

Modified Graph-theoretic Clustering Algorithm for Mining International Linkages of Philippine Higher Education Institutions

Sheila R. Lingaya¹, Bobby D. Gerardo², Ruji P. Medina³
Technological Institute of the Philippines - Quezon City, Philippines^{1,3}
Tarlac Agricultural University¹
West Visayas State University²

Abstract—Graph-theoretic clustering either uses limited neighborhood or construction of a minimum spanning tree to aid the clustering process. The latter is challenged by the need to identify and consequently eliminate inconsistent edges to achieve final clusters, detect outliers and partition substantially. This work focused on mining the data of the International Linkages of Philippine Higher Education Institutions by employing a modified graph-theoretic clustering algorithm with which the Prim's Minimum Spanning Tree algorithm was used to construct a minimum spanning tree for the internationalization dataset infusing the properties of a small world network. Such properties are invoked by the computation of local clustering coefficient for the data elements in the limited neighborhood of data points established using the von Neumann Neighborhood. The overall result of the cluster validation using the Silhouette Index with a score of .69 indicates that there is an acceptable structure found in the clustering result – hence, a potential of the modified MST-based clustering algorithm. The Silhouette per cluster with .75 being the least score means that each cluster derived for $r=5$ by the von Neumann Neighborhood has a strong clustering structure.

Keywords—MST-based clustering; Small World Network; von Neumann Neighborhood; internationalization; Prim's MST

I. INTRODUCTION

Internationalization and partnership development undertakings pave way to establish identity in the international arena. As such, data in the field of internationalization as mirrored by students and international partnerships established by education institutions is growing to be a good interest of researches [1]–[4]. This is since the rate of internationalization increases with the unhindered channels of communications and affordable travel expenses. Universities seek to seize the opportunities from global partnerships and foster relationships with other organizations or institutions. Internationalization is also described to transform into mainstream strategy in higher educations and is increasingly seen as adding value to the life of universities through improving their quality [5]. The definition of internationalization being the process of integrating international, intercultural, or global dimensions into the purpose, functions or delivery of post-secondary education [6] is by common knowledge, the most frequently cited and widely accepted.

Meanwhile, methods and techniques in data mining allow analysis of very large datasets (i.e. big data) to extract and discover previously unknown structures and relations out of huge amount of details [7] for the purpose of knowledge extraction. As such, clustering in the data mining arena aims to establish high intra-cluster and low inter-cluster similarity in data. The high intra-cluster similarity should be based on the derived measurement from the data while the low inter-cluster similarity should maintain that elements in the different clusters should have maximum distance. These are intended to achieve beneficial knowledge from the data [8] for decision making and strategizing. Among different types of clustering, the most conventional distinction is whether the set of clusters is hierarchical or partitional [9] where hierarchical is a set of nested clusters while partitional clustering divides the set of data objects into non-overlapping clusters such that each object is in exactly a single cluster [10]. However in the real world, clusters come in arbitrary shapes, varied densities and unbalanced sizes that is why there is no universal clustering method which can deal with all problems [11].

Since most clustering algorithms' performance is affected by the shape and size of the detectable clusters [12], the requirement of an *a priori* knowledge about the actual number of clusters and the setting of a threshold to obtain adequate clustering results; a number of modifications to the clustering algorithms have emerged and are being explored to cope with said problems. Among which are graph-theoretic or graph-based clustering algorithms where data is represented in an undirected graph denoted as $G=\{V,E\}$ where the set of all data points is V and the set of connections between two distinct data objects (i.e. edges or links are contained in E . This is associated with a distance measure resulting to a connected subgraph or clusters. The use of Minimum Spanning Tree (MST) is one of said methods which either uses the Prim's [13]–[15] or Kruskal's [16], [17]. An MST is constructed for the whole data with a threshold value being set along with a number of steps to terminate the process to form clusters resulting from removing an inconsistent edge whose value is greater than the threshold value. However, this strategy is constrained by the identification and elimination of the inconsistent edge [17], detection of outliers [18], as well as insufficiently evidenced partitioning—hence, having the same weaknesses as other clustering methods that are based on distance measures [19].

