Reliability Engineering and System Safety 116 (2013) 64–74

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

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

Yi-Ping Fang [a],*, Enrico Zio [a,b]

[a] Chair on Systems Science and the Energetic Challenge, Ecole Centrale Paris and Supelec, Paris 92295, France
[b] Energy Department, Politecnico di Milano, Milano, Italy

ARTICLE INFO

Article history:
Received 23 August 2012
Received in revised form
24 January 2013
Accepted 21 February 2013
Available online 16 March 2013

Keywords:
Critical infrastructures
Complex networks
Criticality analysis
Centrality measures
Spectral clustering
Hierarchical modelling

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 structure of many technological networks [24], and can be utilized

* Corresponding author. Tel.: +33 65224 0019.
E-mail address: yiping.fang@ecp.fr (Y.-P. Fang).

0951-8320/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ress.2013.02.021

Nomenclature

V	set of network nodes
E	set of network edges
$G(V, E)$	a network with set of nodes V and edges E
A	adjacency matrix of network
s_{ij}	similarity measure between node i and j
S	similarity matrix of network
L_{sys}	normalized graph Laplacian matrix
Q	network modularity index
SSE	sum of square error
D	network degree matrix
DB	Davies–Bouldin index of clustering
$Dunn$	Dunn index of clustering
C_k	cluster k of network $G(V, E)$
N_k	the central node of cluster k
n	number of network nodes
m	number of network edges
n_k	number of nodes in cluster k
$\Lambda^{(k)}$	set of fictitious nodes at level k
$E^{(k)}$	set of fictitious edges at level k
$G^{(k)}(\Lambda^{(k)},E^{(k)})$	fictitious network at level k of the hierarchy
$E(G)$	topology efficiency of network G
$V_i^{(k)}$	node i of the fictitious network at level k of the hierarchy
$C^I_{V_i^{(k)}}$	information centrality of node $V_i^{(k)}$
d_{ij}	shortest path between node i and j

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 modelling 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 (i,j), 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 $\overline{u}_1,\overline{u}_2,\ldots,\overline{u}_k$of matrix L_{sym}. The first k eigenvalues are such that they are very small whereas λ_{k+1} is relatively large. All eigenvalues are ordered increasingly.
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 $\overline{u}_1,\overline{u}_2,\ldots,\overline{u}_k$ as columns. Form the matrix $T\in\mathbb{R}^{n\times k}$ from U by normalizing the rows to norm 1, that is set $t_{ij}=u_{ij}/(\sum_k u_{ik}^2)^{1/2}$.
5. For $i=1$, …, n, let $y_i\in\mathbb{R}^k$be the vector corresponding to the ith row of T.

