Unsupervised spectral clustering for hierarchical modelling and criticality analysis of complex networks

Yi-Ping Fang, Enrico Zio

To cite this version:


HAL Id: hal-00838315
https://hal-supelec.archives-ouvertes.fr/hal-00838315
Submitted on 25 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Unsupervised spectral clustering for hierarchical modelling and criticality analysis of complex networks

Yi-Ping Fang, Enrico Zio

Abstract

Infrastructure networks are essential to the socioeconomic development of any country. This article applies clustering analysis to extract the inherent structural properties of realistic-size infrastructure networks. Network components with high criticality are identified and a general hierarchical modelling framework is developed for representing the networked system into a scalable hierarchical structure of corresponding fictitious networks. This representation makes a multi-scale criticality analysis possible, beyond the widely used component-level criticality analysis, whose results obtained from zoom-in analysis can support confident decision making.

1. Introduction

Engineered critical infrastructures are a ‘network of independent, large-scale, man-made systems…that function collaboratively and synergistically to produce a continuous flow of essential goods (e.g. energy, data, water…) and services (e.g. banking, healthcare, transportation)” [1] vital to the economy, security and well-being of any country. These systems are exposed to multiple hazards and threats, some of which are even unexpected and emergent, so that a complete analysis by exhaustive treatment cannot be guaranteed. Furthermore, the infrastructure networks consist of a large number of elements whose interactions are not easily modeled and quantified. In practice, then, the performance and reliability assessment of such ‘complex’ systems has proved to be a non-trivial task.

The theory of complex networks has in recent years emerged as a valid tool for describing, modelling and quantifying complex systems in many branches of science [2–5]. Based on the network topology and its treatment by tools of graph theory, various statistical measures have been introduced to evaluate the global structural properties of the network and quantify the importance of the individual elements in the structure of the system [6–8]. While global performance indicators encompass the static characteristics of the whole network, the importance of the different elements in the network can be seen from the point of view of their individual connectivity efficiency and/or their contribution to the propagation of failures through the system network of connections [9–11]. Among these measures, classical and relevant statistics are the network efficiency [12–14], which evaluates the connectivity of the whole network, and the topological centrality measures including degree centrality (CD) [16,17], closeness centrality (CC) [15,17], betweenness centrality (CB) [17] and information centrality (CI) [18,19], which rely on topological information to qualify the importance of individual network elements.

On the other hand, recent studies suggest that many real complex networks exhibit a modularized organization [20]. In many cases, these modularized structures are found to correspond to functional units within networks (ecological niches in food webs, modules in biochemical networks) [21]. Broadly speaking, clusters (also called communities or modules) are found in the network, forming groups of elements that are densely interconnected with each other but only sparsely connected with the rest of the network. The study of the clustered structure of the network of a critical infrastructure is of particular interest because such structure can provide a protection for the system against attacks from an intruder [22], reduce the effects of cascade failures [23] and point at important heterogeneities within the network that may not be registered via network level measures [21]. Finally, hierarchically modularized organization, which is a central idea about the life process in biology, is found to be also an internal structural structure of many technological networks [24], and can be utilized...
Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>set of network nodes</td>
</tr>
<tr>
<td>E</td>
<td>set of network edges</td>
</tr>
<tr>
<td>$G(V, E)$</td>
<td>a network with set of nodes V and edges E</td>
</tr>
<tr>
<td>A</td>
<td>adjacency matrix of network</td>
</tr>
<tr>
<td>$s_{ij}$</td>
<td>similarity measure between node $i$ and $j$</td>
</tr>
<tr>
<td>S</td>
<td>similarity matrix of network</td>
</tr>
<tr>
<td>$L_{sys}$</td>
<td>normalized graph Laplacian matrix</td>
</tr>
<tr>
<td>Q</td>
<td>network modularity index</td>
</tr>
<tr>
<td>SSE</td>
<td>sum of square error</td>
</tr>
<tr>
<td>D</td>
<td>network degree matrix</td>
</tr>
<tr>
<td>$DB$</td>
<td>Davies–Bouldin index of clustering</td>
</tr>
<tr>
<td>Dunn</td>
<td>Dunn index of clustering</td>
</tr>
<tr>
<td>$C_k$</td>
<td>cluster $k$ of network $G(V, E)$</td>
</tr>
<tr>
<td>$N_k$</td>
<td>the central node of cluster $k$</td>
</tr>
<tr>
<td>$n$</td>
<td>number of network nodes</td>
</tr>
<tr>
<td>$m$</td>
<td>number of network edges</td>
</tr>
<tr>
<td>$n_k$</td>
<td>number of nodes in cluster $k$</td>
</tr>
<tr>
<td>$A^{(k)}$</td>
<td>set of fictitious nodes at level $k$</td>
</tr>
<tr>
<td>$E^{(l)}$</td>
<td>set of fictitious edges at level $k$</td>
</tr>
<tr>
<td>$G^{(l)}(A^{(k)}, E^{(l)})$</td>
<td>fictitious network at level $k$ of the hierarchy</td>
</tr>
<tr>
<td>$E(G)$</td>
<td>topology efficiency of network $G$</td>
</tr>
<tr>
<td>$V_{(i)}$</td>
<td>node $i$ of the fictitious network at level $k$ of the hierarchy</td>
</tr>
<tr>
<td>$C_i$</td>
<td>information centrality of node $V_{(i)}$</td>
</tr>
<tr>
<td>$d_{ij}$</td>
<td>shortest path between node $i$ and $j$</td>
</tr>
</tbody>
</table>