This work aims to perform data mining in the data of the international linkages of Philippine Higher Education Institutions (PHEIs) using a proposed modified Prim's MST-based clustering algorithm producing a minimum spanning tree for the dataset infusing the computation of local clustering coefficient for the data points in the limited neighborhood generated by von Neumann Neighborhood.

This paper is organized as follows. Section II presents the conceptual framework of the modified Prim's MST-based clustering algorithm invoking the properties of the small-world network of graph theory. It also highlights the preparation of the International Linkages data. Section III includes the results of the simulation and the cluster validation. Section IV highlights the conclusions and future works of the study.

II. MODIFIED PRIM'S MST-BASED CLUSTERING ALGORITHM

Clustering can be used on many problems as it is helpful to seek and see relationships. It aims to congregate into clusters unlabeled data elements with high similarity based on a measure obtained solely from the data itself [20]. The distance measure defines the radius of membership depending on the type of data on hand. A good cluster is associated with high clustering value in terms of distance so the selection of distance metric is essential in clustering [21] while another clustering algorithm approach is to represent a target data set as a weighted undirected graph [20].

A. Prim's MST-based Clustering Algorithm

Prim's MST Algorithm uses a distance function to specify the closeness of data objects to establish the weight between them by choosing an arbitrary point to the next adjacent point of minimum weight. For clustering, an edge inconsistency measure is defined to identify an inconsistent edge to be removed to partition the whole dataset into clusters. Prim's MST is modified for efficient construction of spanning tree based on the k-nearest neighbor search mechanism during which a new edge weight is defined to maximize the intra-cluster similarity and minimize the inter-cluster similarity [13]. The algorithm can be used for a complete graph while using Fibonacci Heap [19], [22].

In this work, the traditional Prim's MST Algorithm for clustering defined by [18] as shown in Fig. 1 is modified by infusing the local neighborhood search by the von Neumann Neighborhood in order to facilitate the computation of the local clustering coefficients of the data elements in said neighborhood.

Higher clustering coefficient indicates the robustness on an average shortest path between any pair of nodes [23]–[25]. As such, small world networks [26] have the properties of having a small mean of shortest path length and high clustering coefficient. The Local Clustering Coefficient (LCC) quantifies the closeness of the neighbors of a vertex in becoming a clique. A concept in graph theory, LCC is basically computed as the number of triangles connected to a vertex over the number of

triples around a given vertex. It is the probability that duos of neighbors of a vertex are connected by an immediate connection – the value is $0 \leq LCC \leq 1$. Thus,

$$LCC = \frac{\text{number of connected neighbors}}{\text{number of neighbors}} \quad (1)$$

Meanwhile, the von Neumann Neighborhood is one of the most commonly used types of neighborhood for cellular automata of two dimensions [27]. It is also used in pattern generation [28] and operations research [29] as it has been proven to have better performance than other topologies to further improve the quality of local search [30]. It can be extended by taking the set of data objects at Manhattan distance r where $r > 1$ which yields a result of a diamond-shaped region – hence, the neighborhood of data objects. The two-dimensional square lattice is composed of the central cell and the four adjacent cells around achieved through traversing North, East, West and South (NEWS) derived at a Manhattan distance 1. The number of neighbors (i.e. cells) in 2-dimensional by von Neumann Neighborhood of the cellular evolutionary algorithm for range r is defined as:

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

As such, the modified Prim's MST-based Clustering Algorithm establishes the adjacency of the data facilitated by the suitable cellular automaton, the von Neumann Neighborhood which simulates the establishment of neighborhood. This precludes the computation of local efficiency or local clustering coefficient. Thus, the modified Prim's MST construction for clustering is defined by $(u, v, LCC(v), d(u,v))$ such that u is the initial data point and v is the terminal data point.

