

Research Article

Clustering Networks' Heterogeneous Data in Defining a Comprehensive Closeness Centrality Index

Farnaz Barzinpour,¹ B. Hoda Ali-Ahmadi,¹
Somayeh Alizadeh,² and S. Golamreza Jalali Naini¹

¹ Department of Industrial Engineering, Iran University of Science and Technology (IUST), Narmak, Tehran 1684613114, Iran

² Department of Industrial Engineering, K. N. Toosi University of Technology (KNTU), 470 Mirdamad Avenue West, Tehran 19697, Iran

Correspondence should be addressed to B. Hoda Ali-Ahmadi; rahayesh@yahoo.com

Received 12 July 2014; Accepted 11 September 2014; Published 15 October 2014

Academic Editor: Carsten Proppe

Copyright © 2014 Farnaz Barzinpour et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

One of the most important applications of network analysis is detecting community structure, or clustering. Nearly all algorithms that are used to identify these structures use information derived from the topology of these networks, such as adjacency and distance relationships, and assume that there is only one type of relation in the network. However, in reality, there are multilayer networks, with each layer representing a particular type of relationship that contains nodes with individual characteristics that may influence the behavior of networks. This paper introduces a new, efficient spectral approach for detecting the communities in multilayer networks using the concept of hybrid clustering, which integrates multiple data sources, particularly the structure of relations and individual characteristics of nodes in a network, to improve the comprehension of the network and the clustering accuracy. Furthermore, we develop a new algorithm to define the closeness centrality measure in complex networks based on a combination of two approaches: social network analysis and traditional social science approach. We evaluate the performance of our proposed method using four benchmark datasets and a real-world network: oil global trade network. The experimental results indicated that our hybrid method is sufficiently effective at clustering using the node attributes and network structure.

1. Introduction

In recent years, there has been great interest in investigating and understanding the underlying mechanism of complex networks. These systems are generally symbolized using graph techniques that contain sets of nodes, which represent the objects under investigation that are connected together in pairs by links if the corresponding nodes are related by some type of relationship. Complex networks represent a fundamental area of multidisciplinary study and are found in physics, mathematics, chemistry, biology, social sciences, and information sciences; examples of the networks examined in these fields include the Internet, the World Wide Web, social networks, information networks, biological networks, neural networks, food webs, reaction and metabolic networks, and protein-protein interaction networks [1, 2].

In the literature of complex networks, one of the most important and widely used concepts, centrality, has been

defined to quantify the relative importance of a node in a network. This concept can be measured in several different ways; however, not every centrality index is suitable for every application. In general, the result of a centrality index depends solely on the structure of the graph [3–5].

The most common study of complex networks is the analysis of a network for a single context (layer). In a multilayered network, the researchers aim at obtaining a more comprehensive view of the network under study. They analyze different types of relations for each node that is represented in the network. Unfortunately, few articles have considered “multiplexity” in complex networks [6]. Wasserman and Faust [7] recommend the use of centrality measures in each layer because the aggregation of different relations can cause the loss of information due to the union of layers.

Network communities play essential organizational and functional roles in complex networks [8] and have been

increasingly applied in various fields. For example, communities in a social network indicate groups of different interests [9]. In protein-protein interaction networks, communities are expected to group proteins with the identical particular function within the cell. In the graph of the World Wide Web, communities may match groups of pages that address identical or associated topics. In metabolic networks, the communities are likely functional modules, such as cycles and pathways [10–13]. As a result, community detection in complex networks has become one of the most active fields of research in network theory (for review papers see [14, 15], and for a comparison paper see [16]).

In addition to aiding in understanding the network system by providing insights into the structure-functionality relationship [17], identification of communities in complex networks can also reduce complex networks to simpler systems. This ability is critical because there is an increasing requisite to consider extremely large real-world networks, and few presented methods can handle large graphs [18].

Community detection has a key function in determining some centrality measures because these measures assign valid centrality scores on the condition that the network presents a core-periphery structure in which all nodes revolve around a single core [19]. Thus, these types of centrality measures cannot be computed without such a cluster analysis.

Complex network research has been fundamentally concerned with the structure and effects of relations among nodes, rather than on individual attributes of the nodes. However, contribution of such attributes to the development and maintenance of ties among the nodes in networks that would thereby impact the behavior of those networks is disregarded. The development of a new paradigm in which both individual agency and social structure establish the network behavior presents an alternative to a rigorous structural perspective in which action is derived exclusively from the structure of relations in the networks [20].

This work introduces a novel framework for complex network analysis by considering the individual characteristics as well as structure of relations, and it presents a comprehensive and improved algorithm to determine the closeness centrality in multilayer complex networks that has five valuable specifications simultaneously.

- (i) The algorithm can cope with (un)weighted, (un)directed, single/multilayer graphs.
- (ii) It solves the core-periphery assumption problem by clustering as a preprocessing step.
- (iii) It maps the graph to separate nodes that are clustered using data clustering techniques and therefore overcomes some challenges of clustering methods in graphs, such as high computational complexity.
- (iv) Its dimensionality reduction (from $n \times n$ matrix to $n \times k$ matrix, ($k \ll n$)) significantly simplifies the computations.
- (v) It considers heterogeneous data that are related to the network's nodes and edges, which improves the network comprehension and the clustering accuracy.

We empirically assess the performance of such an integrated analysis by experimenting on four benchmark datasets.

