Community detection based on network communicability

Ernesto Estrada
Department of Mathematics and Statistics, Department of Physics, SUPA and Institute of Complex Systems, University of Strathclyde, Glasgow G1 1XQ, United Kingdom

(Received 16 October 2010; accepted 12 January 2011; published online 29 March 2011)

We propose a new method for detecting communities based on the concept of communicability between nodes in a complex network. This method, designated as N-ComBa K-means, uses a normalized version of the adjacency matrix to build the communicability matrix and then applies K-means clustering to find the communities in a graph. We analyze how this method performs for some pathological cases found in the analysis of the detection limit of communities and propose some possible solutions on the basis of the analysis of the ratio of local to global densities in graphs. We use four different quality criteria for detecting the best clustering and compare the new approach with the Girvan–Newman algorithm for the analysis of two "classical" networks: karate club and bottlenose dolphins. Finally, we analyze the more challenging case of homogeneous networks with community structure, for which the Girvan–Newman completely fails in detecting any clustering. The N-ComBa K-means approach performs very well in these situations and we applied it to detect the community structure in an international trade network of miscellaneous manufactures of metal having these characteristics. Some final remarks about the general philosophy of community detection are also discussed. © 2011 American Institute of Physics. [doi:10.1063/1.3552144]

Communities play fundamental organizational and functional roles in many networks representing complex systems. Most of the algorithms that are used to detect these structures use information directly contained in the topology of these networks, such as adjacency and distance relationships. This paper proposes a new technique for identifying communities using the concept of network communicability, which is based on walks on networks. Nodes are grouped into communities according to their capacity of communicating better among them than with outsiders. We analyze the problem of detection limit and propose some possible solutions for the method introduced here based on the analysis of local-to-global densities in networks. A more challenging example consisting of a super-homogeneous network with communities is presented and the advantages of using the current approach are analyzed.

I. INTRODUCTION

Complex networks are the structural skeleton of complex systems, which are ubiquitous in nature, society, and technology. A network is represented by a graph, \( G = (V, E) \), where the set of nodes \( V \) represents the entities of the system and the set of links \( E \) represents the (binary) relationship between these entities. A “microscopic” analysis of a complex network is possible by considering its local topological properties, i.e., those derived from the analysis of close environments around individual nodes and links. Some examples of these local properties are those of centrality, such as degree, closeness, betweenness, etc., or network motifs, i.e., subgraphs more frequently found in a network than in a random graph, and graphlets, which are the number of small subgraphs centered at a given node, which are found in networks. On the other side of the scale we can study some “macroscopic” properties of these complex networks by analyzing their global topological properties. Some examples of these global properties are degree distributions, “small-worldness,” self-similarity, good expansion properties, etc. However, a closer inspection of the structure of many real-world networks gives us insights about a third order of topological organization in these systems. This is a sort of “mesoscopic” organization, understanding it as something between the micro- and macroscopic features of networks, in which nodes and links group together forming some kind of clusters characterized by properties which are more or less independent of the properties of individual nodes and those of the network as a whole. This type of structure is known in network theory as the community structure of a network.

Network communities play important organizational and functional roles in complex networks. Consequently, the identification of communities in complex networks has become one of the most active areas of research in network theory. In a wider context these techniques are used in different disciplines under the umbrella of clustering analysis. Excellent reviews exist in these areas and the reader is directed to them for obtaining an update about these techniques. In complex networks research a community is identified mainly by using different techniques based on the topological information provided by adjacency or shortest-path distance relationships in the network. These techniques...
include the use of link centrality measures like the link betweenness used in the classical Girvan–Newman algorithm, the study of information flow among nodes, and several techniques based on the concept of modularity. The concept of modularity tries to quantify the density of links in a community, which should be significantly larger than the density we would expect if the links in the network were formed by a random process. It has proved to be of great utility as a quality criterion for detecting communities in different scenarios.

A different approach to detect communities was proposed by Estrada and Hatano by using a communicability function between nodes in a network. This function can be interpreted as the thermal Green’s function for the network, and it has found applications in clustering studies of brain networks of patients those have suffered strokes. When used to identify communities, the method uses a very restrictive definition of community. A community is defined in this context as a group of nodes in which every pair has larger intracommunity communicability than intercluster one. This restrictive definition makes that a very large number of highly overlapped communities are detected. Two communities are overlapped if they share at least one node in common. If the proportion of nodes shared between two communities is very high we say that they are highly overlapped. There have been some amendments proposed in the literature to solve this problem, which are basically variations of the criterion for merging communities with a certain degree of overlapping.

Here we propose a relaxation of the definition of community based on the concept of communicability in complex networks. The aim of this work is to develop a method that: (i) solves the problems that arise when communities are defined as cliques in the communicability graph; (ii) allows for community detection in general overcoming some of the problems found by previous methods, such as the detection limit problem; and (iii) solve new challenging problems found in clustering, such as that of detecting communities in homogeneous networks with community structure, where methods like Girvan–Newman fail.

II. NETWORK COMMUNICABILITY

The concept of network communicability was introduced by Estrada and Hatano in 2008. The intuition behind this concept is that in many real-world situations the communication between a pair of nodes in a network does not take place only through the optimal shortest-path routes connecting both nodes. It is possible that the information can flow from one node to another by following all possible routes connecting both nodes in a network. The information can also go back and forth before arriving at the end node of a given route. This immediately invokes the concept of walks in networks. A walk of length \(k\) is a sequence of (not necessarily different) nodes \(v_0, v_1, \ldots, v_{k-1}, v_k\) such that for each \(i = 1, 2 \ldots, k\) there is a link from \(v_{i-1}\) to \(v_i\). Using the concept of walk we define the communicability between two nodes as follows.

**Definition 1.** The communicability between the nodes \(p\) and \(q\) in a network is the weighted sum of all walks starting at node \(p\) and ending at node \(q\), in which the weighting scheme gives more weight to the shortest walks than to the longer ones.

Mathematically, the communicability function can be expressed as follows:

\[
G_{pq} = \sum_{k=0}^{\infty} c_k(A^k)_{pq},
\]

where we have used the fact that the \((p, q)\)-entry of the \(k\)th power of the adjacency matrix, \((A^k)_{pq}\), gives the number of walks of length \(k\) starting at the node \(p\) and ending at the node \(q\). The coefficients \(c_k\) need to fulfill the following requirements: (i) make the series (1) convergent, (ii) giving less weight to longer walks, and (iii) giving real positive values for the communicability. Then, using a factorial penalization we obtain the following communicability function:

\[
G_{pq} = \sum_{k=0}^{\infty} \frac{(A^k)_{pq}}{k!} = (e^A)_{pq},
\]

where \(e^A\) is a matrix function that can be defined using the following Taylor series:

\[
e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots.
\]

Using the spectral decomposition of the adjacency matrix, the communicability function can be expressed as:

\[
G_{pq} = \sum_{j=1}^{n} \lambda_j(p)\varphi_j(q)e^{\lambda_j},
\]

where \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\) are the eigenvalues of the adjacency matrix in a nonincreasing order and \(\varphi_j(p)\) is the \(p\)th entry of the \(j\)th eigenvector which is associated with the eigenvalue \(\lambda_j\).