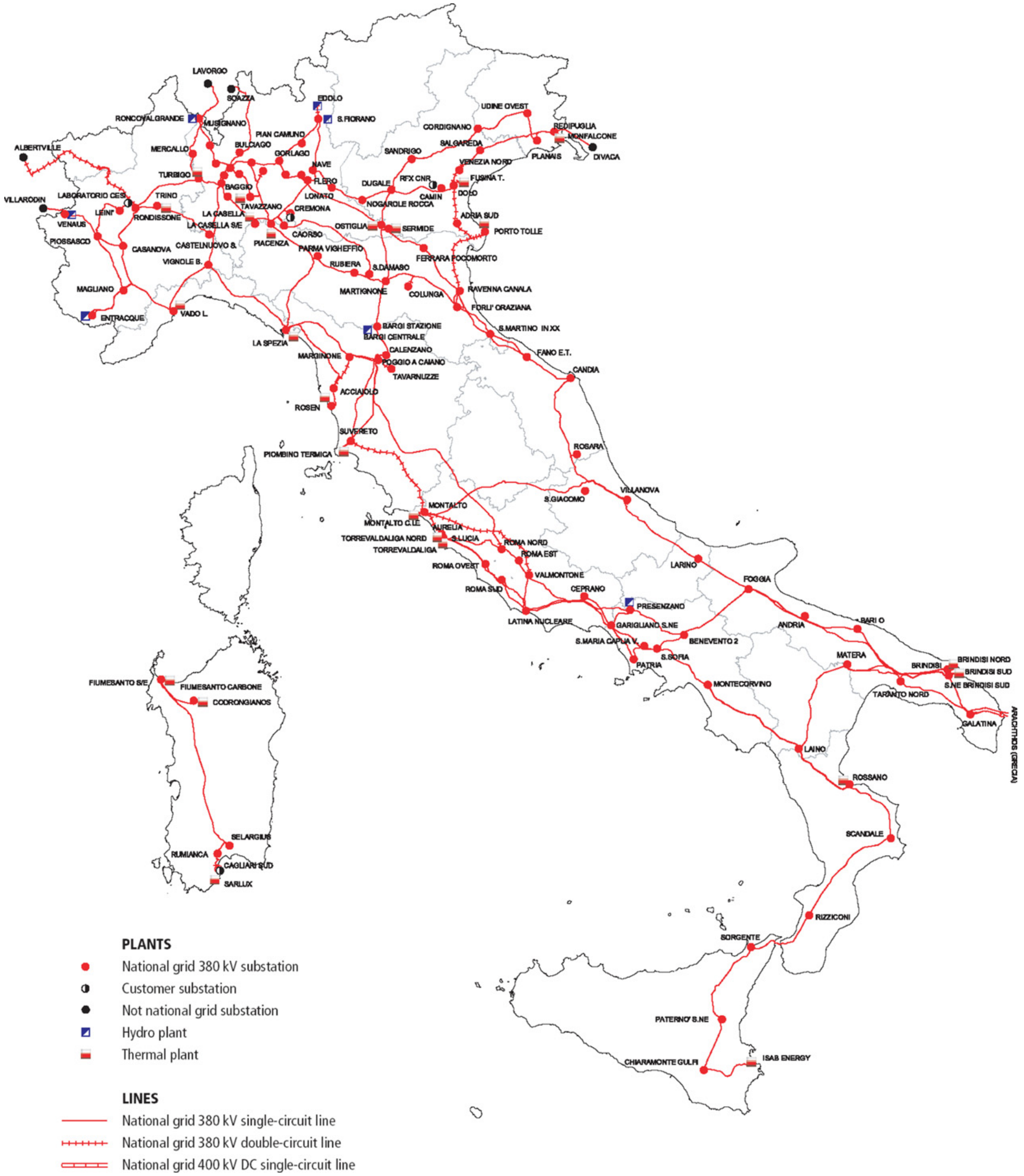


Fig. 1. The 380 kV Italian power transmission network.

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

Output: Clusters $C_1,\ldots,C_k$with $C_i=j|y_j\in A_i$

In the first step, the Laplacian matrix L_{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,…, d_n defined by

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

Then, the normalized graph Laplacian matrix can be obtained:

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

where $L=D\text{–}S$ 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

[32]. 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 intuitionally appropriate to estimate the structure affinity 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:

$$S_2(i,j) = exp(-{d_{ij}}^2/(2\sigma^2)) \quad i,j = 1,2,\ldots,n \tag{3}$$

where σ is a tuning parameter. This parameter can be tuned to scale the Gaussian similarity function, similarly to the parameter ε in the ε-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 σ, which is of the order of the mean distance of a node to its kth nearest neighbor, where k is chosen as $k \sim \log(n)+1$.

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.

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]:

$$Q = \sum_{i=1}^{k} \left(\frac{e_i}{m} - \left(\frac{\varphi_i}{2m} \right)^2 \right) \tag{4}$$

where k is the number of clusters, e_i defines the number of links in cluster i, φ_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

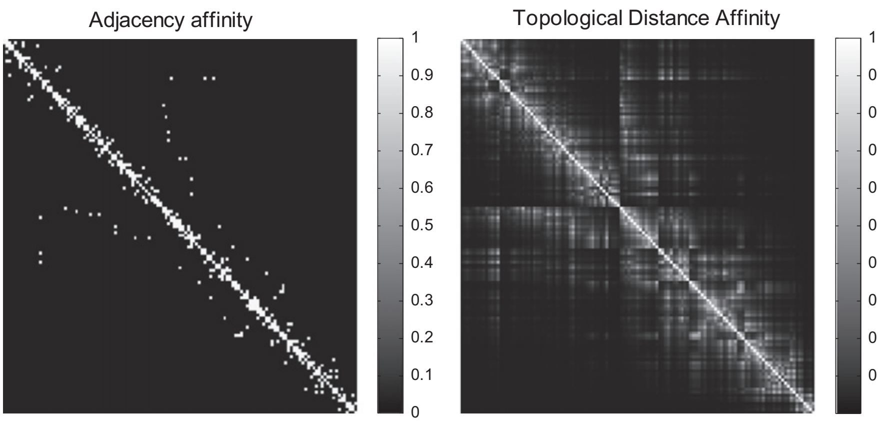


Fig. 2. Adjacency affinity and topological distance affinity matrices for the 380 kV Italian power transmission network.