The remainder of this paper is organized as follows. In Section 2, the necessary background and notations are provided, the current main problems are derived, and their solutions are introduced. Then, the proposed method based on these solutions is represented in Section 3. Section 4 applies the proposed algorithm to the benchmark datasets, with which our algorithm obtains competitive results. The final section of the paper contains brief concluding remarks.

2. Related Concepts and Background

2.1. Centrality. There is no commonly accepted definition of the centrality index; however, the least common ground for all centralities is that it denotes an order of importance on the vertices or edges of a graph by assigning them real values. On the other hand, a centrality index only depends on the graph structure [5]. Although many measures of centrality have been proposed (some revisions of centrality measures may be found in [5, 21]), four measures, degree, closeness, betweenness, and eigenvector centrality, have dominated the empirical usage. Their distinction within the field of network analysis arises from the fact that all of them have strong yet significant theoretical bases and have been considered foundational in the field [22]. Thus, this paper concentrates solely on those measures with special importance and influence.

- (i) Degree centrality measures the number of connections at a specific node and illustrates the level of potential communication activity from a specific node. An individual with greater degree centrality can directly communicate with others more easily [6].
- (ii) Closeness centrality considers being the closest to others on average. It also represents a level of independence for a specific node because a node is more autonomous and thus has higher independence when it can communicate with many other nodes and has a minimum number of intermediaries [23].
- (iii) Betweenness centrality focuses on the importance of a node in the communication between any node pair in the network. If a vertex lies on many shortest paths between other vertices, it plays a central role in information flows and is responsible for the system vulnerability [11].
- (iv) Eigenvector centrality is based on the largest characteristic eigenvalue of the adjacency matrix. In other words, it assesses the centrality of a person as a function of the centrality of the people with whom the person is associated [19].

Some previously considered centrality measures, including all degree, closeness, and eigenvector like measures, calculate the walks that emanate from or terminate with a given node. Social network researchers refer to these measures as radial measures. Another category of centrality measures, such as all betweenness-like measures, which assess the number of walks that pass through a given node, are called medial measures [19].

According to Borgatti and Everett [19], before interpreting a radial centrality measure, one must determine whether the network satisfies the one-group requirement. In other words, the radial centrality indices make sense in networks with at most one center, which would not be partitioned into two or more components.

A common notion in SNA and other fields is the core/periphery structure. Given its wide currency, the lack of an identifying community structure before computing a radial measure of centrality is a gap in the related literature. One point is that the medial centrality measures do not make the same one-group assumption. However, it is difficult to interpret a given value of medial centrality without specifying the group's cohesive structure.

2.2. Community Detection in Multilayer Networks. A network community, which is also called a cluster or module, refers to a group of vertices that likely share common properties and/or play similar roles in the graph [15]. Some important aspects of a community structure include the following.

- (1) Concrete applications, such as topic related Web pages clustering, image segmentation, discovering groups of individuals sharing identical properties, and detecting communities in a protein regulatory network that indicate groups of function-related proteins.
- (2) Classification of vertices based on their structural position in the communities. Determining clusters and their boundaries allows one to identify vertices with a central position in their clusters, namely, sharing many edges with the other cluster members, which may have an important function of control and stability within the cluster. Furthermore, the vertices at the boundaries among the communities are important for mediation and lead the relationships and exchanges among different modules.
- (3) Identifying the hierarchical organization, including communities that are composed of smaller communities, which in turn contain smaller communities. In particular, a system that is organized in interconnected subgroups is generated and evolves more rapidly than an unstructured system [15].

A variety of methods and algorithms for community detection have been developed so far, ranging from traditional clustering methods in computer and social sciences, that is, graph partitioning, hierarchical, partitional, and spectral clustering to modern methods, which are divided into categories based on the type of approach, such as modularity-based methods, and dynamic algorithms [15].

Our learning algorithm is based on spectral clustering, which makes our issue an eigenvalue problem and uses k -means for the final cluster assignments.

In a real network, there are always various types of relations among the members, such as friendships, business relationships, and common interest relationships in a social network. Most existing algorithms assume that there is only one type of connection, which corresponds to a relatively homogenous relationship (such as Web page linkage), and

few studies have considered network descriptors in multilayer case [24–31]. However, a typical network can be analyzed for different “contexts” or at different “layers” which represent different types of relationship among the objects in the network.

For the community detection problem, these different layers can include different communities. To recognize a community with specific properties, it is necessary to identify which relation plays an important role in such a community. As a result, community detection in a multilayer network requires these relations to be combined according to their importance in reflecting the user's information need.

2.3. Considering Nodes Attributes. According to Wasserman and Faust [7], some fundamental principles that underlie the perspective of network analysis are as follows.

- (i) “Actors and their actions are viewed as interdependent instead of independent, autonomous units.
- (ii) Network models focusing on individuals view the network structural environment as providing opportunities for or constraints on individual actions.
- (iii) Network models conceptualize structure (mainly social, economic and political) as lasting patterns of relations among actors.”

Network analysis concerns theory, models, and applications that are presented in terms of relational concepts and mainly disregards individual attributes of the actors or the network members, although such attributes affect the formation and maintenance of relationships among actors in the network and thus influence the network behaviors.