...to model these complex systems for their understanding and analysis [25].

The objective of the work presented in this paper is twofold. First, to propose clustering analysis for extracting some inherent structural properties of a network of a critical infrastructure and, second, to adopt a scheme of successive clustering to obtain a hierarchical model made of different varied-size grained virtual networks which can be exploited to perform zoom-in assessments, focusing on the most relevant clusters in the virtual networks at each level of the hierarchy.

The remainder of this paper is organized as follows: Section 2 presents the proposed spectral clustering analysis, taking the structure of the Italian 380 kV power transmission network as an example for illustration; in Section 3, hierarchical modeling of a complex network is first introduced, and then multi-scaled criticality analyses are performed on the hierarchical model; conclusions are drawn in Section 4.

2. Clustering analysis

2.1. Network representation

Graph theory provides a natural framework for the mathematical representation of complex networks. A graph is an ordered pair $G(V, E)$ comprising a set of vertices (nodes) $V = v_1, v_2, \ldots, v_n$ together with a set of edges (also called arcs or links) $E = e_1, e_2, \ldots, e_m$, which are two-element subsets of $V$. The network structure is usually defined by the $n \times n$ adjacency matrix, which defines which two nodes are connected by assigning a 1 to the corresponding element of the matrix; otherwise, the value in the matrix is 0 if there is no connection between the two nodes. As described, this type of graph is unweighted and undirected. A graph is weighted if a value (weight) is assigned to each edge representing properties of the connection like costs, lengths, capacities, etc. For example, the matrix of physical distances is often used in conjunction with the adjacency matrix to describe a network also with respect to its spatial dimension [12,26].

In this paper, we take an exemplification of the analyses proposed on the 380 kV Italian power transmission network (Fig. 1). This network is a branch of the high-voltage-level transmission, which can be modeled as a graph of $n=127$ nodes connected by $m=171$ links [7][27], defined by its $n \times n$ adjacency (connection) matrix $A$ whose entries $a_{ij}$ are 1 if there is an edge joining node $i$ to node $j$ or 0 otherwise. It is important to underline that only the topology of the physical system is taken as reference and used in the analyses, so that the hierarchical model and clustering relate only on the network structure with no specific relation to the electrical properties of the system. The sub-network for Sardinia is not considered to ensure that the network is connected in the sense of a topological space.

2.2. Unsupervised spectral clustering algorithm

Cluster analysis aims at identifying patterns around which communities of elements in the network can be grouped, emerging implicit information in the network structure [28]. Framed as an unsupervised multiple classification problem [29], clustering has been an essential undertaking in the context of explorative data mining and also a common technique for statistical data analysis used in many fields such as machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [30]. Theoretically, based on a similarity (affinity) measure $s_{ij}$ between pairs of data points ($ij$), which is usually a measure of distance between $i$ and $j$, most clustering approaches seek to achieve a minimum or maximum similarity value through an iterative process of vertex grouping [25,28]. Different similarity definitions can lead to different cluster partitioning of the network.