A. Principles of communicability-based communities

The idea of using communicability for detecting network communities is based on the following intuition. Nodes in a community communicate better among them than with the rest of nodes not included in this community.

Then, our next task is to define what intra- and intercluster communicability means. First, let us write the communicability function for a pair of node in the following way:

\[
G_{pq} = [\varphi_1(p)\varphi_1(q)e^{\lambda_1}]
+ \left[\sum_{2 \leq j \leq n}^{+} \varphi_j(p)\varphi_j(q)e^{\lambda_j} + \sum_{2 \leq j \leq n}^{-} \varphi_j(p)\varphi_j(q)e^{\lambda_j}\right]
+ \left[\sum_{2 \leq j \leq n}^{+} \varphi_j(p)\varphi_j(q)e^{\lambda_j} + \sum_{2 \leq j \leq n}^{-} \varphi_j(p)\varphi_j(q)e^{\lambda_j}\right],
\]

where the signs of the summation sign indicate that the sums are carried out for the components of the eigenvector having
such sign pattern. For instance, $\sum^{++}$ indicates that we are summing only the positive components of $\varphi_j(p)$ and $\varphi_j(q)$.

Now, we can consider each of the signs of the eigenvector components as certain "state" in which the node is. For instance, if $\varphi_j(p) > 0$ we consider that node $p$ is on a "positive" state for the level with energy $\lambda_j$. This state can be simply considered as "pointing" in one direction. Then, $\varphi_j(p) < 0$ means that the node is "pointing" in the opposite direction. We can also interpret these states by considering that an individual in a social network has a "positive" ($\varphi_j(p) > 0$) or "negative" ($\varphi_j(p) < 0$) position with respect to some criterion or point of view or he/she can be completely uninterested ($\varphi_j(p) = 0$) in it. Then, the first term in Eq. (5) represents the consensus configuration in which all the nodes share the same state or point in the same direction. In the second term on the right-hand side of Eq. (5), the nodes $p$ and $q$ have the same sign of the corresponding eigenvector (positive or negative), which means that they will communicate well among them. In other words, they are pointing to the same direction. Consequently, we can consider that $p$ and $q$ are in the same cluster. The last term of Eq. (5), on the other hand, represents a lack of consensus in the states of the nodes $p$ and $q$, i.e., they have different signs of the eigenvector component, which means that they do not communicate well in the network or point to different directions. Then, we can consider that they are in different clusters of the network.

As a consequence of the previous sign pattern analysis we call the second term of Eq. (5) the intracluster communicability and the third term the intercluster communicability between a pair of nodes. The consensus configuration does not give us any information about the community structure of a network. Consequently, we move this term to the other member of Eq. (5) in a way that we obtain the difference between the intra- and intercluster communicability for a network as

$$
\Delta G_{pq} = \sum_{2 \leq j \leq n}^{++} \varphi_j(p) \varphi_j(q) e^{\lambda_j} + \sum_{2 \leq j \leq n}^{-} \varphi_j(p) \varphi_j(q) e^{\lambda_j} \\
+ \sum_{\text{intracluster}}^{+} \varphi_j(p) \varphi_j(q) e^{\lambda_j} + \sum_{\text{intracluster}}^{-} \varphi_j(p) \varphi_j(q) e^{\lambda_j} \\
- \sum_{j=2}^{\text{intercluster}} \varphi_j(p) \varphi_j(q) e^{\lambda_j} - \sum_{j=2}^{\text{intercluster}} \varphi_j(p) \varphi_j(q) e^{\lambda_j},
$$

(6)

where in the last line we used the fact that the intracluster communicability is positive and the intercluster communicability is negative.

We can represent the values of $\Delta G_{pq}$ as the entries of the following matrix:

$$
\Delta G = e^A - e^{\lambda_1} |\varphi_1\rangle \langle \varphi_1|.
$$

(7)

### B. The communicability graph

The first useful transformation that we can carry out for a network based on the communicability function is to condensate the information contained into $\Delta G$ into a new graph. First of all, as we are not interested in the information contained in the main diagonal of this matrix we can simply remove it. That is, we will consider the following matrix instead of $\Delta G$:

$$
C = \Delta G - \text{diag}(\Delta G),
$$

(8)

which will be called the communicability matrix of the network. Secondly, at first instance we can be interested only in the signs of the $\Delta G_{pq}$ entries of this matrix. Consequently, we can introduce the following Heavyside function:

$$
\Theta(x) = \begin{cases} 
1 & \text{if } x > 0 \\
0 & \text{if } x \leq 0,
\end{cases}
$$

(9)

which can be applied in an elementwise way to the $\Delta G$ matrix to obtain a symmetric binary matrix having ones for those pairs of nodes having $\Delta G_{pq} > 0$ and zero otherwise. This new matrix represents a graph, which we call the communicability graph $\Theta(G)$.

### C. Restrictive communicability-based communities

We start here by giving a definition of community based on the concept of communicability.

**Definition 2.** A restrictive communicability-based community is a subset of nodes $C \subset V$ in the network $G = (V, E)$ for which the intracluster communicability is larger than the intercluster one for all $(p, q) \in C$.

Note that because we have moved the consensus state to the other member of the Eq. (5), the fact that $\Delta G_{pq} < 0$ indicates that either the two nodes are not in the same community or that they are in a community formed by all nodes of the graph. This definition is very restrictive in the sense that it only allows nodes $p$ and $q$ in the same community if $\Delta G_{pq} > 0$, such that a community is defined as the subset $C \subset V$, for which $\Delta G_{pq} > 0 \forall (p, q) \in C$. A good consequence of this restrictive definition is that a community can be identified with a clique in the communicability graph. We recall that a clique is a maximum complete subgraph in the network. That is, a subgraph for which every pair of nodes is connected, such as no other node can be included into it keeping this condition. Then, the detection of communities in a network is reduced to finding cliques in the communicability graph. We have used for instance the Bron–Kerbosch algorithm implemented in MATLAB for these purposes.

The problems that arise with this restrictive definition of community based on communicability are evident in the following example. Let us suppose that there is a community $C \subset V$ formed by $n_C$ nodes, and that there is a node $v$ which has $\Delta G_{q,v} > 0$, for 90% of the nodes $q \in C$, but negative with the rest. According to the previous definition of community the node $v$ is not a member of $C$ because $\Delta G_{q,v} \neq 0 \forall q \in C$. As a consequence, there will be communities including $v$ and some of the nodes in $C$, which will have a large degree of overlapping among them. Then, the main disadvantage of this method is a very large proliferation of highly overlapped communities.
III. RELAXED METHOD FOR COMMUNITY DETECTION

In this section we introduce a new paradigm which is aimed to avoid the proliferation of highly overlapped communities produced by the very restrictive definition of communities given before. This paradigm is based on the idea of relaxing the condition \( \Delta G_{pq} > 0 \) \( \forall (p, q) \in C \) for defining communities in a network, so the generic name of “relaxed method.” This definition is given below and it will be the basis of the new approaches studied in this work.