To better investigate an interest, such as community detection or centrality in a network, it is reasonable to consider various data that are related to that interest, which are typically heterogeneous. Network analysis studies the relations, that is, the attribute of pairs of individuals (which is called dyadic attributes), so the canonical dataset can be presented as an entity-by-entity matrix, whereas studying the attributes of individuals (which are called monadic attributes) requires different canonical dataset, which can be considered an entity-by-attribute matrix. For example, in the social context, the first type of data involves a social network analysis, which is concerned with the structure and effects of relations among people, groups, or organizations, whereas the other refers to traditional social science, which examines the influence of psychological individual attributes in the society. In web page analysis, the two types are hyperlink versus textual content; and in gene analysis, they are metabolic pathway and gene expression.

However, the distinction between these two types of data is not always clear because there are methods to convert one type into the other type. Furthermore, in some cases, data can be collected and preserved either as 1-mode or 2-mode, depending on the preference of the researcher [20].

Nevertheless, the information based on the nodes' individual characteristics can occasionally specify similarities that are not visible to techniques that are solely based on the structure of relations and vice versa. For example, consider

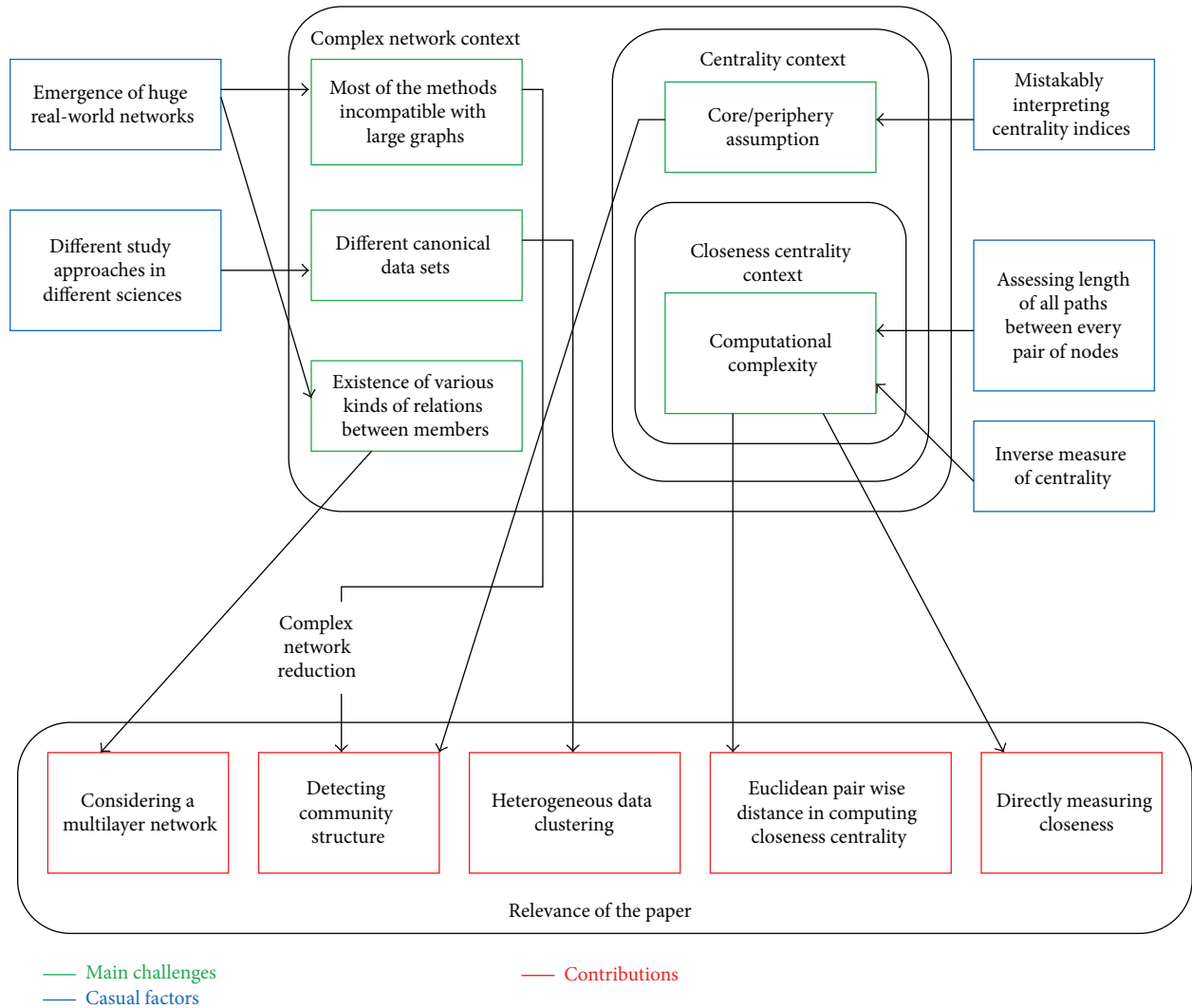


FIGURE 1: An analysis of the main challenges of the context.

the contents of web pages and their links; we typically encounter notably similar web pages even though there are no links among them. Thus, using only an existing graph clustering method cannot solve the problem [32, 33].