The detailed description of the different clustering methods is beyond the scope of this article. For a systematic and synthetic review, the reader is encouraged to look at [28,30,31]. For the purpose of the analyses presented in this paper, we adopt the unsupervised spectral clustering algorithm (USCA) [32], which is invariant to cluster shapes and densities and simple to implement. The USCA makes use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before Fuzzy $k$-means (FKM)-clustering in fewer dimensions. Schematically, it is performed by the following steps [32]:

Unsupervised spectral clustering algorithm

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$

1. Compute the normalized graph Laplacian matrix $L_{sym}$
2. Compute the first $k$ eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ and corresponding eigenvectors $\pi_1, \pi_2, \ldots, \pi_k$ of matrix $L_{sym}$. The first $k$ eigenvalues are such that they are very small whereas $\lambda_{k+1}$ is relatively large. All eigenvalues are ordered increasing.
3. The number of clusters is set equal to $k$, according to the eigengap heuristic theory [32].
4. Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $\pi_1, \pi_2, \ldots, \pi_k$ as columns. Form the matrix $T \in \mathbb{R}^{n \times n}$ from $U$ by normalizing the rows to norm 1, that is set $t_{ij} = u_{ij} / (\sum_{k=1}^{n} u_{ik}^2)^{1/2}$.
5. For $i=1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the $i$th row of $T$. 


6. Resort to the FKM algorithm \cite{33,34} to partition the data points $\{y_i\}_{i=1}^n$ into $k$ clusters $A_1, \ldots, A_k$.

Output: Clusters $C_1, \ldots, C_k$ with $C_i = \{y_j \in A_i\}$

In the first step, the Laplacian matrix $L_{\text{sym}}$ is calculated from the similarity (affinity) matrix as follows. The input similarity matrix $S$ is of size $n \times n$ and its generic element $s_{ij}$ represents the similarity between nodes $i$ and $j$ in the network. The diagonal components $s_{ii}$ are set to 1 and the matrix is symmetric ($s_{ij} = s_{ji}$). The degree matrix $D$ is the diagonal matrix with diagonal entries $d_1, d_2, \ldots, d_n$, defined by

$$d_i = \sum_{j=1}^{N} s_{ij} \text{ for } i = 1, 2, \ldots, n.$$  

(1)

Then, the normalized graph Laplacian matrix can be obtained:

$$L_{\text{sym}} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}SD^{-1/2}$$  

(2)

where $L = D-S$ and $I$ is the identity matrix of size $n \times n$.

It should be noted that the eigengap heuristic theory at the basis of the third step of the algorithm works well when the modularized structure of the data are pronounced whereas the more noisy or overlapping the clusters are, the less effective it is
In those cases, other methods such as the Markov Clustering Algorithm [35] can be used to find the optimal number of clusters.

2.3. Clustering results and analysis

2.3.1. Affinity construction

As mentioned in the previous section, the result of clustering is sensitive to the similarity function which defines the proximity of the nodes in the network. Since network clustering is to group the vertices of the network into clusters taking into consideration the edge structure of the graph in such a way that there should be many links within each cluster and relatively few between the clusters, using topological information itself is intuitively appropriate to estimate the structure of clusters, using topological information itself is intuitively appropriate to estimate the structure of node pairs. In this view, two node affinity definitions representative of the local and global topological properties of the network structure are introduced in this paper to support the successive cluster-level criticality analysis.

Possibly, the most straightforward manner to quantify the affinity between a pair of nodes in a network is to use only the local adjacency information: nodes $i$ and $j$ are seen as similar if they are linked directly, otherwise they are not. The consequent adjacency affinity matrix $S_1$ is identical to the adjacency matrix $A$ of the network.

The adjacency affinity uses only local direct connection information and possibly fails to detect any other structure when a network is not locally dense [24]. Since in this study, we use clustering to decompose the network into topologically dense community structures, for nodes to belong to the same cluster, they should be highly connected to each other, i.e. not necessarily by a direct link but by a short path [36]. For this reason, we introduce the topological distance affinity to drive the clustering. The topological distance (shortest path) $d_{ij}$ between nodes $i$ and $j$ is the minimum number of edges traversed to get from vertex $i$ to vertex $j$. The matrix $D$ of the topological distances can be extracted from the adjacency matrix $A$. Thereafter, the topological distance affinity can then be defined based on the elements $d_{ij}$ of $D$ and the Gaussian similarity function:

$$S_2(i,j) = \exp(-d_{ij}^2/(2\sigma^2)) \quad i,j = 1,2,\ldots,n$$

where $\sigma$ is a tuning parameter. This parameter can be tuned to scale the Gaussian similarity function, similarly to the parameter $\epsilon$ in the $\epsilon$-neighborhood graph [32]. Unfortunately, there are no theoretical results to guide the choice of the parameter, and only some rules of thumb have been suggested in the literature [32]. In our study, we choose a value of 0.8 for $\sigma$, which is of the order of the mean distance of a node to its $k$th nearest neighbor, where $k$ is chosen as $k = \log(n) + 1$.