Definition 3. A relaxed communicability-based community is a subset of nodes \( C \subset V \) in the network \( G = (V, E) \) for which the intracluster communicability is larger than the intercluster one for most of the nodes in \( C \), which are then grouped according to a given quantitative criterion.

This definition introduces the necessity for a quantitative criterion in order to group together the nodes into a given community based on their values of communicabilities. There are many ways of doing this and in fact every single method existing today for clustering or partitioning can be repeated by using the communicability information. Some examples of these methods are resumed below:

(i) Considering the communicability matrix \( C \) as the input for clustering algorithms, such as \( K \)-means clustering.\(^{28} \)
(ii) Considering the communicability matrix \( C \) as the input for hierarchical clustering methods to produce a dendrogram with the hierarchical structure of communities in the graph. Examples are the use of Pearson correlation coefficient or Euclidean distance between rows of \( C \) in combination with complete linkage or other linkage methods.\(^{27} \)
(iii) Dichotomize the matrix \( C \) to obtain the communicability graph. Then, using the communicability graph as the input for any clustering algorithm, such as Girvan–Newman,\(^{19} \) Kernighan-Lin,\(^{29} \) spectral methods,\(^{30} \) \( K \)-means clustering,\(^{28} \) or hierarchical methods\(^{27} \) based on the adjacency or distance matrices of the communicability graph.

For the sake of space and time we propose here to use the well known \( K \)-mean clustering technique, which has been recently reviewed for detecting communities in networks\(^{28} \) to group nodes in communities according to their communicabilities. That is, we are going to develop a strategy to detect communities based on communicabilities of networks using approach of type (i) in the previous examples. We will designate here this method as communicability-based (ComBa) \( K \)-means. The method consists of the following steps:

(1) Constructing the matrix \( C = \Delta G - \text{diag}(\Delta G) \).
(2) Use \( C \) as the input for the \( K \)-means algorithm and select the clusters by sorting distances and taking observations at equal distances in the clusters. Other strategies like choosing observations that maximize the initial between-cluster distance or choosing the first \( N \) (number of clusters) observations are also possible. We will always use here the first of them.
(3) Evaluate the quality of the clusters found so far and determine the best one according to specified criteria. The quality criteria used here are described in the next section.

A. Quality criteria for communities

Here, we briefly review the use of some quality criteria for selecting communities. Most of these quality criteria can be used in a two-way fashion.\(^{31,32} \) First, they can be used as criteria for optimization in a search algorithm for detecting communities, or they can be used to measure the quality of different partitions detected by given methods. We are going to use these measures here as quantitative tests for the quality of partitions found by the ComBa \( K \)-means method introduced here.

1. Silhouette index

This index assigns a quality measure for every node in a given community.\(^{34} \) For instance, let us consider a network that has been clustered into the following communities: \( C_1, C_2, \ldots, C_p \). Now, let us consider that the node \( i \) has been clustered in the community \( C_k \). Let \( \bar{d}_{i,C_k} \) be the average distance between node \( i \) and all nodes in the community \( C_k \). For \( j \neq k \) we compute the minimum among all average distances \( \bar{d}_{i,C_j} \), which we designate as \( \bar{d}_{\text{min}} \). Then, the silhouette of the node \( i \in C_k \) is given by

\[
 s(i) = \frac{\bar{d}_{\text{min}} - \bar{d}_{i,C_k}}{\max\{\bar{d}_{i,C_k}, \bar{d}_{\text{min}}\}}.
\]  

(10)

It is easy to see that \(-1 \leq s(i) \leq 1\), where the upper bound is reached when node \( i \) is “well-clustered” in the community \( C_k \) and the lower bound indicates that the node is misclassified. Values close to zero indicate that the node could be also assigned to the nearest neighboring community.

The quality of a community is then given by the average Silhouette, which characterizes the heterogeneity of that cluster

\[
 s(C_k) = \frac{1}{n_{C_k}} \sum_{i=1}^{n_{C_k}} s(i),
\]

(11)

where \( n_{C_k} \) is the number of nodes in the cluster \( C_k \). Finally, for a clustering of a network into a series of communities \( C_1, C_2, \ldots, C_r \), the following global Silhouette index can be used as an effective validity index:

\[
 GS = \frac{1}{r} \sum_{j=1}^{r} s(C_j),
\]

(12)

2. Modularity of a partition

The idea of using of modularity as a quality criterion for a community is that these structural elements have been formed by processes which are far from random.\(^{21} \)
Consequently, the density of links in a community should be significantly larger than the density we would expect if the links in the network were formed by a random process. Mathematically, the modularity has been defined as21

\[
Q = \frac{1}{4m} \sum_y \left( A_{ij} - \frac{k_i k_j}{2m} \right) S_{ij} S_{jr},
\]

where \( m = |E| \), and \( S_{ij} = 1 \) if \( i \in V_r \) or zero otherwise. This expression can be written as the following form:

\[
Q = \frac{1}{2m} tr(S^T B S),
\]

where \( S \) is a rectangular matrix, whose rows represent nodes and columns represents clusters, and the modularity matrix is defined as

\[
B = A - \frac{1}{2m} K J K,
\]

where \( J \) is an all-ones matrix and \( K \) is the diagonal matrix of node degrees. Values of modularity close to zero indicate that the number of intracluster links is not bigger than the expected value for a random network. The maximum value for the modularity is \( Q = 1 \), which indicates strong community structure. The objective is then to use the values of \( Q \) in order to decide which partitions are particularly satisfactory.

3. Performance

The performance of a clustering \( C \) is defined as34

\[
Per(C) = 1 - \frac{2m[1 - 2 Cov(C)] + \sum |C_i|(|C_i| - 1)}{n(n - 1)},
\]

where the coverage of the clustering is defined as

\[
Cov(C) = \frac{1}{m} \sum_{i=1}^c o(C_i).
\]

Here \( o(C_i) \) and \( |C_i| \) are the number of links and nodes in the cluster \( C_i \), respectively and \( m \) is the total number of links. It is evident that a good clustering is the one having most of the links inside clusters with very few links between them. Consequently, the larger the value of the coverage the better the quality of the clustering \( C \). The smaller the value of the performance of \( C \), the better the quality of the clustering.