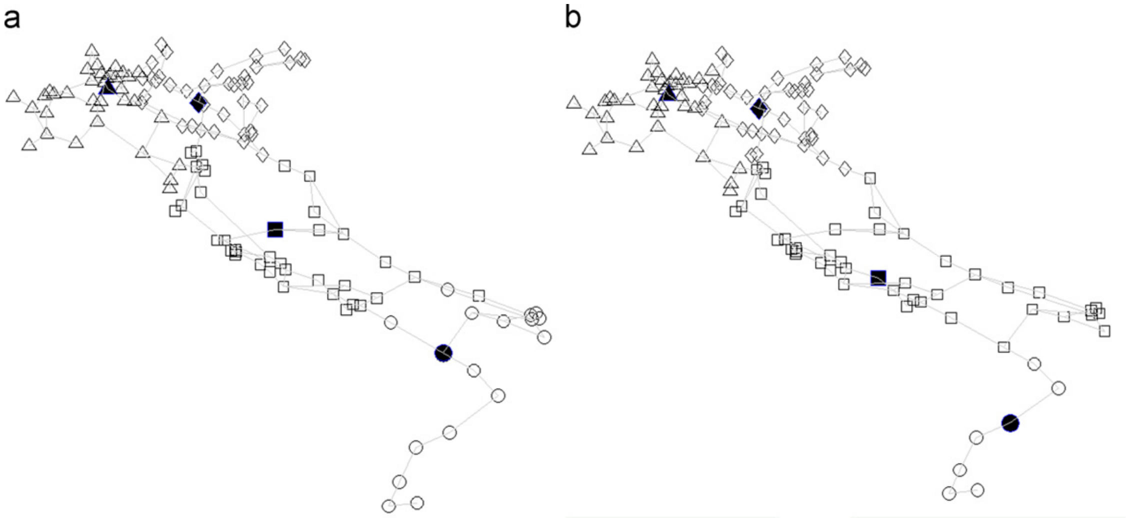


Fig. 3. Clustering results for the adjacency affinity and the topological distance affinity on the 380 kV Italian power transmission network.

from two different clustering procedures, the one with smaller *SSE* is preferable since this means that the prototypes (centroids) of this clustering are superior representations of the points in the clusters. *SSE* is formally defined as follows:

$$SSE=\sum_{i=1}^{k}\sum_{j\in A_i} dist(c_i,j)^2 \quad (5)$$

where *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.

2.3.2.3. Davies–Bouldin (DB) index [41]. The Davies–Bouldin (*DB*) index introduced in [41] is formulated as follows:

$$DB=(1/k)\sum_{i=1}^{k}\left(\max_{i\neq j}\left\{\frac{S_i+S_j}{d(c_i,c_j)}\right\}\right) \quad (6)$$

where S_i is the scatter within the ith 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.

2.3.2.4. 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:

$$Dunn=\min_{1\leq i\leq k}\left\{\min_{1\leq j\leq k, j\neq i}\left\{\frac{\delta(C_i,C_j)}{\max_{1\leq p\leq k}\Delta(C_p)}\right\}\right\} \quad (7)$$

where k is the number of clusters, the function δ gives the distance between two clusters C_i and C_j (the shortest path between two centroids) and Δ 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
Comparison of the clustering results for adjacency affinity and topological distance affinity.

Comparison items	Adjacency affinity	Topological distance affinity
Q	0.664	0.640
Number of cluster	4	4
Cluster central nodes (N_1, N_2, N_3, N_4)	23, 40, 86, 119	23, 40, 99, 121
Cluster size (n_1, n_2, n_3, n_4)	36, 38, 36, 17	36, 41, 43, 7
DB	0.883	0.987
Dunn	0.455	0.455
SSE	1585	1867

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.