2.3.2. Cluster evaluation

The assessment of the quality of the clustering results is a non-trivial task because of the unsupervised nature of the analysis. The clustering structure itself and the relational characteristics of the dataset are often utilized as the measurement information for clustering evaluation [25]. In our study, the evaluation of the clustering is based on four representative indices capturing complementary characteristics of the clusters found: the modularity index ($Q$) as an indicator of the presence of a modularized structure; the Sum of Squared Error (SSE) to quantify the cohesion of clusters; the Davies–Bouldin index (DB) and Dunn index (Dunn) to evaluate high intra-cluster similarity and low inter-cluster similarity, with different metrics.

2.3.2.1. Modularity index. The modularity index $Q$, introduced by Newman and Girvan [37], attempts to measure how well a given partition of a network compartmentalizes its communities and is defined as [38]:

$$Q = \sum_{i=1}^{k} \left( \frac{e_i}{m} - \frac{\phi_i}{2m^2} \right)$$

where $k$ is the number of clusters, $e_i$ defines the number of links in cluster $i$, $\phi_i$ is the sum of the degrees of the nodes in cluster $i$, and $m$ represents the total number of links in the whole network. Note that when $Q$ = 0, all the nodes are in one single community while $Q$ > 0 indicates the existence of some kind of inherent cluster structure. Modularity measures the difference between the total fraction of edges that fall within clusters versus the fraction one would expect if edges were placed at random. Thus, high values of $Q$ represent network partitions in which more of the edges fall within clusters than expected by chance [39]. Moreover, Newman and Girvan [37] suggest that values of $Q$ in the range of 0.2–0.7 designate the presence of cluster structures.

2.3.2.2. Sum of squared error (SSE). Sum of squared error (SSE) measures the cohesion of clusters without respect to external information, i.e. quantifies how closely related are the elements in a cluster. SSE is suitable for comparing two clustering partitions or two clusters [40]. Given two different sets of clusters resulting in $m_1$ and $m_2$ clusters, the SSE is defined as:

$$SSE = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (m_i - c_{ij})^2$$

where $m_1$ and $m_2$ are the sizes of clusters $i$ and $j$, respectively, and $c_{ij}$ is the number of edges that connect nodes in cluster $i$ to nodes in cluster $j$.

Fig. 2 gives out the value landscape of both adjacency affinity matrix $S_1$ (left) and topological distance affinity matrix $S_2$ (right) for the 380 kV Italian power transmission network. One can notice the difference in value scale: the adjacency affinity is a sparse matrix with only values 0 and 1, whereas the topological distance affinity measure shows that nodes in local neighborhoods have relatively high similarity value while affinity values between far-away nodes are weak, although not necessarily negligible.

![Fig. 2. Adjacency affinity and topological distance affinity matrices for the 380 kV Italian power transmission network.](image-url)
from two different clustering procedures, the one with smaller \(\text{SSE}\) is preferable since this means that the prototypes (centroids) of this clustering are superior representations of the points in the clusters. \(\text{SSE}\) is formally defined as follows:

\[
\text{SSE} = \sum_{i=1}^{k} \sum_{j \neq i} \text{dist}(c_i, c_j)^2
\]  

where \(\text{dist}\) represents the topological distance (shortest path) between node \(j\) and the central node \(c_i\) of the cluster \(A_i\) which node \(j\) belongs to.

2.3.2.4. Dunn index [42]. The Davies–Bouldin (DB) index introduced in [41] is formulated as follows:

\[
\text{DB} = (1/k) \sum_{i=1}^{k} \left( \max_{j \neq i} \left\{ \frac{S_i + S_j}{d(c_i, c_j)} \right\} \right)
\]

2.3.2.5. Bounding (DB) index [41]. The Bounding–Bouldin (DB) index introduced in [41] is formulated as follows:

\[
\text{DB} = (1/k) \sum_{i=1}^{k} \left( \max_{j \neq i} \left\{ \frac{S_i + S_j}{d(c_i, c_j)} \right\} \right)
\]

where \(S_i\) is the scatter within the \(i\)th cluster, i.e. the average distance of all elements in cluster \(i\) to its centroid \(c_i\), and \(d(c_i, c_j)\) is the distance between clusters \(i\) and \(j\). A clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion.