4. Davies–Bouldin validity index

This index is defined as35

\[
DB(C) = \frac{1}{c} \sum_{i=1}^c \max_{i \neq j} \left( \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right),
\]

where \( \Delta(C_i) \) represents a measurement of the intracluster distance of cluster \( C_i \) and \( \delta(C_i, C_j) \) defines a distance function between clusters \( C_i \) and \( C_j \). There are several possible definitions of these intra- and intercluster distance functions.36 Here we use

\[
\Delta(C_i) = \frac{1}{|C_i|(|C_i| - 1)} \sum_{r \neq s} d_{rs},
\]

and

\[
\delta(C_i, C_j) = \max_{r \in C_i, s \in C_j} \{d_{rs}\}.
\]

Compact clusters with centers which are far away from each other display small values of the Davies–Bouldin (DB) index. Consequently, among two partitions the one displaying the smallest value of the DB index is considered the one of highest quality. However, the DB index is biased when there are communities with only one node as there are no internal distances in this cluster and the index tends to be too small. Then, we exclude this index when there are such kinds of clusters.

B. Resolution limit

An interesting and very challenging problem for detecting communities in networks has been identified and studied by Fortunato and Barthélemy,37 which is known nowadays as the problem of resolution limit. This problem is particularly marked when the modularity is used for detecting communities as it has been found that the resolution limit may prevent the detection of communities which are comparatively small with respect to the whole network, even if they are tightly connected communities, such as cliques. So, in other words the resolution limit consists in the existence of certain pathological cases when a quality measure is fooled by the structure of the network by detecting some communities which are counterintuitive. This problem is not exclusive of the modularity index. For instance, in their seminal paper “On clustering: Good, Bad and Spectral,” Kannan et al.17 have identified some of these pathological cases for minimal cuts, for the diameter and for the 2-median measures. Fortunato11 has remarked that for quality functions which are based on a null model the problem of resolution limit is one to be considered.

Let us take one of these pathological examples affected by the resolution limit for the case,37 at least, of the modularity measure. It consists of two cliques of \( r \) nodes connected to each other and one of them is also connected to two small cliques of, let say 5 nodes each, which at the same time are connected to each other (see Fig. 1). The “natural” communities here are obviously the four cliques. However, it was shown that for \( r = 20 \) the modularity identifies the tripartition consisting of the two cliques of 20 nodes plus a community formed by the joining of the two cliques of 5 nodes as the best partition. This situation is repeated again for \( r = 30 \) as we can see in Table I. The performance index systematically identifies the bipartition as the best one and the Davies–Bouldin index identifies correctly the best clustering for \( r = 20 \) but fails for \( r = 30 \), where it identifies the bipartition.
as the best one. However, the Silhouette index identifies correctly the tetrapartition as the best clustering in both cases. The results obtained in Table I do not mean that the Silhouette index will fail in identifying the tetrapartition when the size of the largest clique increases very much. However, instead of asking for the resolution limit of the method we propose here to change the question we need to ask. That is, is the maximum value of $r$ for which the existence of the four communities makes sense? We all agree with the intuition that as the density of links between a group of $n_C$ nodes tends to one, this group can be considered as a community. The density is simply the number of links between these nodes divided by the maximum possible number of links, which is $n_C(n_C - 1)/2$. Let us concentrate in the three possible communities formed by $K_r$ and the two 5-node cliques $K_5$ in the pathological graph we have been analyzing so far. The density of the unique cluster $\Omega = K_r \cup K_5 \cup K_5$ is given by

$$d(\Omega) = \frac{r^2 - r}{r^2 + 45r + 506}. \tag{21}$$

Then, for the cases analyzed here and elsewhere, $r = 20$ and $r = 30$, the densities are very small, i.e., $d(\Omega) = 0.24$ and $d(\Omega) = 0.35$, respectively. However, for $r \geq 50$ we obtain $d(\Omega) \geq 0.5$, reaching values larger than 0.90 for $r \geq 400$. Finally, $d(\Omega) \to 1$ as $r \to \infty$. Obviously, at this point we are dealing with one single community instead of three. The limit were this transition between three to one communities takes place is something, we are afraid to say, that only experience in specific application fields can provide.

Despite the use of the ComBa $K$-means method identifies the clusters given in Table I, we would like to analyze the communicability graph obtained by dichotomizing $C$ as a possible source of important structural hints. In this case the communicability graph consists of two disconnected clusters, one formed by a clique of 20 nodes and the other consisting of a clique of 30 nodes, which represents the union of the three remaining communities. This situation can well be produced by the “decompensation” in the node degrees of the $K_{20}$ and $K_5$ cliques, which makes that the large clique swallows the two smaller ones in the communicability graph. Obviously, if we would like to use the adjacency matrix of the communicability graph instead of the matrix $C$ to detect communities, only two communities were identifiable. A possible solution for this problem is to normalize the adjacency matrix before obtaining the communicability function. Normalization is a common practice in graph clustering algorithms and several approaches have been tested in the literature.\textsuperscript{38} One of the most appealing normalization procedures consists on replacing the adjacency matrix by the closest (under relative entropy error) doubly stochastic matrix.\textsuperscript{38} This normalization can be mathematically expressed as follows:

$$\tilde{A} = K^{-1/2}AK^{-1/2}, \tag{22}$$

where $K$ is as before, the diagonal matrix of node degrees. It has been proved that for any nonnegative matrix $M$, iterating the process $M^{r+1} \leftarrow D^{-1/2}MD^{-1/2}$ with $D = \text{diag}(M^1)$, converges to a doubly stochastic matrix.\textsuperscript{38}

The communicability matrix is built now from the normalized adjacency matrix using

$$\Delta G' = e^A - e^{i(\tilde{A})}\langle \varphi_1(\tilde{A}) \rangle \langle \varphi_1(\tilde{A}) \rangle^\dagger. \tag{23}$$

Using this normalization the new communicability graph obtained for the graph under study is illustrated in Fig. 2, where it can be seen that it consists of two connected components, one formed by a clique of 20 nodes and the other consisting of a 20-nodes clique and two 6-nodes cliques fused together. This communicability graph clearly indicates the possibility of the existence of four communities, two communities of 20 nodes each and two overlapped communities of 6 nodes each (see Fig. 2 bottom). Then, the normalization process can be a necessary step in order to improve the quality of the clustering obtained from the communicability-based methods. In the rest of this work we are going to use the normalized adjacency matrix. In this case and to avoid any confusion we are going to designate this method as N-ComBa $K$-means. It simply adds to step (1) of the

![FIG. 1. (A) Schematic example of one of the “pathological” graphs considered for the resolution limit problem. (B) Best partition found by the modularity consisting of joining the two small cliques into one community plus two communities formed by each of the largest cliques. (C) Schematic representation of the communicability graph based on the unnormalized adjacency matrix of the graph depicted in (A). (D) Partition suggested by the communicability representation of the communicability graph based on the unnormalized adjacency matrix of the graph depicted in (A).](image)

**TABLE I.** Results obtained by using the N-ComBa $K$-means method for the “pathological” graph illustrated in Fig. 1(A) and four quality criteria for selecting the best clustering. The best clustering according to every quality criterion is boldfaced.