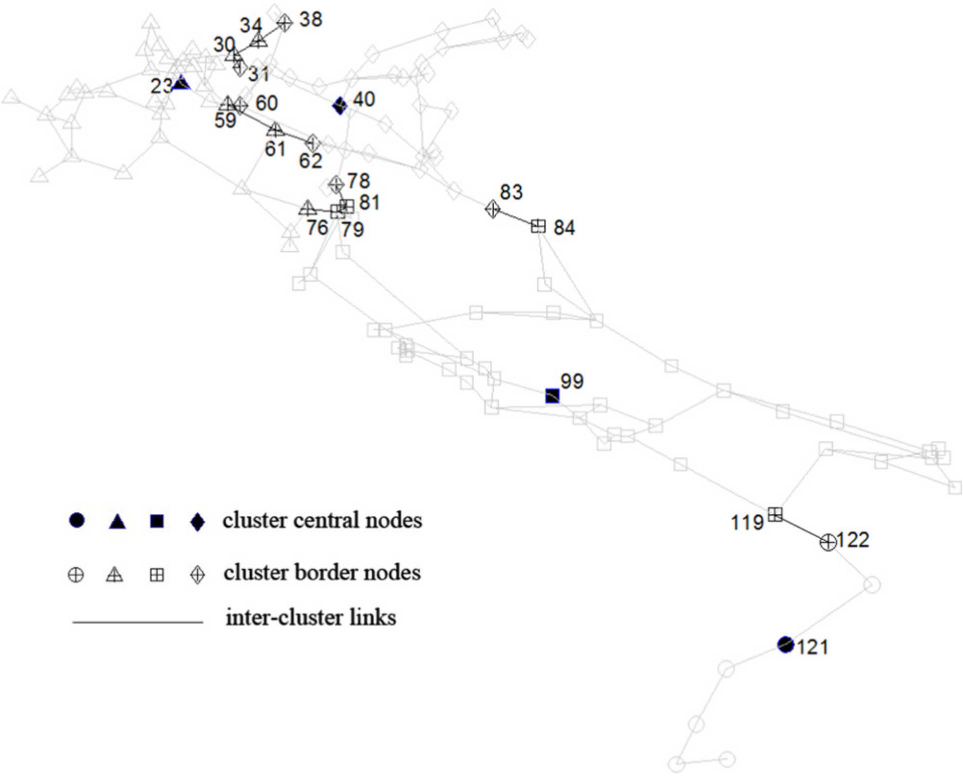


Fig. 4. Inter-cluster links, cluster-border nodes, and central nodes for the 380 kV Italian power transmission network.

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': {(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
Cluster membership value (MV), 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.

Cluster	Critical node	MV	Rank C^I	Rank C^D	Rank C^C	Rank C^B
1	**23**	**0.9999**				
	30	0.7296				
	59	0.7768	13	4	17	8
	61	0.7606	20		9	11
	76	0.5527	15		11	7
2	**40**	**1.0000**			24	18
	31	0.7373				
	34	0.7948				
	60	0.8699		4	15	22
	62	0.8114			8	
	64	0.8394	5	2	1	4
	71	0.9054	22		14	15
3	**86**	**0.9998**			21	
	78	0.4772	10		6	21
	79	0.9198	8	3	3	5
	83	0.4775			22	16
	107	0.7442				24
	110	0.8203	10			10
	112	0.5442				
4	**119**	**0.9993**	4			
	109	0.9466				
	111	0.5724				
	114	0.7314				

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

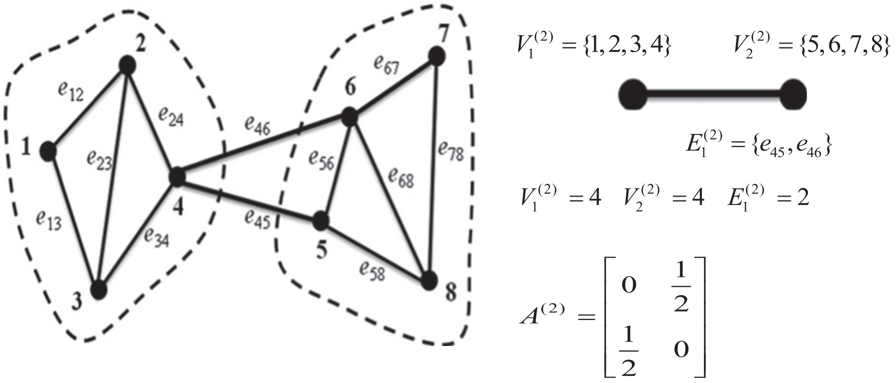

1
2
3
4
5
6
7
8
e_{12}
e_{13}
e_{23}
e_{24}
e_{34}
e_{46}
e_{45}
e_{56}
e_{58}
e_{67}
e_{68}
e_{78}
$V_1^{(2)} = \{1,2,3,4\}$
$V_2^{(2)} = \{5,6,7,8\}$
$E_1^{(2)} = \{e_{45}, e_{46}\}$
$V_1^{(2)} = 4 \quad V_2^{(2)} = 4 \quad E_1^{(2)} = 2$
$A^{(2)} = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$

Fig. 5. Illustrative example of the construction of fictitious networks.

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 cluster-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$G^{(k)}(\Lambda^{(k)},E^{(k)})$. Let us denote as $V_i^{(k)}(i=1,\ldots,n^{(k)})$ 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 $V_i^{(k)}$. These fictitious nodes are connected by $m^{(k)}$ fictitious edges$E^{(k)}=E_1^{(k)},E_2^{(k)},\ldots,E_{m^{(k)}}^{(k)}$. Considering parallel connections, $E_i^{(k)}$ 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 $A^{(k)}\left(V_p^{(k)},V_q^{(k)}\right)=1/|E_{pq}^{(k)}|$if the fictitious nodes $V_p^{(k)}$and $V_q^{(k)}$are connected by fictitious edge $E_{pq}^{(k)}$ 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.