2.3.2.6. Dunn index [42]. The Dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance:

\[
\text{Dunn} = \min_{1 \leq i \leq k} \left\{ \min_{1 \leq j \leq k, j \neq i} \left\{ \frac{\text{dist}(C_i, C_j)}{\max_{1 \leq s \leq k} \text{dist}(C_s, C_t)} \right\} \right\}
\]

where \(k\) is the number of clusters, the function \(\text{dist}\) gives the distance between two clusters \(C_i\) and \(C_j\) (the shortest path between two centroids) and \(\Delta\) represents the diameter of a cluster \(C_p\) (the maximum shortest path between any node pairs within the cluster). Since internal criterions seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

2.3.3. Clustering analysis of the 380 kV Italian power transmission network

We applied the USCA for performing the clustering analysis of the 380 kV Italian power transmission network. Both adjacency affinity and topological distance affinity were considered. The resulting partitions are showed in Fig. 3(a) and (b), respectively. Different shapes represent different clusters. The filled nodes locate the clusters centers, which are the physical node nearest to the centroids of the clusters based on the Euclidean distance measure. The two different affinity definitions produce somewhat similar partitions in four clusters, though some differences exist. The clusters in both cases exhibit not only physical proximity but also intensity of the relationship in terms of the network connectivity, which results from the fact that generally only nodes with geographical closeness are connected in the power transmission network.

Table 1 represents the comparison results of the two partitions. The \(Q\) values for adjacency affinity and topological distance affinity are both within the range of \([0.2, 0.7]\), which designates the existence of a modularized structure within the 380 kV Italian power transmission network. Partitioning into four clusters is confirmed for both affinities. The size and central node for cluster 1 (whose elements are represented as squares in Fig. 3) are identical and cluster 2 (circles) has same centroid but different size, whereas cluster 3 (triangles) and 4 (diamonds) have neither the same size nor identical central nodes. This discrepancy is probably due to the fact that the nodes in the north part of the Italian transmission network (composed by clusters 1 and 2) are densely connected and their modularized structure is more prominent compared with the south part (composed by clusters 3 and 4), thus both local and global topological affinities can achieve the overall maximum of the modularity. Actually, the \(Q\) values of the north part of the network (composed by cluster 1 and 2), i.e. 0.443 for adjacency affinity and 0.444 for topological distance affinity, are both higher than those of the south part (composed by clusters 3 and 4), i.e. 0.314 and 0.119 for adjacency affinity and topological distance affinity, respectively.

![Fig. 3. Clustering results for the adjacency affinity and the topological distance affinity on the 380 kV Italian power transmission network.](image)
In addition, the partitions obtained exhibit $DB=0.883$, $SSE=1585$ for adjacency affinity, and $DB=0.99$, $SSE=1867$ for topological distance affinity. In both evaluation indexes $DB$ and $SSE$, clustering by adjacency affinity outperforms that by topological distance affinity. Furthermore, the clusters from adjacency affinity are relatively more balanced in size. For the above reasons, the adjacency affinity is retained for the analyses of the following sections.

2.4. Component importance by clustering

A previous study [11] defined the community-level vulnerability based on the reciprocal of the number of inter-cluster links, thus showing that the modularized structure could be leveraged to the criticality analysis of network elements. In this study, two types of elements in the clustering are paid special attention to (Fig. 4). First, the elements (links and vertices) which are in the periphery and connect different clusters (hereafter called inter-cluster links and cluster-border nodes, respectively) intuitively play a critical role in the complex interaction and communication occurring between different modules of the whole network. In this sense, the so-called overlapping nodes [43,44] are similar to our cluster-border nodes. Second, the central nodes within each cluster, which own highest membership to the cluster, are expected to have a dense pattern of local connections and their failures could possibly propagate to a severe damage to the network.

Fig. 4 represents the inter-cluster links (black lines), cluster-border nodes (nodes with ‘+’ symbol inside) and the central nodes (nodes filled with black color) obtained from the adjacency affinity clustering of the 380 kV Italian power transmission network. The inter-cluster links set $E$ is $\{(30–31), (30–34), (59–60), (61–62), (64–78), (71–83), (76–79), (107–109), (110–111), (112–114)\}$. Coincidently, the three lines identified as the most critical triplet of lines in [45], because their removal would result in a huge efficiency drop for the whole network, are among the inter-cluster links set $E\prime$: $\{ (64–78), (71–83), (76–79) \}$. This shows the importance of these types of elements for the structured robustness of a network, and the usefulness of clustering analysis for their identification.