While the traditional Prim's MST considers only the next minimum distance $d(u,v)$ between data u in the MST being built T and the adjacent data point v in V ; the modified algorithm initially considers the LCC of the adjacent data point $LCC(v)$ to ensure a high clustering coefficient for the whole cluster – hence, pursuing clusters of density. As Prim grows the MST one edge at a time, it should be noted that the next candidate edge or connection of data point must respect the partition or cut of the set of points in the minimum spanning tree T and V to avoid having a cycle.

```
Pseudocode for FMST for Clustering
procedure MST Clustering (V: set of data points v )
construct a fully connected graph G of V such that
the
    edge weights are the distances between data
    points
construct Prim's minimum spanning tree T of G
maintain disjoint sets V and T
select minimum d(v,u) where v ∈ V and u ∈ T
check for cycle
find all inconsistent edges of T
remove inconsistent edges to get a set of connected
components
define the connected components as clusters
```

Fig. 1. Prim's Minimum Spanning Tree for Clustering.

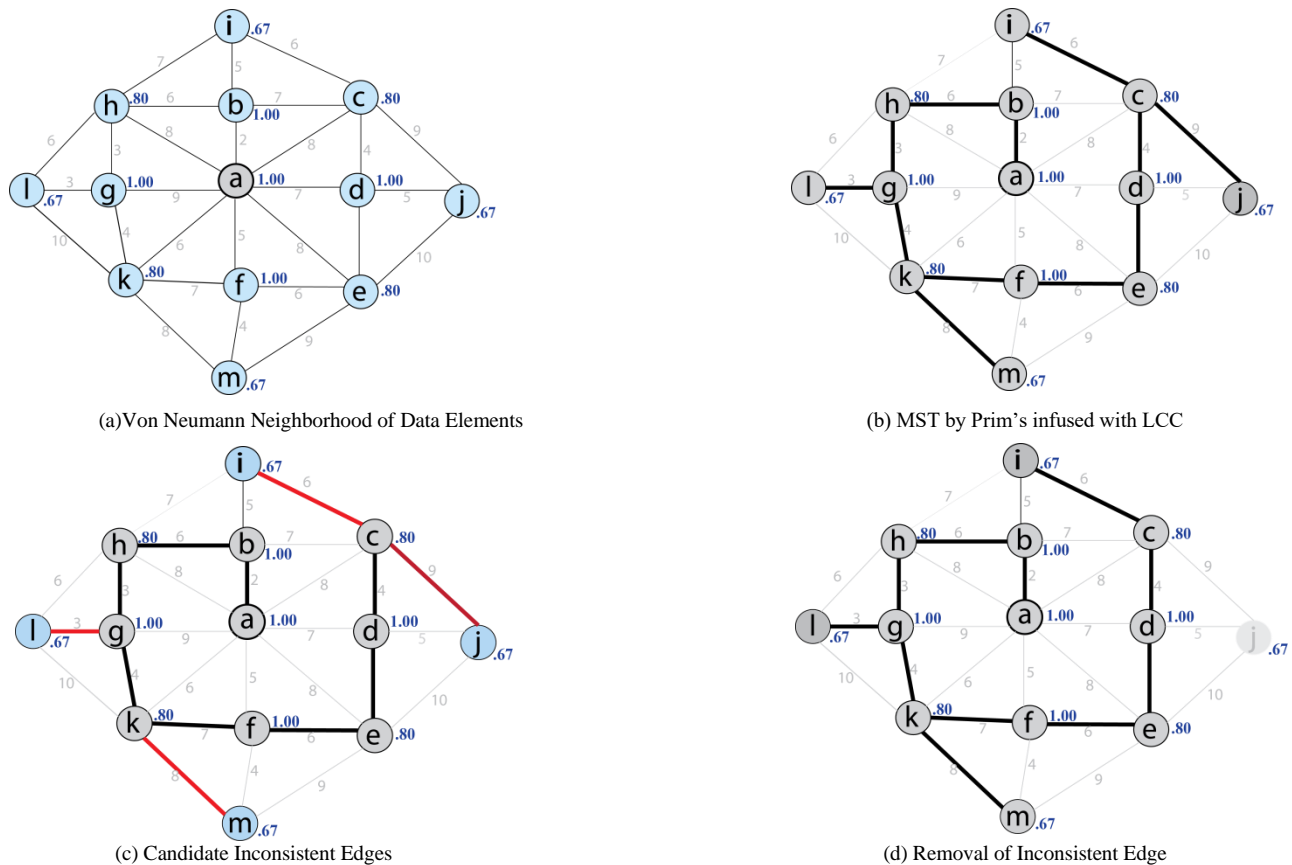


Fig. 2. Prim's Minimum Spanning Tree Construction for a Local Neighborhood Established using Von Neumann.