<table>
<thead>
<tr>
<th>$r$</th>
<th>No. of clusters</th>
<th>GS</th>
<th>$Q$</th>
<th>performance</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4</td>
<td><strong>0.605</strong></td>
<td>0.541</td>
<td>0.997</td>
<td><strong>0.533</strong></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.525</td>
<td><strong>0.543</strong></td>
<td>0.978</td>
<td>0.864</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>0.566</td>
<td>0.496</td>
<td><strong>0.818</strong></td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td><strong>0.586</strong></td>
<td>0.519</td>
<td>0.998</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.581</td>
<td><strong>0.520</strong></td>
<td>0.989</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.584</td>
<td>0.499</td>
<td><strong>0.866</strong></td>
<td><strong>0.546</strong></td>
</tr>
</tbody>
</table>

Downloaded 30 Oct 2011 to 130.159.234.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions
algorithm developed in the previous section the normalization of the adjacency matrix.

The other “pathological” case that has been found by Fortunato and Barthélemy\(^37\) consists of a system of \(r\) cliques of size \(r\), which are connected forming a ring as illustrated in Fig. 3(A). One specific case analyzed so far in the literature is the one for which \(p = 30\) and \(r = 5\). The problem arising here with the modularity index is in some way similar to the one previously analyzed. That is, modularity identifies the clustering in which pairs of \(r\)-cliques are considered as communities instead of the most intuitive partition on single cliques. Fortunato and Barthélemy\(^37\) have found that this pathology appears as soon as \(p > r(r - 1) + 2\). For instance, for \(r = 5\) this means that we need a ring with more than 22 cliques. In this simple finding we think that it is hidden the “secret” about the causes of this pathology. This graph is locally very dense (see further) and this local density remains constant if we increase the number of cliques in the ring. However, the global density is very small and it decreases very fast as the number of cliques increases in the ring. Then, what appears to be happening here is a large “decompensation” between the local and global densities in this graph. Let see this in a more quantitative way.

The global density of a graph is given by the ratio of the number of links \(m\) in the graph of size \(n\) to the maximum possible number of such links\(^39\)

$$d_G = \frac{2m}{n(n - 1)}.$$  \hfill (24)

Analogously we can introduce here a local measure of density. Let \(N_a(i) \subseteq V\) be the \(x\)-neighborhood of the node \(i\), defined here as the set of nodes which are at a geodesic distance smaller or equal than \(x\), i.e., \(v_j \in N_a(i)\) iff \(d(v_i, v_j) \leq x\), \(\forall v_j \in V\). This neighborhood of course includes the node \(i\) as well. Let \(n_i = |N_a(i)|\) be the number of nodes and \(m_i\) the number of links in \(N_a(i)\). Then, the local density of node \(i\) in the graph is defined as

$$d_s(i) = \frac{2m_i}{n_i(n_i - 1)}.$$  \hfill (25)

The average local density is then the average of the local density over all nodes in the graph;

$$\bar{d}_s = \frac{1}{n} \sum_{i=1}^{n} d_s(i).$$  \hfill (26)

We will use here the ratio \(\bar{d}_s/d_G\) as a measure of how compensated a graph is in its local and global densities. In general we will use only \(x = 2\) here.

If we consider the graph in Fig. 1 we can see that it has \(d_G = 0.3298\) and \(\bar{d}_s = 0.7573\), which makes the graph compensated as \(\bar{d}_s/d_G = 2.296\). However, the graph in Fig. 3 for \(r = 5\) and \(p = 30\) has \(\bar{d}_s/d_G = 17.029\) because it is very dense locally \(d_s = 0.503\) but very sparse globally \(d_G = 0.0295\).

Then, because the graph is so locally dense in relation with its global density the modularity index, and probably other quality measures, is fooled and identifies pairs of cliques as a better partition than the one single clique in order to try to compensate “artificially” for the local-to-global densities. The communicability-based strategies allow a possible solution to this problem. In the definition of the

FIG. 2. (A) Schematic representation of the communicability graph based on the normalized adjacency matrix of the graph depicted in Fig. 1(A). (B) Representation of the overlapped communities suggested by the communicability graph based on the normalized adjacency matrix.

FIG. 3. (A) Schematic representation of one of the “pathological” graphs considered for the resolution limit problem, and the best partition found by the modularity index. (B) Illustration of the communicability graph based on the normalized adjacency matrix of the graph depicted in (A). (C) Illustration of the type of overlapped communities suggested by the communicability graph depicted in Fig. 3(B). (D) Communicability graph based on normalized adjacency matrix of the graph in (A) but considering that two nodes are connected if \(\Delta G_{pq} > 0.2\).
communicability graph we have used the strict condition that two nodes \( p \) and \( q \) are connected in the communicability if and only if \( \Delta G_{p,q} > 0 \). This produces, in the case of decompensated graphs like the one in Fig. 3(A), communicability graphs formed by many overlapped communities [see Figs. 3(B) and 3(C)]. However, the relaxation of this condition to \( \Delta G_{p,q} > \vartheta \) produces a disentanglement of these communities. This disentanglement generally appears in the form of a disconnected communicability graph in which every connected component is a clique or quasiclique easily identifiable as a community. In the case of the graph in Fig. 3(A) the communicability graph obtained for \( \Delta G_{p,q} > 0.2 \) produces a communicability graph consisting of 30 isolated 5-cliques [see Fig. 3(D)], which avoids the possibility that modularity or any other quality function can identify pairs of cliques as communities.

In conclusion, from the analysis of the pathological cases arising in the resolution limit problem we conclude that when using communicability-based methods for detecting communities we have that: (i) the use of the Silhouette index emerges as the best candidate for analyzing the quality of the partitions found, (ii) the use a normalization of the adjacency matrix is recommended to obtain better cluster separation, and (iii) is the graph is very much decompensated in its local/global densities it is recommended to explore the use of the relaxed condition \( \Delta G_{p,q} > \vartheta \) instead of \( \Delta G_{p,q} > 0 \) in generating the communicability graph, in case such graph is to be used.

IV. TESTING OF THE METHODS

Here we are going to analyze two real-world networks for which there is some “experimental” evidence about their community structures. The first network represents the friendship ties among individuals in a karate club in USA, which is known as the Zachary karate club network. At some point in time the members of this social network were polarized into two different factions due to an argument between the instructor and the president. These two factions act as cohesive groups which can be considered as independent entities in the network. The global and local densities for this network are \( d_G = 0.139 \) and \( d_s = 0.233 \), respectively, which makes the network compensated as \( d_s/d_G = 1.675 \). The other network is formed by 62 bottlenose dolphins living in Doubtful Sound, New Zealand, where links are considered between two animals if they are seen together more frequently than expected at random. These dolphins were split into two cohesive groups after one of them abandoned the place for some time. The global and local densities for this network are \( d_G = 0.084 \) and \( d_s = 0.268 \), respectively, which makes the network compensated as \( d_s/d_G = 3.182 \). These two networks are classically analyzed in most works dealing with the topic of network community due mainly to the existence of these experimental evidences about their cluster structure.