Table 2 reports the membership values of these cluster-border nodes and cluster central nodes (bold), and their rank positions according to the information, degree, closeness and betweenness centrality measures for cluster-border and central nodes (bold) of each cluster; only the 24 top-ranked are reported.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Critical node</th>
<th>MV</th>
<th>Rank $C^I$</th>
<th>Rank $C^D$</th>
<th>Rank $C^C$</th>
<th>Rank $C^B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>30</td>
<td>0.7296</td>
<td>13</td>
<td>4</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>59</td>
<td>0.7768</td>
<td>13</td>
<td>4</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>0.7606</td>
<td>20</td>
<td>9</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>0.5527</td>
<td>15</td>
<td>11</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1.0000</td>
<td></td>
<td>24</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.7373</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.7948</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>0.8699</td>
<td>4</td>
<td>15</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>0.8114</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>0.8394</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>0.9054</td>
<td>22</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>0.9998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>0.4772</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>0.9198</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>0.4775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>0.7442</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>0.8203</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>0.5442</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>0.9993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>0.9466</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>0.5724</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>0.7314</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. Inter-cluster links, cluster-border nodes, and central nodes for the 380 kV Italian power transmission network.
centrality measures based on the results in [7]. Detailed definition and explanation of these four centrality measures can be found in the literature [7,15–19]. One can see that most of the nodes found important by clustering, because cluster-border or central, are ranked among the top 24 with highest centrality values, although specific exceptions exist such as the nodes 23, 30, 31, 34, 112 in clusters 1, 2, 3 and the nodes 109, 111 and 114 in cluster 4. This difference is due to the fact that the “clustering-important” nodes are identified based only on regional topological information and not on any other consideration on the role in the whole network.

3. Hierarchical modelling and zoom-in assessment of the network

3.1. Hierarchical model of the network

If one looks closely at the individual clusters in Fig. 3, it may notice that some of them exhibit a modularized structure, and hence can be decomposed further into sub-clusters. Indeed, many real networks reveal a hierarchical organization, where vertices divide into groups that further subdivide into groups of groups, and so forth over multiple scales [4]. On this basis, a framework for hierarchical system modelling has recently been proposed in [25] aiming at reducing the computational burden of modelling the entire system.

For illustration of the potential of the hierarchical modelling framework for complex system analysis, by analogy one may think of the electronic maps such as those provided by Google Maps; the tools are powerful because they present information in a scalable manner—despite the decrease in the amount of information as we “zoom in”, the representation shows the information that is relevant at the new scale.

In the same spirit, a hierarchical model representing the whole system at the top and individual elements at the bottom could be obtained via successively performing unsupervised spectral clustering algorithm on the network. Then, based on the hierarchical network representation, fictitious networks can be defined in each level, from which the analyst can extract relevant information at the suitable level of the hierarchy. Fictitious networks are clusters—simplified representations of the real network and can facilitate the understanding and analysis of the network properties by focusing on the relevant information that emerges at the different levels.

Following a similar formulation as in [46], the fictitious network at level \( k \) is denoted by a graph \( G^{(k)} \), \( A^{(k)}, E^{(k)} \). Let us denote as \( V^{(k)}(i=1,...,n^{(k)}) \) the node \( i \) of the fictitious network at level \( k \) of the hierarchy and associate a weight to it which is equal to the number of actual nodes which compose \( V^{(k)} \). These fictitious nodes are connected by \( m^{(k)} \) fictitious edges \( E^{(k)} = E^{(k)}_1, E^{(k)}_2, ..., E^{(k)}_{m^{(k)}} \). Considering parallel connections, \( E^{(k)}_i \) is weighted by the reciprocal of the number of actual edges it contains. Then, the fictitious network is represented by a weighted adjacency matrix \( A^{(k)} \) whose element \( A^{(k)}(i,j) \) for the fictitious nodes \( V^{(k)}_i \) and \( V^{(k)}_j \) are connected by fictitious edge \( E^{(k)}_p \) and 0 otherwise. This definition accounts for the fact that a fictitious edge embracing several real links has that number of paths available between the two communities it connects, thus holding more interaction efficiency and smaller weight viewed as the physical distance between the two communities connected by the virtual edge.