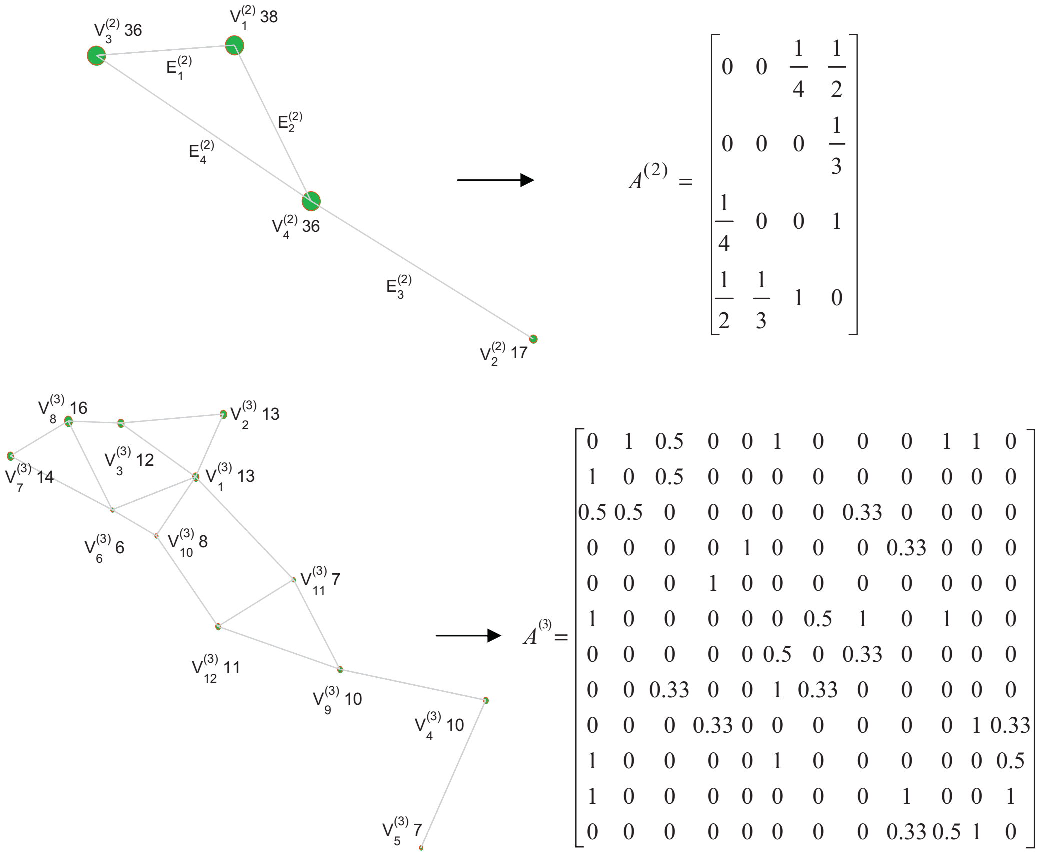
The 380 kV Italian power transmission network has been modeled as a five levels hierarchy (to which correspond five fictitious networks) by successively applying USCA. In Fig. 6, the weighted fictitious networks and their corresponding weighted adjacency matrices at the levels 2 and 3 of the hierarchy are presented for illustration. The number beside the fictitious node $V_i^{(k)}$represents its weight (number of actual nodes included in the virtual node): for example, the weight of $V_3^{(2)}$ 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, ε_{ij}, 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



$$A^{(2)} = \begin{bmatrix} 0 & 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{3} & 1 & 0 \end{bmatrix}$$

$$A^{(3)} = \begin{bmatrix} 0 & 1 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.33 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.33 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.33 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.5 & 1 & 0 \end{bmatrix}$$

Fig. 6. Fictitious networks and their corresponding weighted adjacency matrices at levels 2 and 3 of the hierarchical model for the 380 kV Italian power transmission network.

weighted network. Then, the efficiency of the whole network is given by:

$$E(G) = \frac{\sum_{i \neq j \in G} \varepsilon_{ij}}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}. \tag{8}$$

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^{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 $E(G^{ideal})$ 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 $V_p^{(k)}$and$V_q^{(k)}$: 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^{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.

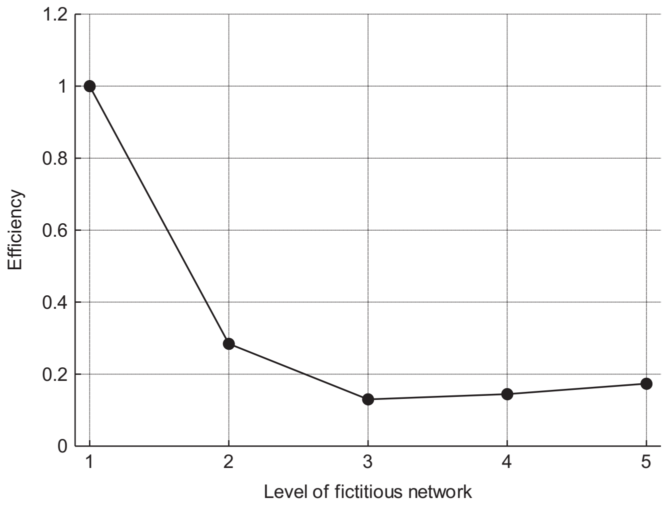


Fig. 7. Network efficiency of fictitious networks at each level of the hierarchy.