First of all, we illustrate in Fig. 4(A) the Zachary karate club network in which we identify both communities by using different symbols for the nodes in the network. In Fig. 4(B) we illustrate the communicability graph obtained from the unnormalized adjacency matrix of the Zachary karate club network. As is seen in Fig. 4(B) the communicability graph based on the unnormalized adjacency matrix emphasizes the existence of two major communities, which basically correspond to those identified experimentally by Zachary. The natural way for dealing with those communities is by identifying all cliques present in the unnormalized communicability graph. For the sake of completeness we reproduce here the results obtained by using the clique method for finding overlapped communities. Using this approach we have previously identified five communities, from which the first three are very much overlapped, sharing more than 85% of nodes among them. The communities are given below using the same numbering as given in Fig. 4(A):

\[
A := \{10, 15, 16, 19, 21, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34\}; \\
B := \{9, 10, 15, 16, 19, 21, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34\}; \\
C := \{10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34\}; \\
D := \{1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22\}; \\
E := \{3, 10\}.
\]

We have also previously shown that by merging communities which have more than 90% of overlap we reduce this number of communities to only two, i.e., \( C_1 = A \cup B \cup C \cup E \) and \( C_1 = D \cup E \), which have a small overlap among them. However, this situation in which so many highly overlapped communities arise from the analysis of the unnormalized communicability graph can be impossible to deal with for more complex networks. Consequently, we propose here to use the normalized communicability matrix as a source of information for identifying communities in networks. In Fig. 4(C) we illustrate the communicability graph obtained from the normalized adjacency matrix just to emphasize the fact that the structural information revealed by the normalization is completely different from that of the unnormalized case. We remark again that we are not dealing with the use of this normalized communicability graph to detect communities but with the matrix \( \Delta G' = e^{\lambda \mathbf{A}} - e^{\lambda \mathbf{1}(\mathbf{A})} \langle \phi_1(\mathbf{A}) \rangle \langle \phi_1(\mathbf{A}) \rangle \) from which this graph was obtained.

The advantage of using the normalization of the adjacency matrix can be better appreciated when analyzing the network of bottlenose dolphins. In Fig. 5(A) we illustrate the network of bottlenose dolphins indicating the two communities observed so far in it. The communicability graph of this network based on the unnormalized adjacency matrix [Fig. 5(B)] reveals the existence of at least three highly overlapped communities. However, when we use the algorithm based on clique detection in the communicability graph we observe a serious difficulty in using this method. In fact, we detect 47...
FIG. 4. Representation of the Zachary karate club network with nodes of different shapes according to the communities they belong to (A), its communicability graph based on the unnormalized (B) and normalized (C) adjacency matrices.
FIG. 5. Representation of the bottlenose dolphins network with nodes of different shapes according to the communities they belong to (A), its communicability graphs based on the unnormalized (B) and normalized (C) adjacency matrices.
overlapped communities in this network, which range from communities having 25 nodes to small clusters of just 4 nodes. Those communities display a very large overlapping among them emphasizing the previously mentioned problem existing with this restrictive definition of communities. On the other hand, a visual inspection of the communicability graph based on the normalized adjacency matrix [Fig. 5(C)] makes clear the presence of the two main communities but at the same time reveals some more complex internal structure of these two communities.

We now turn our attention to the analysis of the N-ComBa K-means method we have proposed in the previous section. We compare this method with the Girvan–Newman algorithm by using four quality criteria for the communities: modularity, global silhouette, performance, and the Davies–Bouldin index. For the Zachary karate club network we study the clustering of two, three, and four communities, while for the bottlenose dolphins network we explore clustering from two to six communities. The results are given in Table II. For the case of the Zachary karate club network it is known that the Girvan–Newman method produces a bipartition in which all nodes except node 3 are correctly assigned in comparison with the observed partition of this network. This partition is identified by the global silhouette, performance, and the Davies–Bouldin index as the best one among all partitions found by this method. Girvan–Newman is known to identify a tetrapartition of this network as the best one with a value of $Q = 0.409$. The best modularity reported so far for this network is $Q = 0.419$ for a tetrapartition reported by Newman. The N-ComBa K-means approach introduced here correctly classify all nodes into the two observed groups and this bipartition is identified as the best one according to the global silhouette, performance and the Davies–Bouldin index but not by the modularity, which again identifies a tetrapartition as the best one. There are several things to be remarked here. First, the global silhouette index does not distinguish between the bipartitions in which node 3 is swapped from one community to another. The performance index is better for the “wrong” placement of node 3 in the bipartition than when it is placed in the “right” community. Only the Davies–Bouldin index is able to identify the bipartition with node 3 in the right community as slightly better than the other in which this node is in the wrong community. In addition, the tri- and tetrapartitions obtained by using the N-ComBa K-means display better quality parameters than the ones obtained by using the Girvan–Newman method. This includes the case of the modularity where even the tetrapartition obtained by the new method displays slightly improved modularity than the one obtained by using the Girvan–Newman method. The two tetrapartitions are displayed in Fig. 6 for the sake of visual comparisons.

In the case of the bottlenose dolphins network the Girvan–Newman method places correctly all nodes into the two partitions observed experimentally, but this partition is not identified as the best one neither by modularity nor by the global silhouette index. Modularity is known to identify a partition into five clusters as the best one, with a value of $Q = 0.519$. The global silhouette index identifies a tripartition as the best one, but this value of the index is exaggerated by the fact that one of these communities detected by the Girvan–Newman index has only two nodes which are connected to each other. In that case the average intracluster distance is equal to one and this produces that the index artificially increases its value. The N-ComBa K-means method, however, identifies the bipartition as the best one according to three of the four quality methods used. The exception is again the modularity, which in this case identifies a tetrapartition of the network as the best one. We note in passing that the value of the modularity for this tetrapartition is better than the one obtained for the pentapartition detected by the

<table>
<thead>
<tr>
<th>Method</th>
<th>Clusters</th>
<th>$Q$</th>
<th>GS</th>
<th>Performance</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zachary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GN</td>
<td>2</td>
<td>0.360</td>
<td>0.348</td>
<td>0.611</td>
<td>0.742</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.395</td>
<td>0.202</td>
<td>0.729</td>
<td>1.057</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.409</td>
<td>0.211</td>
<td>0.804</td>
<td>1.099</td>
</tr>
<tr>
<td>Comm</td>
<td>2</td>
<td>0.372</td>
<td>0.348</td>
<td>0.617</td>
<td>0.741</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.399</td>
<td>0.271</td>
<td>0.700</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.411</td>
<td>0.258</td>
<td>0.791</td>
<td>1.017</td>
</tr>
<tr>
<td>Dolphins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GN</td>
<td>2</td>
<td>0.379</td>
<td>0.454</td>
<td>0.533</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.381</td>
<td>0.527</td>
<td>0.571</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.458</td>
<td>0.447</td>
<td>0.680</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.519</td>
<td>0.409</td>
<td>0.792</td>
<td>0.734</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.514</td>
<td>0.459</td>
<td>0.810</td>
<td>0.728</td>
</tr>
<tr>
<td>Comm</td>
<td>2</td>
<td>0.379</td>
<td>0.454</td>
<td>0.533</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.449</td>
<td>0.304</td>
<td>0.659</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.524</td>
<td>0.272</td>
<td>0.784</td>
<td>0.772</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.516</td>
<td>0.263</td>
<td>0.832</td>
<td>0.810</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.506</td>
<td>0.294</td>
<td>0.838</td>
<td>0.771</td>
</tr>
</tbody>
</table>

*This value is exaggerated by the fact that one community has only two connected nodes, which makes the average intracluster distance equal to 1.
Girvan–Newman method, i.e., $Q = 0.524$. The best partitions according to the modularity measure obtained by both methods are displayed in Fig. 7.