Fig. 5 gives an example of the construction of a fictitious network.

The 380 kV Italian power transmission network has been modeled as a five levels hierarchy (to which correspond five fictitious networks) by successively applying USCA. In Fig. 6, the weighted fictitious networks and their corresponding weighted adjacency matrices at the levels 2 and 3 of the hierarchy are presented for illustration. The number beside the fictitious node \( V^{(k)}_i \) represents its weight (number of actual nodes included in the virtual node); for example, the weight of \( V^{(3)}_5 \) is 36. The fictitious network at level 1 is a single fictitious node whose size is 127, the total number of nodes in the network, whereas at the last level 5 the fictitious network corresponds to the actual physical network.

3.2. Centrality analysis on fictitious networks

Based on the hierarchical representation of the network, problems such as reliability assessment and damage propagation [25] can be swiftly unraveled with low complexity at the expense of low specificity. In this section, we carry out centrality analysis on the fictitious networks, focusing step-wise on the most critical clusters (fictitious nodes) at each scale of the hierarchy. This is valuable for decision makers when they want to allot limited investments to a regional part of the network, which is usually operated by local organizations, to improve the vulnerability of the overall network system.

3.2.1. Efficiency modelling

Network topological efficiency introduced in [46] allows a quantitative analysis of the information flow, and works both in the unweighted abstraction and in the more realistic assumption of weighted networks. This measure is based on the assumption that the information (communication) in a network travels along the shortest routes, and that the efficiency in the communication between two nodes \( i \) and \( j \) is inversely proportional to their shortest path length \( d_{ij} \) which is defined as the smallest sum of the physical distances throughout all the possible paths in the
weighted network. Then, the efficiency of the whole network is given by:

\[ E(G) = \frac{1}{\sum_{i \neq j \in G} d_{ij}} \sum_{i \neq j \in G} \frac{1}{d_{ij}}. \]  \tag{8}

This formula produces a value of \( E \) that can vary in the range of \([0, \infty) \). \( E(G) \) is defined as 1 in the case of \( n = 1 \), i.e., there is only one single node in the network. It is more practical to have \( E \) normalized to be in \([0, 1]\). For this reason, we consider the ideal case \( G_{\text{ideal}} \) in which the network has all the \( n(n-1) \) possible links among its nodes. In such a case, the information is propagated in the most efficient way since \( d_{ij} \) equals the physical distance between nodes \( i \) and \( j \) and \( E \) assumes its maximum value. The efficiency \( E(G) \) considered in the following of the paper is always divided by \( E(G_{\text{ideal}}) \) and therefore \( 0 \leq E(G) \leq 1 \).

Notice that, for our analysis of fictitious networks modelling of the Italian power transmission network, the physical distance exists even if there is no fictitious edge between two nodes \( V_p \) and \( V_q \). For generality, their physical distance is defined as the reciprocal of the minimum size of the two fictitious nodes if there is not fictitious edge connecting them. By this definition, the physical distance of nodes in the bottom level fictitious network, i.e., the actual network, coincides with that obtained by considering it as an unweighted network.

Fig. 7 plots the efficiency values of the fictitious networks at each level of the hierarchy. It can be observed that as the evaluation moves down in the hierarchy, the efficiency difference between the fictitious network and the actual network decreases as expected. Note that the minimum efficiency at level 3 stems from the fact that the ideal fictitious networks \( G_{\text{ideal}} \) have different topologies and link weights at different levels of the hierarchy.

Thus, it is not necessary that the curve of network efficiency decreases monotonically. Fig. 7 is used to qualitatively show that as the evaluation moves down in the hierarchy, the efficiency approximation gets closer to the efficiency of the actual network.

### 3.2.2. Zoom-in criticality analysis

The hierarchical model makes a multi-scale criticality analysis possible, beyond the widely studied component-level criticality analysis. This zoom-in criticality analysis is analogous to the procedure of locating a specific site in a scalable electronic map.
where $G$ network built by clustering in Section 3.1 is presented in Fig. 8. By CI as the information centrality of its corresponding network topological efficiency caused by the removal of all the fictitious edges incident in $V_i^{(k)}$: 

$$C_{i}^{C} = \frac{\Delta E(V_i^{(k)})}{E} = \frac{E[C_i^{(k)}]-E[C_i^{(i)}]}{E[C_i^{(i)}]} \quad (9)$$

where $C_i^{(k)}$ is the network obtained by removing from the original fictitious network the fictitious edges incident in node $V_i^{(k)}$.