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 manually: a large area is first fixed at the coarse granular scale of the map based on the limited information at that level, and then the user can zoom in on that area to get a relatively fine-grained view which offers more local information, based on which a narrower region can be identified, repeating this operation until the desired scale of the map.

Information centrality is used as an illustration to quantify the importance criticality of a cluster on the network. Parallel with the component information centrality definition [18,19,47], we define the information centrality for cluster $V_i^{(k)}$at level k of the hierarchy as the information centrality of its corresponding fictitious node in the fictitious networks, i.e. the relative drop in the fictitious network topological efficiency caused by the removal of all the fictitious edges incident in $V_i^{(k)}$:

$$C^I_{V_i^{(k)}} = \frac{\Delta E(V_i^{(k)})}{E} = \frac{E[G^{(k)}]-E[G_r^{(k)}]}{E[G^{(k)}]} \tag{9}$$

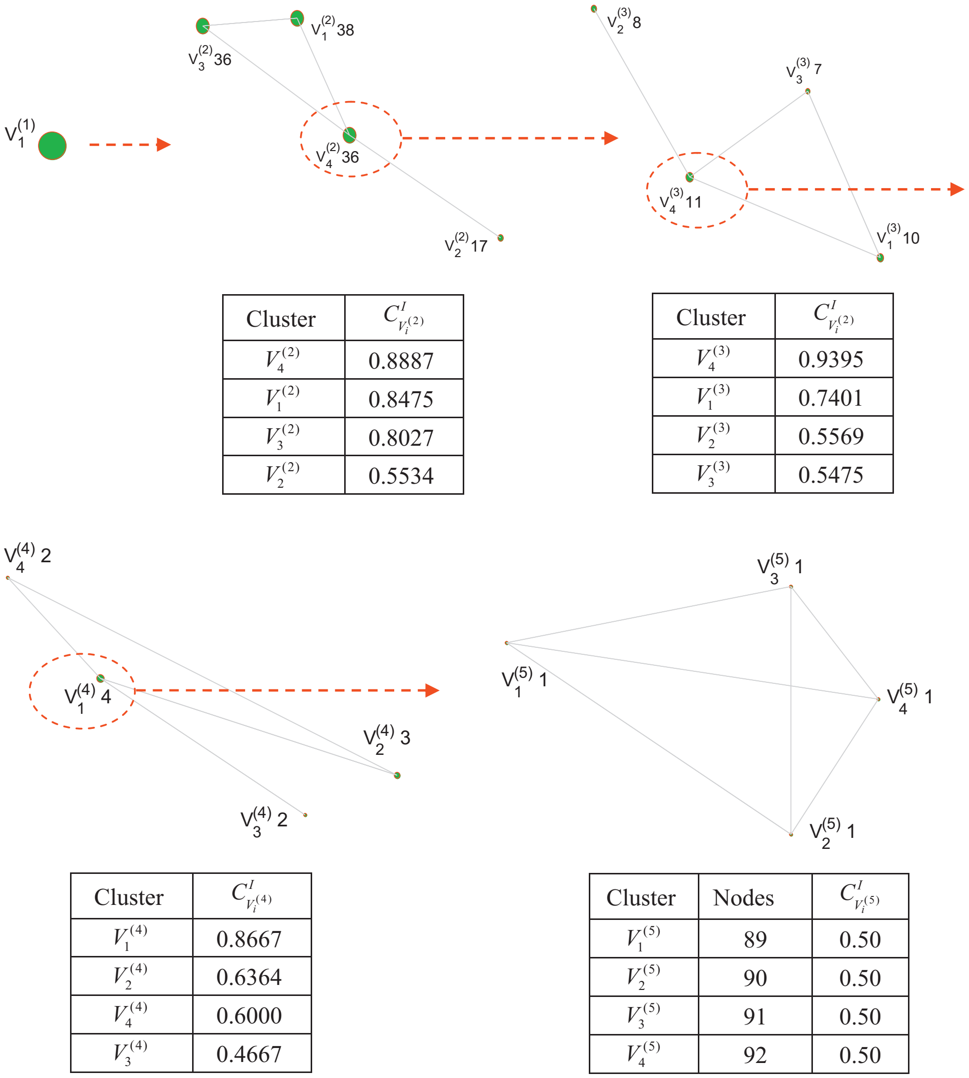
where $G_r^{(k)}$ is the network obtained by removing from the original fictitious network the fictitious edges incident in node $V_i^{(k)}$.