Being the data element having the least $LCC(v)$ and maximum distance d is the criterion set for identifying the inconsistent edge. Data elements l, j, k and m in Fig. 2(a) are all $LCC=.67$ – hence, their distances to the data points in the MST were considered as indicated in Fig. 2(b). As such, $d(c,j)=9$ indicated in Fig. 2(c) as the connection with the greatest distance is the inconsistent edge. The algorithm will herein iterate and continue on the other data points of the data set. The resulting MST must have $N-1$ edges for N number of data points without cycle – hence, the cluster as seen in Fig. 2(d).

B. Data Cleaning and Preparation

The PHEI International Linkages data contains the actual and essential records for the international linkages of Philippine Higher Education Institutions. It is consist of partnerships entered by PHEIs with foreign universities and/or organizations transpiring into different internationalization activities including student exchange, faculty exchange, academic collaboration, research collaboration and other activities across different disciplines. The dataset is summarized in Table I.

An integral part of the data mining process is the data to which knowledge discovery is applied. The International Linkages data contains instances of inconsistencies, incompleteness and variations in the essence of data mining. As such, entries or values were simplified and prepared such that the proposed algorithm is able to process it. In the original

data, the field for partnership form has duplicative entries and no defined options. A particular discipline is mentioned in several groups with each specific to a particular partnership. A similar case can be observed with an area of partnership (e.g. Faculty Exchange) being included and specific in a number of partnerships. Hence, in Table II are the disintegrated attributes rooted from the form of linkages attribute of the original data.

TABLE I. LIST OF PHEI LINKAGES DATASET FEATURES AND DESCRIPTION

Name	Definition	Example
country	where foreign university or organization partnered is located	Indonesia
continent	where country of foreign university or organization partnered is located	Asia
phei	the Philippine Higher Education Institution (e.g. SUC, HEI)	TAU
partner	name foreign university or organization partnered	CRRU
p_form	form of partnership	Bilateral
p_area	area of internationalization activities	Faculty Exchange
p_discipline	field of discipline covered by the partnership	Education
d_signed	date when partnership was signed	02/06
p_year	year when partnership was signed	2017
p_status	if active or inactive	Active

TABLE II. CONTENT RELATED FEATURE OF ATTRIBUTES DISINTEGRATED FROM FIELD

Name	Definition
p_type	Bilateral; Multilateral
p_area	Faculty Exchange; Student Exchange; Research(er) Exchange/Collaboration; Academic Collaboration; Joint Publication
p_discipline	Accounting, Arts, Education, Fashion and Textiles, Social Studies

The conversion of the textual values was necessary since most instances are texts and multiple values are specific to one entry. The data cleaning and preparation executed is where each distinct group is coded. For instance, in the area of partnership terms, the PHEI can either use its own nomenclature but certainly, it may also use the terms of reference by the prospect foreign partner university or organization. Hence, all attributes were coded and assigned a numerical value to discretize the data so that the clustering algorithm will be able to process it.

III. RESULTS AND DISCUSSION

The cluster analysis of the data on international linkages of PHEIs aimed to gain valuable insights of the data to see what groups the data elements belong to while having the modified clustering algorithm to define instances with similar properties as a group. Data may come into mix type in the real world such that one attribute may be expressed in ration and others in terms of categorical that adjustment may be hard in terms of the algorithm because some specific algorithms can only be applicable to certain types of data. There may be a need for some data transformation or preprocessing to do so that the algorithm will work. Data cleaning and preprocessing was an integral part of the data mining process to make adjustments and the data be made suitable with the proposed algorithm as it cleaned and prepared the data for the algorithm to be able to process it.