In addition, when used alone, each data source may suffer from critical shortcomings, such as noisy and unreliable data, incompleteness because of missing data, and bias. For example, in gene application, accurate gene clustering is essential for predicting the gene function; however cDNA microarray data are noisy and unreliable. Hence, as expected, another data source, such as a metabolic or gene regulatory network, is often required [32]. Some other limitations occur in applications related to mapping the science, including sparse matrices, documents with an insufficient number of references (such as letters), and the bias toward high-impact journals [33]. A potential approach to solve the problem is to integrate two data sources, which improves the reliability of the clustering results and thus the centrality measure. The idea of heterogeneous data clustering is not new [34–39]; however, most previous applications were designed to map

the cognitive structure of science and its adjustment over time by combining bibliometric or citation information with the textual content [33]. However, our focus is wider; namely, we consider clustering by combination of two heterogeneous data sources, the attributes of the nodes and the edges in a multilayer complex network, and subsequently finding the closeness centrality. Our integrating technique is clearly different from those that provide the opportunity to consider different layers and node attributes.

3. Proposed Method

3.1. Analysis of the Main Challenges of the Context. The following structure in Figure 1, which is based on conclusions that are reached by reviewing the related literature, classifies a series of challenges in the context and their casual factors and offers their corresponding solutions. The proposed solutions are applied to the proposed method, which is introduced for community detection and to compute the closeness centrality in multilayer complex networks.

The resulting structure illustrates the main challenges and gaps in the literature. In the complex network context, the first challenge is that motivating and challenging new cases of complex networks in recent years have resulted in increased attention and activity regarding the structure and dynamics of complex networks. In addition, network models have become standard tools in economics, social science, and the design of transportation and communication systems, which typically have multilayers as well. Because these networks are generally highly complex and most methods cannot address such large graphs, it is helpful to determine whether they can be reduced to simpler structures [18]. In fact, significant effort has been dedicated to dividing the networks into small numbers of communities. However, there are various approaches in different fields which study the same phenomenon; for example, the traditional social science studies the personal attributes of the members of a network, and social network analysis studies the attributes of their relations in the network under consideration. Matching different canonical datasets and considering both individual agency and social structure is the recommended solution to this challenge, which also increases the comprehension of the phenomenon.

Another considerable challenge that occurs in the centrality context is that radial measures may be incorrectly computed and interpreted due to the core/periphery assumption. Thus, the one-group requirement must be first satisfied, and if the network includes more than one component, the subgraph indices should be determined. The computational complexity in the closeness centrality context is the next challenge, which has been described in our previous work [20].

3.2. Community Detection and Closeness Centrality Index in Multilayer Complex Networks by Considering Both Structure of Relations and Individual Characteristics. Based on the analysis in Section 3.1, we introduce our proposed method, which consists of the following steps.

- (i) Checking the core-periphery assumption based on the spectral analysis [20].
 - (1) Given the input datasets (in the form of adjacency matrix such that $A_{ij} = 1$ if nodes i and j are connected by an edge and $A_{ij} = 0$ otherwise; or its weighted counterpart W).
 - (2) Define D to be the diagonal matrix whose (i, i) -element is the sum of A 's i th row (W for weighted graphs), which represents the degree of the node i . Then, form the normalized Laplacian, which is defined as follows:

$$L_{rw} = D^{-1}L. \quad (1)$$
 - (3) Find k nontrivial eigenvalues of the normalized Laplacian, L_{rw} , such that the first k eigenvalues of L_{rw} are close to zero and the $(k + 1)$ th is relatively large.
 - (4) If $k = 1$, then the network has a core-periphery structure, and the closeness centrality is computed using the common methods.

- (5) If $k > 1$, then construct a $n \times k$ matrix, V , whose columns are eigenvectors of L_{rw} .
- (ii) Considering individual characteristics and other layers.
 - (1) Add l attributes of individual nodes to have a $n \times (k + l)$ matrix Q as an input for clustering.
 - (2) If the 1-mode relation data that correspond to the other layers can be converted to 2-mode data, insert them as columns into matrix Q .
- (iii) Clustering using data clustering methods.
 - (1) Given the input matrix, Q , and a range of number of clusters, cluster the nodes using typical techniques, such as k -means.
 - (2) If the data that are related to any layer cannot be converted into 2-mode data, regard them as a type of dissimilarity among the nodes and insert them at pairwise distances throughout the clustering procedure.
- (iv) Computing the closeness centrality within the clusters.
 - (1) Determine the center of the clusters as nodes that have the highest closeness centrality in the clusters because they have the smallest pairwise distance among the cluster members.
- (v) Computing the closeness centrality of the network.
 - (1) Determine the closeness centrality among the cluster centers.

4. Experimental Evaluation

To study the behavior of a proposed community detection algorithm, it is necessary to have adequate benchmark networks in which the ground truth is known. A few benchmarks have been proposed to test communities in networks with node attributes. There is no such benchmark dataset with multiple layers. Thus, we had to restrict the empirical part of our research to single layer attributed networks. In our previous study [20], we examined the proposed algorithm using Zachary's karate club network, which is often used as a benchmark for community detection in networks. The preliminary results indicate that the proposed method efficiently detects both a good intercluster closeness centrality and an appropriate number of clusters.

Here, we evaluate the performance of the proposed method using four real-world networks from the WebKB dataset [40], for which we have network topology data and node attributes [41]. Moreover, explicit ground-truth community labels are accessible. The availability of such ground truth labels helps us quantify the degree of consistency between the detected and ground truth communities. The Rand index is used to compare the results obtained with the classification scheme.