V. A CHALLENGE: “SUPER-HOMOGENEOUS” NETWORKS WITH COMMUNITIES

In all previous examples, and most, if not all, of the ones treated in the literature, the networks analyzed certainly contain some structural “bottlenecks.” By structural bottlenecks we understand a small subset of nodes/links which by removal separate the network into isolated chunks. This can be seen by simple inspection in the Zachary karate club or dolphins networks as well as in the artificial networks used to analyze the resolution limit problem. We have seen that even those networks we have called here “compensated” display larger local density than the global one. For instance, despite in the Zachary karate club network the ratio between both densities is “only” 1.67, it indicates that the local density almost double the global one. In the dolphins network the local density triplicates the global one, and we have seen such cases in which the local density is much larger than the global one. In all these cases the network has some “rugosities” or heterogeneities that allow most of the methods designed so far to detect communities in their structures. But, what if we consider a network for which the local and global densities are the same, i.e., $d_l = d_G$? It is not difficult to build such kind of networks. For instance, for $x = 2$ any network having diameter smaller or equal to two displays this characteristic. However, the challenge here is to build such a network that also contains community structure. We will call these networks super-homogeneous with communities. Note that these networks are expected to be good expanders, which are networks without bottlenecks. For a formal definition of good expansion networks and how to identify them using the spectral scaling method are discussed in Ref. 42. In Fig. 8 we create a simple model that allows us to build a network having $d_l = d_G$ and presenting two identifiable communities. The process starts by considering two clusters here named $C_1$ and $C_2$, which are highly connected internally. That is, most of the nodes in one cluster are connected to the other members of the cluster. In addition, there

is at least 1 node which is connected to all the nodes in that cluster. Let designate a pair of these nodes as \( v_p \in C_1 \) and \( v_q \in C_2 \). Now, let us connect \( v_p \) to all nodes of \( C_2 \) and \( v_q \) to all nodes of \( C_1 \) as indicated in Fig. 8 by using dotted lines.

Let us now use the Girvan–Newman algorithm to detect communities in this network. What happen is that this algorithm fails completely in identifying any community structure in this network. All partitions from 2 to 12 found with this algorithm display modularity smaller than 0.01. This failure is not exclusive of the Girvan–Newman algorithm but there are others which are also unable to detect any community structure in this graph. For instance, the use of the adjacency matrix of the graph for building a similarity matrix based on the Pearson correlation coefficient produces a complete linkage dendrogram that groups 4 nodes in one cluster and 8 in another, and the use of \( K \)-means clustering based on the adjacency matrix is not better than that. In contrast with these frustrating results, the N-ComBa \( K \)-means method introduced here identifies the two communities formed by 6 nodes each. The modularity for this partition is \( Q = 0.138 \) and the global silhouette is 0.365. But instead of continuing the analysis of this artificial graph and its possible variations, think that the number of communities can be increased in a similar way as explained in Fig. 8, we would turn our attention to a real network with this property.

The first question that arises in this context is where to find real-world networks that display this structural homogeneity. Good candidates are networks of international trade. International trade is organized into certain geopolitical groups which can define the communities of the whole network. However, there are some hubs in these communities that have trade with most of the countries in all other communities, resembling very much the model we have previously described in Fig. 8. Here we study a trade network of miscellaneous manufactures of metal (MMM) among 80 countries in 1994. The data were compiled for all countries with entries in the paper version of the Commodity Trade Statistics published by the United Nations. For some countries the authors used the 1993 data (Austria, Seychelles, Bangladesh, Croatia, and Barbados) or the 1995 data (South
Africa and Ecuador) because they were not available for the year 1994. Countries which are not sovereign are excluded because additional economic data were not available: Faeroe Islands and Greenland, which belong to Denmark, and Macau (Portugal). Most missing countries are located in central Africa and the Middle East, or belong to the former USSR. Despite the network is weighted and directed here, we use an undirected and symmetrized version of this network. The symmetrization of the adjacency matrix is very much justified by the fact that in general, all countries that export MMM also import them from other countries. However, there are 24 countries that are only importers and have no export of MMM at all. They are: Kuwait, Latvia, Philippines, French Guiana, Bangladesh, Fiji, Reunion, Madagascar, Seychelles, Martinique, Mauritius, Belize, Morocco, Sri Lanka, Algeria, Nicaragua, Iceland, Oman, Pakistan, Cyprus, Paraguay, Guadalupe, Uruguay, and Jordan. If the degree of the nodes is analyzed it is observed that the countries having the larger number of exports are (in order): Germany, USA, Italy, U.K., China, Japan, France, Belgium/Luxemburg, The Netherlands, and Sweden. When an undirected version of this network is considered the same countries appear as the larger exporters, with only tiny variations in the order: Germany, USA, Italy, U.K., Japan, China, France, The Netherlands, Belgium/Luxemburg, and Sweden. On the other hand, the in-degree is practically the same for every country. For instance, the average in-degree is 12.475 and its standard deviation is only 3.15. Then, we can consider here only the undirected and unweighted version of this network. As usual the nodes represent the countries and a link exists between two countries if one of them imports miscellaneous manufactures of metal (MMM) from the other. The undirected network is depicted in Fig. 9.

One characteristic feature of this network is that it has diameter equal to two and consequently $\frac{d_{G}}{c_{G}} = 0.277$. We have explored the identification of communities using the N-ComBa K-means method described in this work. In Table III we illustrate the results obtained by using the four quality criteria we previously described and used in this work. As can be seen, modularity and global silhouette identify the partition into three clusters as the best partition, while the performance and the Davies–Bouldin index identify the bipartition, closely followed by the tripartition, as the best ones. Then, we select the partition into three communities as the best one in order to further investigate it in more details. Before that, we would like to explore what the Girvan–Newman algorithm produces for this network. The results are quite disappointing as no single partition is obtained among all possible ones between 2 and 20. In all cases the modularity of the partitions found is no larger than 0.001, and the partitions consist of many single-node communities.