A. Simulation

The algorithm was implemented through the following Pseudocode in Fig. 3 and simulated on the discretized Internationalization data set.

```
Pseudocode for Modified Prim's MST-based Clustering with Local Efficiency  
  
procedure MST Clustering (V: set of data points v )  
  Remove all redundant data  
  set arbitrary data point  
  get arbitrary data point's Neighborhood  
  generate connection for each data point in neighborhood;  
  set LCC for each data point in neighborhood  
  construct Prim's minimum spanning tree T of G  
  maintain disjoint sets V and T  
  set data point with least LCC and maximum distance as  
  inconsistent edge  
  remove inconsistent edges to get a set of connected  
  components  
  define the connected components as clusters
```

Fig. 3. Modified Prim's MST-based Clustering with Local Efficiency.

The International Linkages data set is composed of 12 attributes with 748 instances. With a random value $r=5$, seven clusters were generated. The attributes with only at most 2 possible values were not used for the experiment.

Two attributes (e.g. continent, pheI) were used to define an instance-hence to illustrate, data point (x, y) defines one data element by its value on attributes continent and pheI as x and y , respectively. The neighborhood of said data points determined by NEWS was derived with the nearest higher value in x for north, nearest lower value in x for south, nearest higher value in y for east, and nearest lower value in y for south until the prescribed number of neighbors of the arbitrarily chosen value through von Neumann's Neighborhood is derived.

An observation on the result of the presented data mining procedure is that the generation of edge or connection between the data points to form the neighborhood impacts the processing time of the algorithm. The complexity of this part of the modified algorithm is also challenged when the data points are not linear. The choice of value for r also is also critical as a minimum choice will produce more clusters which impact the inter-cluster separation.

B. Cluster Validation

As there is no optimal clustering algorithm [31], it is necessary to evaluate the generated clusters of the mining process on the International Data. One approach is an internal validation with which the concentration is the partitioned data such that the compactness and separation of the clusters are measured. The Silhouette index [32] is where the silhouettes show which objects lie well within their partition and which are somewhere between clusters. The silhouettes herein were formed basically by knowing the clusters or partitions generated by the modified clustering algorithm and the distance between the data points-hence, a data point i 's distance to other points within the cluster it belongs to and to other data points in other clusters.

The average distance $a(i)$ of a data point i to all other objects in the cluster it belongs to is computed in the same manner that the average distance $b(i)$ to other objects in other clusters is also derived. Hence, the silhouette score is derived as:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (3)$$

The Silhouette index is chosen for the validation of the resulting clusters of the proposed graph-theoretic clustering algorithm in order to observe how well the algorithm partitioned the data set [33]. The focus is also on the quality of the clustering structure being measured only using information or feature intrinsic to the data set [34]. Another salient point in choosing the Silhouette index for cluster validity is because it measures attributes taken from the data, itself and the clusters found [35]. The silhouette scores ranging from $-1 \leq s(i) \leq 1$ can be interpreted in Table III.

The validation on the clustering result generated by the modified graph-theoretic clustering algorithm infused with small-world network structure based on the Silhouette score is presented in Table IV which presents the silhouette score of the clustering result. The average intra-cluster distance was derived

from calculating the distance of a random point (x, y) in a cluster towards all other data points in the same cluster to which it belongs to. Inter-cluster distance is the distance of this (x, y) towards the other data points in other clusters.

TABLE III. SILHOUETTE SCORE INTERPRETATION

Range	Description	Interpretation
0.71 – 1.00	Strong	A strong structure has been found.
0.51 – 0.70	Reasonable	Reasonable structure has been found.
0.26 – 0.50	Weak	Structure is weak and could be artificial.
≤ 0.25	Not Substantial	No substantial structure has been found.