The WebKB dataset consists of 877 web pages and 1,608 hyperlinks among them, which were gathered from different

TABLE 1: Dataset statistics.

| Dataset | Node | Edge | Node attribute | Community |
|------------|------|------|----------------|-----------|
| Cornell | 195 | 304 | 1703 | 5 |
| Texas | 187 | 328 | 1703 | 5 |
| Washington | 230 | 446 | 1703 | 5 |
| Wisconsin | 265 | 530 | 1703 | 5 |

university websites. These web pages are classified into one of the following five classes: course, faculty, student, project, or staff. Each web page in the dataset is described by a 0/1-valued word vector, which indicates the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1,703 unique words. All words with a document frequency less than 10 were removed. Table 1 lists the networks and their properties.

The main computational stages of our algorithm are performed using MATLAB; the Cluster Validity Analysis Platform (CVAP) [42] MATLAB tool was applied to evaluate the clustering results and compute the validity indices. Because eigenvalue and eigenvector calculation is required in primary steps of the proposed method, the adjacency matrix with zero rows or columns is not accepted. Thus, to overcome this challenge, we were obliged to ignore the link directions and make the adjacency matrices symmetric.

The results of the hybrid clustering analysis were compared to the predefined classification or the ground truth. The Rand index [43] was used to quantify the correspondence between the obtained clustering results (from 2 to 10 clusters) and the ground truth categorization. Table 2 illustrates that the Rand index in all cases is greater than 0.6 on average. The Rand index is greater in those bench mark networks, which have less sparse adjacency matrices and thus requires less adjustment.

5. Real-Life Example: Crude Oil Global Trade Network

The crude oil network data were gathered using the World Integrated Trade Solution (WITS) software [44], which was developed by The World Bank in collaboration with the United Nations Conference on Trade and Development (UNCTAD) and in consultation with various organizations, such as the International Trade Center, the United Nations Statistical Division (UNSD), and the World Trade Organization (WTO). This software provides access to international merchandise trade, tariff, and nontariff measure data. Data on crude oil bilateral trade around the world, including 175 countries, were provided from 2007 to 2010. The values were recorded in US dollars, and the computations were based on the mean bilateral trade that occurred in the study period. Due to missing bilateral trade data, 66 key countries were selected by consultant with oil trade experts, and the adjacency matrix of the network was formed.

According to the gravity model from international economics, which is claimed to be one of the great success stories in empirical economics [45], the trade between two countries is positively related to both of their economic

sizes and negatively related to the distance between them [46]. Therefore, we select the GDP per population, which represents the economic size of the countries as an individual characteristic of the nodes in the network, and consider the distance layer as a new effective layer in the bilateral trade network to form a multilayer attributed network.

The entire study consists of two steps. First, to analyze total bilateral trade between pairs of countries, the community structure of the network was detected and the clustering quality and number of optimized clusters were examined using internal, external, and relative validity indices. Second, to explain the finding (key countries in the global trade) from the analysis in the first step, we further determined the closeness centrality in the multilayer attributed oil trade network and analyzed the bilateral trade at the cluster level. This study helps oil trade policy makers, research analysts, academia, trade professionals, and others to better comprehend and navigate the world of energy trade.

Because clustering methods discover community structures, which are not known a priori, the concluding partitions of a dataset require some type of evaluation. Cluster validity indices are developed to quantitatively evaluate the results of the clustering algorithms. There are three approaches to investigating the cluster validity [43]. The first approach, based on external criterion, evaluates the results of a clustering algorithm based on a prespecified structure. The second approach is based on internal criteria, which implies that the clustering result is evaluated using only quantities and features that are inherent to the dataset. The optimal number of clusters is often determined by the internal validity indices. The third approach, based on relative criterion, evaluates a clustering structure by comparing it to other clustering schemes.

In the previous section, because the cluster label of the data was available, we assessed the clustering quality using an external validity index, that is, the Rand index. Here, the internal validity indices were used to determine the optimal number of clusters. Then, the external criteria were used to investigate the clustering quality using in-depth interviews with experts of the oil global trade on the optimal detected community structure. However, the relative criterion approach was examined by comparing the results of the proposed clustering method to other clustering methods with different parameters.

In the first step, the primary computational stages of the proposed algorithm were performed using MATLAB. CVAP was applied to cluster the network in 2 to 10 communities and compute the internal validity indices. Four clusters were proven to be optimum using three internal validity indices: the C, Calinski-Harabasz, and Hartigan indices [43]. Figure 2 presents the indices and optimal number of clusters. Then, with the aim of external validation, we examined the significance of the obtained results in four clusters by consulting experts in the field. Ninety percent of the experts indicated that the quality of the clustering results using the proposed method was satisfactory and within acceptable limits. The list below and Figure 3 present the clustering results of the proposed method.