An illustration of the process of zoom-in criticality analysis on the 5-levels 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 $V_4^{(2)}$ owns the highest C^I value; then, the internal topology of $V_4^{(2)}$ at level 3 of the hierarchy is unraveled by zooming into $V_4^{(2)}$. Similarly, the most critical clusters at levels 3 and 4 can be determined as $V_4^{(3)}$ and $V_1^{(4)}$, 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 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



Cluster	$C^I_{V_i^{(2)}}$
$V_4^{(2)}$	0.8887
$V_1^{(2)}$	0.8475
$V_3^{(2)}$	0.8027
$V_2^{(2)}$	0.5534

Cluster	$C^I_{V_i^{(2)}}$
$V_4^{(3)}$	0.9395
$V_1^{(3)}$	0.7401
$V_2^{(3)}$	0.5569
$V_3^{(3)}$	0.5475

Cluster	$C^I_{V_i^{(4)}}$
$V_1^{(4)}$	0.8667
$V_2^{(4)}$	0.6364
$V_4^{(4)}$	0.6000
$V_3^{(4)}$	0.4667

Cluster	Nodes	$C^I_{V_i^{(5)}}$
$V_1^{(5)}$	89	0.50
$V_2^{(5)}$	90	0.50
$V_3^{(5)}$	91	0.50
$V_4^{(5)}$	92	0.50

Fig. 8. The process of zooming-in analysis of information centrality in the hierarchy.

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 $O(n^3)$[48]. Thus, the computation cost of constructing the hierarchical model is$O(n^3l)$, 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

[1] Ellis J, Fisher D, et al. Report to the President's Commission on critical infrastructure protection. S. E. Institute, Editor Carnegie Mellon University, 1997.
[2] Wasserman S, Faust K. Social network analysis. Cambridge: Cambridge University Press; 1994.
[3] Albert R, Barabasi A-L. Statistical mechanics of complex networks. Reviews of Modern Physics 2002;74:47–97.
[4] Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature 2008;453:98–101.
[5] Fortunato S. Community detection in graphs. Physics Reports 2010;486 (75):174.
[6] Scott J. Social network analysis: a handbook. 2nd ed. London: Sage; 2000.
[7] Zio E, Sansavini G. Component criticality in failure cascade processes of network systems. Risk Analysis 2011;31:1196–210.
[8] Zio E, Golea LR, Rocco CM. S. Identifying groups of critical edges in a realistic electrical network by multi-objective genetic algorithms. Reliability Engineering and System Safety 2012;99:172–7.
[9] Zio E. From complexity science to reliability efficiency: a new way of looking at complex network systems and critical infrastructures. International Journal of Critical Infrastructures 2007;3(3):488–508.
[10] Yazdani, A, P Jeffrey. A note on measurement of network vulnerability under random and intentional attacks. arXiv:1006.2791v1 [physics.comp-ph] 14 Jun 2010.
[11] Rocco S, Claudio M, Ramirez-Marquez José Emmanuel. Vulnerability metrics and analysis for communities in complex networks. Reliability Engineering & System Safety 2011;96(10):1360–6.
[12] Latora V, Marchiori M. Economic small-world behavior in weighted networks. European Physical Journal B: Condensed Matter and Complex Systems 2003;32:249–63.
[13] Criado, R, J Flores, A Garcıa del Amo, J Pello, M Romance, M Vela-Pérez. Understanding complex networks through the study of their critical nodes: efficiency, vulnerability and dynamical importance. In: Proceedings of the 2007 international conference on modelling and computation on complex networks and related topics Net-Works 2007 Aranjuez, Spain, 2007.
[14] Criado, R, J Pello, M Romance, M Vela-Pérez. Structural analysis and optimality of vulnerability and efficiency in artificial networks. In: Proceedings of the 2007 international conference on modelling and computation on complex networks and related topics Net-Works 2007 Aranjuez, Spain, 2007.
[15] Sabidussi G. The centrality index of graphs. Psychometrika 1966;31(4): 581–603.
[16] Nieminen J. On the centrality in a graph. Scandinavian Journal of Psychology 1974;15(1):332–6.
[17] Freeman LC. Centrality in social networks conceptual clarification. Social Networks 1978;1(3):215–39.
[18] Latora V, Marchiori M. A measure of centrality based on the network efficiency. New Journal of Physics 2007;9:188 (12 pages).
[19] Little RG. Controlling cascading failure: understanding the vulnerabilities of interconnected infrastructures. Journal of Urban Technology 2002;9(1):109–23.
[20] Mason A, Onnela J, Mucha P. Communities in networks. Notices of the American Mathematical Society 2009;56:9.
[21] Karrer B, Levina E, Newman M. Robustness of community structure in networks. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2008;77:046119.
[22] Eum S, Arakawa S, Murata M. Traffic dynamic in modularity structure of complex networks. In: Proceedings of the fifth international conference on broadband communications, networks and systems, 2008. BROADNETS2008. Issue Date: 8–11Sept; 2008, p. 390–395.
[23] Wu J, Gao Z, Sun H. Cascade and breakdown in scale-free networks with community structure. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2006;74:066111.
[24] Sales-Pardo M, Guimerá R, Moreira A A, Moreira, Amaral LAN. Extracting the hierarchical organization of complex systems. Proceedings of the National academy of Sciences of the United States of America 2007;104:15224–9.
[25] Gómez C, Sánchez-Silva M, Duenas-Osorio L. (2011). Clustering methods for risk assessment of infrastructure network systems. Applications of statistics and probability in Civil Engineering. Faber, Köhler and Nishijima (Eds). ISBN:978-0-415-66986-3.
[26] Brandes U, Erlebach T. Network analysis: methodological foundations. LNCS. Berlin: Springer; 2005 3418.
[27] Rosato V, Bologna S, Tiriticco F. Topological properties of high-voltage electrical transmission networks. Electric Power Systems Research 2007;77: 99–105.
[28] Filippone M, Camastra F, Masulli F, Rovetta S. A survey of kernel and spectral methods for clustering. Pattern Recognition 2008;41(1):176–90.
[29] Scholkopf B, Smola AJ, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 1998;10(5):1299–319.
[30] Jain A K, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys (CSUR) 1999;31(3):264–323.
[31] Schaeffer SE. Graph clustering. Computer Science Review 2007;1(1):27–64.
[32] Von Luxburg Ulrike. A tutorial on spectral clustering. Statistics and Computing 17.4 2007:395–416.
[33] Leguizamon S, HP, Azzali S. Unsupervised Fuzzy *C*-means classification for the determination of dynamically homogeneous areas. p. 851-856, 1996.
[34] Alata M, Molhim M, Ramini A. Optimizing of Fuzzy *C*-means clustering algorithm using GA. World Academy of Science, Engineering and Technology 2008:224–9.
[35] van Dongen, Stijn Marinus. Graph clustering by flow simulation. (2000).
[36] Edachery, Jubin, Arunabha Sen, Franz Brandenburg. Graph clustering using distance-*k* cliques. In: Graph drawing, pp. 98–106. Springer Berlin/Heidelberg; 1999.