![FIG. 8. Example of the construction of a super-homogeneous network with communities. (A) Example of the two clusters in which nodes $v_p \in C_1$ and $v_q \in C_2$ are connected to all members of their respective communities. (B) Second step of the construction in which the nodes $v_p \in C_1$ and $v_q \in C_2$ are now connected to the remaining nodes of the graph.](image)

![FIG. 9. Network of international trade of miscellaneous metal manufactures (MMM) in which nodes represent countries and two nodes are connected if one of them imports MMM from the other.](image)
The three communities found by the N-ComBa K-means are displayed in a pictorial way in Fig. 10. The first community, colored in clear gray, is formed by all Latin-American countries plus Canada, USA, and Spain. The presence of USA and Canada in this community is plenty justified due to the geographical closeness of these two countries to central and South-American ones, while the inclusion of Spain is understood by the long tradition of trade between this country and its former colonies in America. The second community, colored in dark gray, consists of most European countries, except Spain, and many of their former colonies or regions of influence in Africa, Middle East, and South America (French Guiana). The third community, colored in black, is formed by Asian and Oceanic countries. The countries which are the global hubs of this trade network are identified as follow: USA, which is connected to 72 of the 80 countries in the web is the main global connector placed in the American cluster; Germany, Italy, and U.K. connected to 77, 72, and 63 countries, respectively are the main global hubs located in Europe; and Japan and China, connected to 58 and 57 countries, respectively, are the global hubs in the third community. These results are encouraging for the use of the communicability matrix for identifying communities in networks.

VI. ABOUT COMPUTATIONAL COMPLEXITY

An important aspect of the methods used for finding communities in networks is that of their computational complexity. This is of vital importance when considering the application of these methods to search for communities in huge complex networks. The Kernighan–Lin algorithm for instance, is known to be very slow with complexity which is $O(n^4)$ for dense networks, and $O(n^3)$ for sparse ones. The Girvan–Newman algorithm is known to take time $O(n^2)$ in sparse networks due to the fact that betweenness-based methods are quite slow. The spectral modularity maximization method proposed by Newman performs in $O(n^3)$ for very dense graphs and in $O(n^2)$ for sparse ones. The reader can find the analysis of the complexity of these algorithms in Ref. 39.

In this context we analyze briefly the complexity of the N-ComBa K-means method we proposed here. The most time consuming operation in this method is to find the exponential adjacency matrix. In the worst case scenario, computing $\exp(A)$ is $O(n^3)$, which is the case for dense graphs. However, for sparse adjacency matrices, which contains only $O(n)$ nonzero entries, it is possible to compute $\exp(A)$ in less than $O(n^3)$. It is certain than diagonalization-based methods are slow but Benzi and Razouk have analyzed the existence of polynomial algorithms which perform in $O(n)$. The most popular of them is probably the expansion of $\exp(A)$ in the Chebyshev polynomial basis: since the coefficients of the expansion decay faster than exponentially, this series converges extremely fast. Benzi and Boito have also found excellent approximations for the communicability function, which can be exploited in order to dramatically reduce the computational time in extremely large networks. All in all, the current method performs better or as good as some of the best ones described in the literature for both dense and sparse networks. Other approaches that use

<table>
<thead>
<tr>
<th>Clusters</th>
<th>$Q$</th>
<th>GS</th>
<th>Perf</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.172</td>
<td>0.135</td>
<td>0.599</td>
<td>1.080</td>
</tr>
<tr>
<td>3</td>
<td>0.215</td>
<td>0.193</td>
<td>0.689</td>
<td>1.304</td>
</tr>
<tr>
<td>4</td>
<td>0.144</td>
<td>0.157</td>
<td>0.711</td>
<td>1.647</td>
</tr>
<tr>
<td>5</td>
<td>0.146</td>
<td>0.159</td>
<td>0.748</td>
<td>1.651</td>
</tr>
<tr>
<td>6</td>
<td>0.063</td>
<td>0.012</td>
<td>0.702</td>
<td>1.531</td>
</tr>
</tbody>
</table>

The three communities found by the N-ComBa K-means are displayed in a pictorial way in Fig. 10. The first community is represented by coloring countries with clear gray, the second by coloring them with dark gray, and the last by coloring the countries with black.

![Fig. 10. Representation of the three communities obtained as the best partition by using N-ComBa K-means method for the MMM network represented in Fig. 9. The first community is represented by coloring countries with clear gray, the second by coloring them with dark gray, and the last by coloring the countries with black.](image)
different matrix functions for the communicability have been explored in the recent literature\textsuperscript{47,48} and they can be considered as alternatives to expand the methods proposed here to improve results as well as computational efficiency.

VII. CONCLUSIONS AND OUTLOOK

We have analyzed here the use of the communicability matrix of a network with normalized adjacency matrix as a source of methods for detecting communities in networks. We have studied here one method, which is based on the use of K-means clustering using a normalized communicability matrix. As we have previously stated, the communicability matrix can be used in other clustering algorithms or even as a preprocessing step for detecting communities in networks. We have seen here that the relation between the local and global densities in networks appears as a good indicator of the type of strategies that we should use for detecting communities. More work on this ratio of densities should be done to clarify its use as a network parameter with possible utility in detecting the community structure of networks. However, the results found here indicates that methods based on the normalized communicability can play an important role in networks with a large variety of density ratios, ranging from those super-homogeneous in which the ratio is 1 to those which are significantly much more dense locally than globally. In a more specific context, we would like to remark that the N-ComBa K-means method developed here solve the previous problems that have arisen when a very restrictive definitions of community, like the one using the cliques of the communicability graph, were used. In the current scenario we do not need the clique detection method, which is also a very time consuming procedure. In addition, the current method is based on very solid pattern-recognition techniques for which there is vast experience in both theory and computational applications.

Finally, we would like to make some remarks about the general philosophy of community detection in networks. We know that there are many methods and algorithms described today to detect communities in networks. This number increases dramatically if we also consider those methods of clustering developed and used today in the field of pattern recognition. The existence of this large number of methods is an indication that there is no such thing as “the best clustering method.” Probably, this method is not realizable and we have to deal with some methods that solve one problem and other methods that solve others. We have found here one kind of network, the super-homogeneous network with communities, which are not well clustered by the Girvan–Newman algorithm and probably by other methods. The N-ComBa K-means method proposed here appears to solve this problem effectively. However, it could fail in some other cases in which the Girvan–Newman algorithm is successful. The moral is that we should not discard one method simply because it fails for one type of networks, if it works perfectly for others in which other methods fail. What we have to know, and currently do not know, is which methods perform well for certain kind of networks and fail for others. This could be an area in which more investigation should be done and probably the necessity of new methods arises for certain particular classes of networks. Perhaps we are lucky enough and the existing arsenal of methods is enough to solve all the clustering problems in all the existing types of networks. In such case it should be important to classify the methods according to the type of problems they are suitable for.

ACKNOWLEDGMENTS

Constructive comments and suggestions of two anonymous referees contributed significantly to the improvement of the material presented here. Professor M. Benzi (Emory Univ.) is thanked for his help with the discussion about the complexity of matrix exponentials. Partial financial support from the Principal of the University of Strathclyde under the New Professor’s Fund is also thanked.