TABLE IV. SILHOUTTE VALIDATION ON CLUSTERING RESULT

Cluster ID	Average Intra-Cluster Distance	Average Inter-Cluster Distance	Silhouette Score
1	2.57	14.65	0.81
2	2.46	9.65	0.75
3	3.09	15.82	0.80
4	3.20	19.83	0.84
5	2.86	16.64	0.83
6	2.70	15.82	0.83
7	-	15.11	0.00

The average Silhouette score derived as 0.69 indicates an acceptable structure was found which is also manifested in the scores of all the clusters which derived scores not lower than 0.75 which means that each cluster has a strong structure except for Cluster 7 which has only one (1) data point – hence, silhouette score is 0. Such constraint is present to prevent the number of groups from significantly increasing [36]. Consequently, when a clustering result is interpreted based on Table III, the clustering is acceptable when the score is at least 0.50 [37].

IV. CONCLUSION

This work performed data mining in the international linkages of Philippine Higher Education Institutions (PHEIs) data using a proposed modified Prim’s MST-based clustering algorithm producing a minimum spanning tree for the data set infusing the computation of local clustering coefficient for the data points in the limited neighborhood generated by von Neumann’s Neighborhood.

An integral part of this work was the preparation of the raw data to achieve the dataset that is ready for processing by the modified Prim’s MST-based clustering algorithm. The numerical attributes of the International Linkages dataset were used for the clustering to work on similarity on a particular parameter.

The results of the study show that there is an acceptable structure found in the clustering result with silhouette score 0.69 and 0.75 being the least score for the 6 out of 7 clusters derived for r=5 of the von Neumann Neighborhood.

However, the algorithm is still bound by the *a priori* input value of r which dictates the number of possible neighbors in one cluster for the von Neumann Neighborhood. As such the

optimum number of clusters and most ideal value of r for a particular size of data are interesting.

Also for future works, the interest is also centered on the cluster validation utilizing external validity indices particularly those which works or are specific to graph-theoretic clustering algorithms. The data can also be refined more and subjected to clustering process to compare the performance of the traditional and the modified clustering algorithm.

ACKNOWLEDGMENT

The authors would like to extend gratitude to the Commission on Higher Education International Affairs Staff (Philippines) for cooperation in the realization of this work by providing the data necessary for the study. Appreciation is also extended to the Tarlac Agricultural University as well as the Technological Institute of the Philippines – Quezon City.

REFERENCES

- [1] U. Teichler, “Internationalisation of higher education: European experiences,” *Asia Pacific Educ. Rev.*, vol. 10, no. 1, pp. 93–106, 2009.
- [2] J. Lawrence, “Internationalization of Higher Education in the United States of America and Europe: A Historical, Comparative, and Conceptual Analysis (review),” *Rev. High. Educ.*, vol. 27, no. 2, pp. 281–282, 2004.
- [3] D. Dutschke, “Campus Internationalization Initiatives and Study Abroad,” *Coll. Forum*, vol. 45, no. October, pp. 67–73, 2009.
- [4] W. Green and C. Whitsed, “Critical perspectives on internationalising the curriculum in disciplines,” *Crit. Perspect. Int. Curric. Discip. Reflective Narrat. Accounts from Business, Educ. Heal.*, no. February, 2015.
- [5] A. Aerden, F. De Decker, J. Divis, M. Frederiks, and H. de Wit, “Assessing the internationalisation of degree programmes: Experiences from a Dutch-Flemish pilot certifying internationalisation,” *Compare*, vol. 43, no. 1, pp. 56–78, 2013.
- [6] J. Knight, “Internationalization Remodeled: Definition, Approaches, and Rationales,” *J. Stud. Int. Educ.*, vol. 8, no. 1, pp. 5–31, 2004.
- [7] V. Vijay, V. P. Raghunath, A. Singh, and S. N. Omkar, “Variance based moving k-means algorithm,” *Proc. - 7th IEEE Int. Adv. Comput. Conf. IACC 2017*, no. i, pp. 841–847, 2017.
- [8] B. Kenidra, M. Benmohammed, A. Beghriche, and Z. Benmounah, “A Partitional Approach for Genomic-Data Clustering Combined with K-Means Algorithm,” *2016 IEEE Intl Conf. Comput. Sci. Eng. IEEE Intl Conf. Embed. Ubiquitous Comput. 15th Intl Symp. Distrib. Comput. Appl. Bus. Eng.*, pp. 114–121, 2016.
- [9] J. Chang, J. Luo, J. Z. Huang, S. Feng, and J. Fan, “Minimum Spanning Tree Based Classification Model for Massive Data with MapReduce Implementation,” *2010 IEEE Int. Conf. Data Min. Work.*, pp. 129–137, 2010.
- [10] P.-N. Tan, M. Steinbach, and V. Kumar, “Chap 8: Cluster Analysis: Basic Concepts and Algorithms,” *Introd. to Data Min.*, p. Chapter 8, 2005.
- [11] R. Xu and D. Wunsch II, “Survey of clustering algorithms for MANET,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.
- [12] 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.
- [13] 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.
- [14] L. Galluccio, O. Michel, P. Comon, M. Klinger, and A. O. Hero, “Clustering with a new distance measure based on a dual-rooted tree,” *Inf. Sci. (Ny)*, vol. 251, pp. 96–113, 2013.
- [15] 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.