TABLE 2: Rand external validation indices which compare the clustering result to the known standard partition.

| Dataset | 2 clusters | 3 clusters | 4 clusters | 5 clusters | 6 clusters | 7 clusters | 8 clusters | 9 clusters | 10 clusters | Average |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|----------|
| Cornell | 0.531959 | 0.574993 | 0.591647 | 0.613904 | 0.645519 | 0.64885 | 0.631245 | 0.638118 | 0.652868 | 0.614345 |
| Texas | 0.405497 | 0.648554 | 0.652924 | 0.654534 | 0.649186 | 0.684434 | 0.697085 | 0.690587 | 0.664654 | 0.638606 |
| Washington | 0.676514 | 0.6543 | 0.64925 | 0.716233 | 0.707651 | 0.694817 | 0.706246 | 0.717904 | 0.705259 | 0.692019 |
| Wisconsin | 0.515209 | 0.707033 | 0.747256 | 0.74777 | 0.756118 | 0.752945 | 0.745798 | 0.700114 | 0.723756 | 0.710666 |

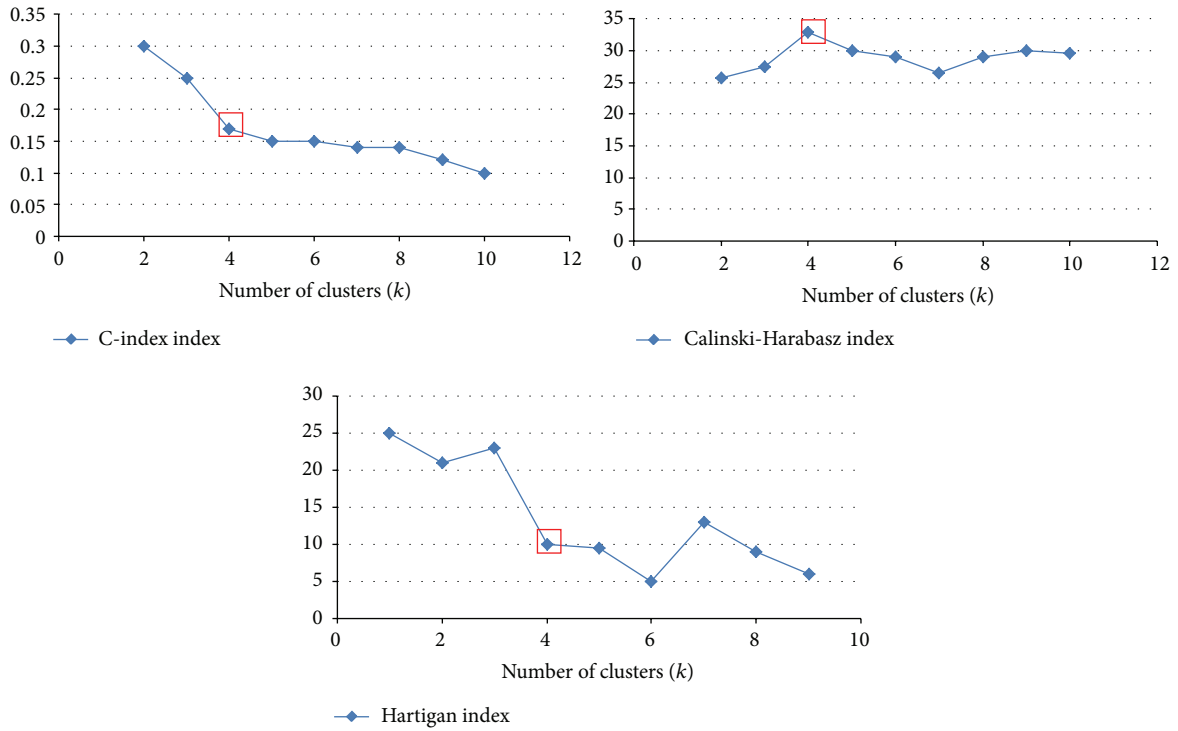


FIGURE 2: Internal validity indices and the optimal number of clusters.

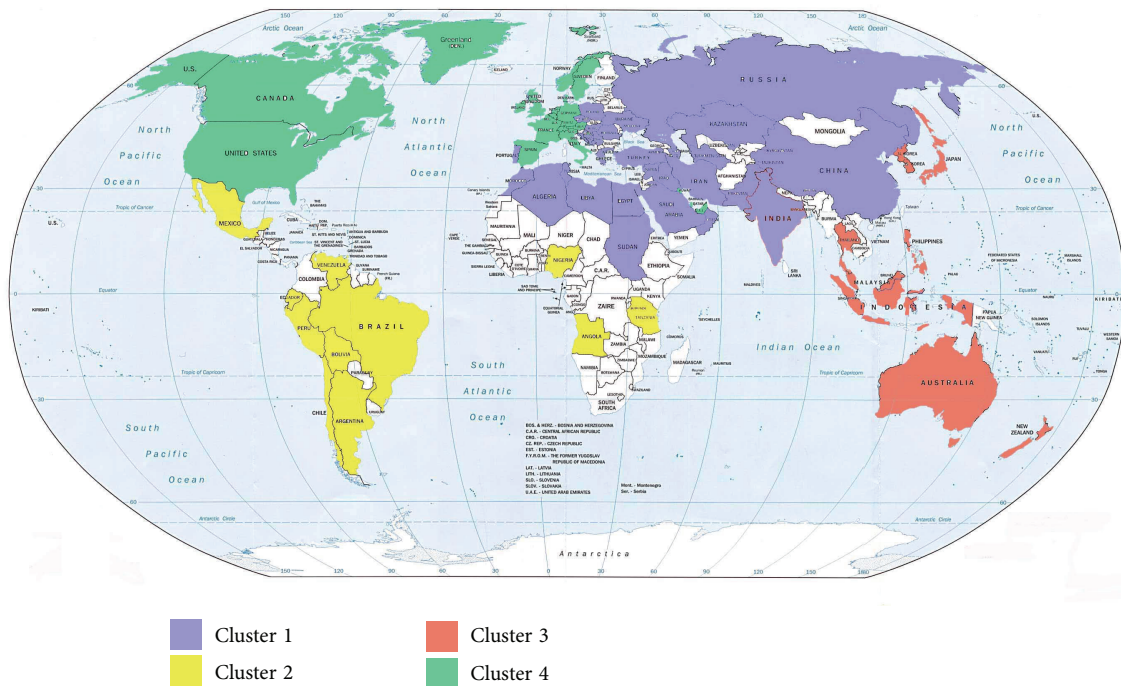


FIGURE 3: Geographical representation of the clustering results of the proposed method.