An illustration of the process of zoom-in criticality analysis on the 5-level hierarchical model of the 380 kV Italian power transmission network built by clustering in Section 3.1 is presented in Fig. 8. By first ‘opening’ the single unit at level 1, a weighted fictitious network with 4 nodes at level 2 is achieved, in which the information centrality of each fictitious node is calculated according to Eq. (9) and is presented in the corresponding Table. It shows that node $V_i^{(2)}$ owns the highest $C_i^C$ value; then, the internal topology of $V_i^{(2)}$ at level 3 is unraveled by zooming into $V_i^{(3)}$. Similarly, the most critical clusters at levels 3 and 4 can be determined as $V_i^{(3)}$ and $V_i^{(4)}$, which include 11 and 4 actual nodes, respectively. In level 5, which represents the real network, however, the four nodes have the same values of information centrality since they are completely connected and the removal of all the edges incident in any one of the four nodes would result in the equal relative drop in the network topological efficiency.

Note that the difference of cluster-level information centrality is quite pronounced for the 380 kV Italian power transmission network, compared to the node-level information centrality reported in [7] where the difference between the biggest and smallest $C_i^C$ values is only 0.0194; then, the analyst may have more confidence to make clear-cut, relevant decisions based on the cluster-level criticality results of the 380 kV Italian power transmission network.

4. Conclusions

In this article, the feasibility of extracting cluster-level structural properties for a realistic-size network by clustering analysis...
has been first investigated, taking as reference example the 380 kV Italian power transmission network structure. Then, the hierarchical modelling framework has been utilized to represent the networked system, forming a scalable hierarchical structure of corresponding fictitious networks. In the context of the hierarchical representation of the network, zoom-in criticality analysis has been proposed to identify the most relevant clusters at the desired level of the hierarchy.

For clustering analysis, both adjacency affinity and topological affinity have been considered when applying USCA on the 380 kV Italian power transmission network structure, and their results have been compared to those of four classic centrality measures. For the considered network, the adjacency affinity has turned out to give superior partition. Also, the inter-cluster links, cluster-border nodes and central nodes of each cluster, have been identified as critical: most of the nodes found important by clustering, because cluster-border or central, have turned out to be ranked among the top 24 with highest centrality values (CI, CD, CC and CB) and the most critical triplet of lines identified in [45] is contained within the inter-cluster links set. This confirms the importance of these types of elements for the structural robustness of a network and the usefulness of clustering analysis for their identification.

Then, the systemic hierarchical representation has been introduced for modelling and analysis of complex network systems, with the objective of rendering more manageable the treatment of real-world critical infrastructures. A five-level hierarchical model of the 380 kV Italian power transmission network structure has been obtained by successively applying USCA. The cluster-level information centrality has been proposed and used as an illustration to quantify the importance criticality of a cluster in the network. The most critical clusters at each level of the hierarchy have been identified with high confidence for decision making.

Finally, a comment is in order with respect to the computational complexity of the approach proposed. The complexity depends primarily on the computational cost of spectral clustering, where a large number of eigenvectors have to be computed for large graph Laplace matrices (step 2 of the algorithm), whose time complexity of computing eigenvectors is \( \Theta(n^3) \) [48]. Thus, the computation cost of constructing the hierarchical model is \( \Theta(n^3 l) \), where \( l \) is the number of hierarchical levels. In general, the high-quality clustering of the spectral method is at the expense of its comparatively demanding computation cost. In this study, the spectral clustering is adopted as one possible way to extract some inherent cluster-level structural properties and derive the hierarchical modelling which sets the base for a multi-scale criticality analysis, which is our main objective. Furthermore, as many real adjacency matrices are sparse in nature, efficient existing methods to compute the eigenvectors of sparse matrices need to be adopted [49]. Finally, some improvements of spectral clustering have been proposed in Statistics and Data Mining such as parallel spectral clustering [50], distributed method [51] and fast approximation [52] to make it scalable to large network problems.

Acknowledgements

The authors are thankful to Dr. Giovanni Sansavini of Politecnico di Milano for supporting the work with relevant information on the network structure uses as reference system and to Dr. Franck Marle for fruitful discussions.

References