- [16] 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.
- [17] D. R. Edla, S. Machavarapu, and P. K. Jana, "An Improved MST-based Clustering for Biological Data," pp. 42–47, 2012.
- [18] N. Paivinen, "Clustering with a minimum spanning tree of scale-free-like structure," vol. 26, pp. 921–930, 2005.
- [19] 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.
- [20] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013.
- [21] C. B. Abhilash, K. Rohitaksha, and S. Biradar, "A Comparative Analysis of Data sets using Machine Learning Techniques," *Adv. Comput. Conf.*, pp. 24–29, 2014.
- [22] 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.
- [23] L. H. Yen and Y. M. Cheng, "Clustering coefficient of wireless ad hoc networks and the quantity of hidden terminals," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 234–236, 2005.
- [24] M. R. Brust, D. Turgut, C. H. C. Ribeiro, and M. Kaiser, "Is the clustering coefficient a measure for fault tolerance in wireless sensor networks?," *IEEE Int. Conf. Commun.*, pp. 183–187, 2012.
- [25] C. H. Zeng and K. C. Chen, "Clustering coefficient analysis in large wireless ad hoc network," 2017 IEEE/CIC Int. Conf. Commun. China, ICC 2017, vol. 2018–Janua, no. Iccc, pp. 1–6, 2018.
- [26] 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.
- [27] D. A. Zaitsev, "A generalized neighborhood for cellular automata," *Theor. Comput. Sci.*, vol. 666, no. November, pp. 21–35, 2017.
- [28] U. Sahin, S. Uguz, H. Akin, and I. Siap, "Three-state von Neumann cellular automata and pattern generation," *Appl. Math. Model.*, vol. 39, no. 7, pp. 2003–2024, 2015.
- [29] N. Y. Soma and J. P. Melo, "On irreversibility of von Neumann additive cellular automata on grids," *Discret. Appl. Math.*, vol. 154, no. 5 SPEC. ISS., pp. 861–866, 2006.
- [30] X. Min, X. Xu, and Z. Wang, "Combining von neumann neighborhood topology with approximate-mapping local search for ABC-based service composition," *Proc. - 2014 IEEE Int. Conf. Serv. Comput. SCC 2014*, pp. 187–194, 2014.
- [31] O. Arbelaitz, I. Gurrutxaga, and J. Muguerza, "An extensive comparative study of cluster validity indices," vol. 46, pp. 243–256, 2013.
- [32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.
- [33] J. Shen, S. I. Chang, E. S. Lee, Y. Deng, and S. J. Brown, "Determination of cluster number in clustering microarray data," *Appl. Math. Comput.*, vol. 169, no. 2, pp. 1172–1185, 2005.
- [34] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis," *Bioinformatics*, vol. 22, no. 19, pp. 2405–2412, 2006.
- [35] D. N. Campo, G. Stegmayer, and D. H. Milone, "A new index for clustering validation with overlapped clusters," vol. 64, pp. 549–556, 2016.
- [36] F. Wang, J. D. Kelleher, J. Pugh, and R. Ross, "An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity," 2017.
- [37] R. C. De Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," vol. 324, no. 2015, pp. 126–145, 2016.