Cluster 1: Croatia, Greece, Portugal, Poland, Hungary, Czech Rep, Romania, Ukraine, Yugoslavia, Azerbaijan, Kazakhstan, Turkmenistan, Russian Fed, China, India, Pakistan, Turkey, Iran-Islam.R, Iraq, Saudi Arabia, Oman, Syrian A.R., Algeria, Egypt, Libya, Morocco, Tunisia, and Sudan.

Cluster 2: Argentina, Bolivia, Brazil, Chile, Ecuador, Mexico, Peru, Venezuela, Trinidad Tbg, Angola, Nigeria, and Tanzania.

Cluster 3: Japan, Korea Rep., Indonesia, Malaysia, Philippines, Thailand, Australia, and New Zealand.

Cluster 4: Austria, Belgium, Denmark, France + Monaco, Germany, Ireland, Italy, Netherlands, Norway, Spain, Sweden, Switz. Liecht, Untd. Kingdom, USA, Canada, Kuwait, Qatar, and Untd. Arab Em.

Then, for relative validation, the behavior of the proposed clustering method was studied by comparison with the existing clustering methods. The gathered data were clustered once based on the individual characteristics of the nodes, particularly the economic size and geographic coordinates, using a data-mining clustering method, and again based on the characteristics of the edges, that is, bilateral-trade data via a community detection method in graphs; the quality of the aforementioned clustering methods was compared with the proposed method, which simultaneously considers the characteristics of the nodes and edges using the C-index. The results are presented in Table 3, illustrating that the proposed method acquires the minimum and thus the best level of the index.

The cluster formation was based on three key pillars. The first pillar is the energy market environment. The clusters are affected by cluster-specific business environment conditions, which result from the bilateral oil trade of countries. The second pillar is geography. The clusters are driven by proximity and often concentrated in a region within a larger zone, which is often in one continent. The third pillar is the economic size criterion. The clusters may include countries in different regions with similar economic sizes.

In the second step, the closeness centrality was computed according to Section 3.2. The findings indicate that Turkey, Venezuela, Malaysia, and Belgium are the key countries in the first, second, third, and fourth clusters, respectively; that is, they are the closest countries to all other members of each cluster based on the bilateral-trade rate, geographic coordinates, and GDP-per-population criterion.

6. Conclusions

Unlike most social network analysis studies, this study assumes that there are heterogeneous data that are related to a network, namely, the network structure and node attributes. Clustering algorithms traditionally consider only the node attributes, whereas community detection methods mainly focus on the network structure. Therefore, our approach to community detection in network analysis represents a key shift in methodology from the traditional approach because it can overcome the shortcomings of its two components when

TABLE 3: Comparison of the clustering qualities of the proposed method and existing methods.

| Method | Optimal Number of Clusters | C-index |
|--|----------------------------|---------|
| Clustering based on node attributes | 5 | 0.4 |
| Clustering based on structure of relations | 3 | 0.2 |
| Proposed method | 4 | 0.17 |

applied separately. In addition, our proposed method can consider multilayer networks, which may provide a better fit to real-world multilayer complex networks. The hybrid clustering method, which uses the node attributes and topology structure of the network, provided promising results when applied to four benchmark networks and a real-life example.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] Y. Hu and D. Zhu, "Empirical analysis of the worldwide maritime transportation network," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 10, pp. 2061–2071, 2009.
- [2] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality and clustering in complex hyper-networks," *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 581–594, 2006.
- [3] B. A. Mello, L. H. Batistuta, R. Boueri, and D. O. Cajueiro, "Measuring the flow of information among cities using the diffusion power," *Physics Letters, Section A: General, Atomic and Solid State Physics*, vol. 374, no. 2, pp. 126–130, 2009.
- [4] M. G. Everett and S. P. Borgatti, "Extending centrality," in *Models and Methods in Social Network Analysis*, P. J. Carrington, J. Scott, and S. Wasserman, Eds., Cambridge University Press, 2005.
- [5] D. Koschutzki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, "Centrality indices," in *Network Analysis: Methodological Foundations*, U. Brandes and T. Erlebach, Eds., vol. 3418, Springer, Berlin, Germany, 2005.
- [6] M. Guzman, *A probabilistic programming approach in the analysis of social networks [Ph.D. thesis]*, University of Arkansas, 2008.
- [7] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8, Cambridge University Press, Cambridge, UK, 1994.
- [8] E. Estrada, "Community detection based on network communicability," *Chaos*, vol. 21, no. 1, 2011.
- [9] K. Xu, J. He, S. Zou, H. Zhang, T. Yan, and X. Wei, "A cohesive subgraph visualization-based approach to efficiently discover large k-clique community," *Arabian Journal for Science and Engineering*, vol. 37, no. 7, pp. 1959–1968, 2012.