[37] Newman M, Girvan M. Finding and evaluating community structure in networks. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2004;69(2):026113.

[38] Good B, de Montjoye Y, Clauset A. The performance of modularity maximization in practical contexts. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2010;81:046106.

[39] Mason A, Onnela J, Mucha P. Communities in networks. Notices of the American Mathematical Society 2009;56:9.

[40] Tan P-N, Steinbach M, Kumar V. Introduction to data mining: chapter 8 cluster analysis—basic concepts and algorithms. Addison-Wesley; 2005.

[41] Davies DL, Bouldin DW. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1979;2:224.

[42] Dunn JC. A Fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 1973;3(3):32–57.

[43] Lázár A, Ábel D, Vicsek T. Modularity measure of networks with overlapping communities. 2010 EPL 90 18001.

[44] Gregory Steve. Fuzzy overlapping communities in networks. Journal of Statistical Mechanics: Theory and Experiment 2011 ISSN:1742-5468, pp. P02017. February.

[45] Crucitti P, Latora V, Marchiori M. Locating critical lines in high-voltage electrical power grids. Fluctuations and Noise Letters 2005;5(2):L201–8.

[46] Latora V, Marchiori M. Efficient behavior of small-world networks. Physical Review Letters 2001;87:198701.

[47] Latora V, Marchiori M. Vulnerability and protection of infrastructure networks. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2005; 71(1):015103.

[48] Schaeffer, Satu Elisa. Graph clustering. Computer Science Review 2007; 1(1):27–64.

[49] Golub G, Van Loan D. Matrix computations. Baltimore: Johns Hopkins University Press; 1996.

[50] Song Yangqiu, Wen-Yen Chen, Hongjie Bai, Chih-Jen Lin, Chang Edward. Parallel spectral clustering. Machine Learning and Knowledge Discovery in Databases 2008:374–89.

[51] Kempe David, Frank McSherry. A decentralized algorithm for spectral analysis. Journal of Computer and System Sciences 2008;74(1):70–83.

[52] Yan, Donghui, Ling Huang, Michael I. Jordan. Fast approximate spectral clustering. In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 907–916. ACM; 2009.