhu, "A new fast minimum spanning tree-based clustering technique," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2015–Janua, no. January, pp. 1053–1060, 2015.

[7] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *IEEE Access*, vol. 5, pp. 4991–5000, 2017.

[8] N. Paivinen, "Clustering with a minimum spanning tree of scale-free-like structure," *Pattern Recognition Letters*, vol. 26, pp. 921–930, 2005.

[9] D. Elsayad, A. Khalifa, M. E. Khalifa, and E. S. El-Horbaty, "An improved parallel minimum spanning tree based clustering algorithm for microarrays data analysis," *No. Infos*, pp. 66–72, 2012.

[10] D. R. Edla, S. Machavarapu, and P. K. Jana, *An Improved MST-based Clustering for Biological Data*, pp. 42–47, 2012.

[11] C. Zhong, D. Miao, and R. Wang, "A graph-theoretical clustering method based on two rounds of minimum spanning trees," *Pattern Recognit.*, vol. 43, no. 3, pp. 752–766, 2010.

[12] L. Galluccio, O. Michel, P. Comon, and A. O. Hero, "Graph based k-means clustering," *Signal Processing*, vol. 92, no. 9, pp. 1970–1984, 2012.

[13] G. W. Wang, C. X. Zhang, and J. Zhuang, "Clustering with prim's sequential representation of minimum spanning tree," *Appl. Math. Comput.*, vol. 247, pp. 521–534, 2014.

[14] J. J. Kponyo, Y. Kuang, E. Zhang, and K. Domenic, "VANET cluster-on-demand minimum spanning tree (MST) prim clustering algorithm," in *Proc. 2013 Jt. Conf. Int. Conf. Comput. Probl. Int. High Speed Intell. Commun. Forum*, 2013, pp. 101–104.

[15] C. Zhong, M. Malinen, D. Miao, and P. Fränti, "A fast minimum spanning tree algorithm based on K-means," *Inf. Sci. (Ny).*, vol. 295, October, pp. 1–17, 2015.

[16] J. D. Medaglia, "Graph theoretic analysis of resting state functional MR imaging," *Neuroimaging Clin. N. Am.*, vol. 27, no. 4, pp. 593–607, 2017.

[17] P. Das and K. A. A. Nazeer, "A novel clustering method to identify cell types from single cell transcriptional profiles," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 983–992, 2018.

[18] X. Wang, X. L. Wang, Y. Ma, and D. M. Wilkes, "A fast MST-inspired kNN-based outlier detection method," *Inf. Syst.*, vol. 48, pp. 89–112, 2015.

[19] Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, and P. J. Laurienti, *The Ubiquity of Small-World Networks*, vol. 1, no. 5, 2011.

[20] D. J. J. Watts and S. H. H. Strogatz, "Collective dynamics of 'small-world' networks.," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[21] R. Jothi, S. K. Mohanty, and A. Ojha, "Fast approximate minimum spanning tree based clustering algorithm," *Neurocomputing*, vol. 272, pp. 542–557, 2018.

[22] J. J. Whang, X. Sui, and I. S. Dhillon, "Scalable and memory-efficient clustering of large-scale social networks," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 705–714.
[23] Z. Yu, X. Wang, H.-S. Wong, and Z. Deng, "Pattern mining based on local distribution," in *Proc. the International Joint Conference on Neural Networks*, 2008, pp. 584–588.
[24] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Trans. Comput.*, vol. C-20, no. 1, pp. 68–86, 1971.

**Sheila R. Lingaya** is a current student with a doctor in information technology program in the Technological Institute of the Philippines, Quezon City. She has also got a master of science in information technology from the Tarlac State University, Tarlac City, Philippines in 2011. She is also an assistant professor at the Tarlac Agricultural University, Tarlac, Philippines.

**Bobby D. Gerardo** is a lecturer at the Technological Institute of the Philippines and also connected to the College of Information and Communications Technology of the West Visayas State University, La Paz, Iloilo City, Philippines. His research interests include distributed systems, telematics systems, CORBA, data mining, web services, ubiquitous computing and mobile communications. Dr. Gerardo is also referee to international conferences and journal publications such as in *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Knowledge and Data Engineering*.

**Ruji P. Medina** is currently the dean of the Graduate Programs of the Technological Institute of the Philippines, Quezon City. He has got the doctor of philosophy in environmental engineering from the University of the Philippines. His reseach areas are in environmental modeling and mathematical modeling using multivariate analysis.