- [10] Y. Q. Niu, B. Q. Hu, W. Zhang, and M. Wang, "Detecting the community structure in complex networks based on quantum mechanics," *Physica A: Statistical Mechanics and Its Applications*, vol. 387, no. 24, pp. 6215–6224, 2008.
- [11] P. Pollner, G. Palla, D. Ábel et al., "Centrality properties of directed module members in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 19–20, pp. 4959–4966, 2008.
- [12] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A: Statistical Mechanics and Its Applications*, vol. 374, no. 1, pp. 483–490, 2007.
- [13] T. Ghosh, S. Sengupta, M. Chattopadhyay, and P. K. Dan, "Metaheuristics in cellular manufacturing: a state-of-the-art review," *International Journal of Industrial Engineering Computations*, vol. 2, no. 1, pp. 87–122, 2011.
- [14] M. E. J. Newman, "Detecting community structure in networks," *European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.
- [15] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.
- [16] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, Article ID P09008, 2005.
- [17] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 1, pp. 187–197, 2010.
- [18] J. Liu and T. Liu, "Detecting community structure in complex networks using simulated annealing with k-means algorithms," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 11, pp. 2300–2309, 2010.
- [19] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466–484, 2006.
- [20] F. Barzinpour and B. H. A. Ahmadi, "A novel framework in complex network analysis: considering both structure of relations and individual characteristics in closeness centrality computation," *International Journal of Industrial Engineering Computations*, vol. 4, no. 2, pp. 227–240, 2013.
- [21] C. Kiss and M. Bichler, "Identification of influencers—measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.
- [22] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, no. 2, pp. 124–136, 2006.
- [23] D. Gómez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada, "Centrality and power in social networks: a game theoretic approach," *Mathematical Social Sciences*, vol. 46, no. 1, pp. 27–54, 2003.
- [24] D. Cai, Z. Shao, X. He, X. Yan, and J. Han, "Community mining from multi-relational networks," in *Knowledge Discovery in Databases: PKDD*, pp. 445–452, Springer, Berlin, Germany, 2005.
- [25] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.
- [26] S. Gómez, A. Díaz-Guilera, J. Gómez-Gardeñes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas, "Diffusion dynamics on multiplex networks," *Physical Review Letters*, vol. 110, Article ID 028701, 2013.
- [27] M. De Domenico, A. Sole, S. Gomez, and A. Arenas, "Random walks on multiplex networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 23, pp. 8351–8356, 2013.
- [28] C. Granell, S. Gómez, and A. Arenas, "Dynamical interplay between awareness and epidemic spreading in multiplex networks," *Physical Review Letters*, vol. 111, no. 12, Article ID 128701, 2013.
- [29] E. Cozzo, R. A. Baños, S. Meloni, and Y. Moreno, "Contact-based social contagion in multiplex networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 88, no. 5, Article ID 050801, 2013.
- [30] L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti, "Eigenvector centrality of nodes in multiplex networks," *Chaos*, vol. 23, no. 3, 2013.
- [31] M. De Domenico, A. Solé-Ribalta, E. Cozzo et al., "Mathematical formulation of multilayer networks," *Physical Review X*, vol. 3, no. 4, Article ID 041022, 2014.
- [32] M. Shiga, I. Takigawa, and H. Mamitsuka, "A spectral clustering approach to optimally combining numerical vectors with a modular network," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*, pp. 647–656, ACM, August 2007.
- [33] F. Janssens, W. Glänzel, and B. de Moor, "A hybrid mapping of information science," *Scientometrics*, vol. 75, no. 3, pp. 607–631, 2008.
- [34] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM '13)*, pp. 1151–1156, December 2013.
- [35] F. Janssens, L. Zhang, B. De Moor, and W. Glänzel, "Hybrid clustering for validation and improvement of subject-classification schemes," *Information Processing & Management*, vol. 45, no. 6, pp. 683–702, 2009.
- [36] M. Arab and M. Afsharchi, "Community detection in social networks using hybrid merging of sub-communities," *Journal of Network and Computer Applications*, vol. 40, no. 1, pp. 73–84, 2014.
- [37] Y. Zhou, H. Cheng, and J. X. Yu, "Clustering large attributed graphs: an efficient incremental approach," in *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM '10)*, pp. 689–698, IEEE, December 2010.
- [38] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pp. 927–936, ACM, July 2009.
- [39] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, "Centrality in interconnected multilayer networks," <http://arxiv.org/abs/1311.2906>.
- [40] <http://www.cs.umd.edu/~sen/lbc-proj/LBC.html#>.
- [41] Q. Lu and L. Getoor, "Link-based classification," in *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, vol. 3, pp. 496–503, August 2003.
- [42] K. Wang, B. Wang, and L. Peng, "CVAP: validation for cluster analyses," *Data Science Journal*, vol. 8, pp. 88–93, 2009.
- [43] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2–3, pp. 107–145, 2001.
- [44] <https://wits.worldbank.org/>.

- [45] R. C. Feenstra, J. R. Markusen, and A. K. Rose, "Using the gravity equation to differentiate among alternative theories of trade," *Canadian Journal of Economics*, vol. 34, no. 2, pp. 430–447, 2001.
- [46] M. Zhou, "Intensification of geo-cultural homophily in global trade: evidence from the gravity model," *Social Science Research*, vol. 40, no. 1, pp. 193–209, 2011.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

