Detecting network communities beyond assortativity-related attributes

Xin Liu,[1,2,3,*] Tsuyoshi Murata,[1] and Ken Wakita[1,2]
[1]*Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan*
[2]*CREST, Japan Science and Technology Agency, K's Gobancho, 7 Gobancho, Chiyoda, Tokyo, 102-0076 Japan*
[3]*Wuhan University of Technology, 122 Luoshi Road, Wuhan, Hubei, 430070 China*
(Received 28 November 2013; revised manuscript received 7 April 2014; published 10 July 2014)

In network science, assortativity refers to the tendency of links to exist between nodes with similar attributes. In social networks, for example, links tend to exist between individuals of similar age, nationality, location, race, income, educational level, religious belief, and language. Thus, various attributes jointly affect the network topology. An interesting problem is to detect community structure beyond some specific assortativity-related attributes ρ, i.e., to take out the effect of ρ on network topology and reveal the hidden community structures which are due to other attributes. An approach to this problem is to redefine the null model of the modularity measure, so as to simulate the effect of ρ on network topology. However, a challenge is that we do not know to what extent the network topology is affected by ρ and by other attributes. In this paper, we propose a distance modularity, which allows us to freely choose any suitable function to simulate the effect of ρ. Such freedom can help us probe the effect of ρ and detect the hidden communities which are due to other attributes. We test the effectiveness of distance modularity on synthetic benchmarks and two real-world networks.

DOI: 10.1103/PhysRevE.90.012806 PACS number(s): 89.75.Fb, 89.75.Hc

I. INTRODUCTION

Many social, biological, and information systems can be described by networks, where nodes represent fundamental entities of a system, such as individuals, users, genes, web pages, and links represent relations or interactions between the entities [1]. In recent years, there has been a surge of interest in the analysis of networks. A highly discussed topic is community detection—the detection of groups of network nodes, known as communities, within which links are dense, but between which links are sparse [2]. Community detection [3–7] is considered as a crucial step toward inferring function units of the underlying system, such as collections of pages on closely related topics on the web or groups of people with common interest in social media.

People observed that in real-world networks links tend to exist between nodes with similar attributes. For example, in social networks individuals commonly choose to associate with others of similar age, nationality, location, race, income, educational level, religious belief, and language as themselves. This tendency is known as *assortativity* (also known as assortative mixing or homophily) [8–11]. To see whether an attribute is assortativity related, or whether it is correlated with the network topology, Bavaud proposed a modes permutation test [12].

Assortativity indicates that various attributes jointly affect the network topology, either directly or indirectly. An interesting problem is to detect community structure beyond some specific attributes, represented by a variable vector ρ. In other words, the goal is to take out the effect of ρ on network topology and reveal the hidden community structures which are due to other attributes, represented by a variable vector $\bar{\rho}$. Note that ρ is observable, while $\bar{\rho}$ may contain some latent variables. For example, one may be interested in studying the community structure, which is not due to age, but due to religious belief, educational level, income, and some attributes which are not observed in social networks. A challenge of this problem is that we do not know to what extent the network topology is affected by ρ and $\bar{\rho}$. Moreover, some attributes in ρ and $\bar{\rho}$ can be correlated. This makes the problem even more complex, since it becomes more difficult to disentangle ρ and $\bar{\rho}$ [13].

The popular community detection method which relies on optimization of a quantity measure called NG modularity (Newman-Girvan modularity) [14] cannot solve the above problem. This is because the definition of NG modularity does not take node attributes ρ into account. The definition of NG modularity involves a comparison between the observed network and a *null model*. In order to solve the problem, this null model should be redefined to simulate the effect of ρ on network topology, so that such effect can be taken out when compared to the observed network. Following this idea, Expert *et al.* defined a new null model based on an empirically determined probability distribution and proposed spatial modularity [15]. However, recent experiments by Cerina *et al.* showed that spatial modularity still cannot handle this problem well, especially when there is a correlation between ρ and $\bar{\rho}$ [13].

In this paper, we extend NG modularity and propose distance modularity. In particular, we define a general null model which allows us to freely choose any suitable function to simulate the effect of ρ on network topology. Such freedom can help us probe the effect of ρ and detect the hidden communities which are due to $\bar{\rho}$. We use synthetic networks and two real-world networks to demonstrate the effectiveness of distance modularity. We analyze the reasons why spatial modularity fails in these examples.

The rest of the paper is organized as follows. Section II reviews NG modularity and spatial modularity. Section III introduces our null model and distance modularity. Section IV presents experiments, followed by a conclusion in Sec. V.

*To whom correspondence should be addressed: tsinllew@ai.cs.titech.ac.jp

II. RELATED WORK

In this section, we give a review of NG modularity and spatial modularity. Before this, let us first introduce the prototype of these two measures—the modularity [14,16].

A. Modularity

Modularity is a quantity measure for evaluating the quality of a partition of a network into communities. The definition of modularity compares the fraction of within-community links in the observed network minus the expected value of that fraction in some equivalent randomized network. This randomized network is called the null model, which serves as a reference. The mathematical expression of modularity in an undirected network reads

$$Q(\mathcal{L}) = \frac{1}{2m} \sum_{i,j=1}^{n} (A_{ij} - P_{ij})\,\delta(l_i,l_j), \tag{1}$$

where n is the number of nodes, m is the number of links, $\mathcal{L} = \{l_1,l_2,\cdots,l_n\}$ is a partition with element l_i indicating the community membership of the ith node v_i, A_{ij} is the number of links between v_i and v_j in the observed network, P_{ij} is the expected value of that number in the null model, and δ is Kronecker's delta.

To make Eq. (1) significant, the number of links in the null model should equal that number in the observed network. That is,

$$\sum_{i,j=1}^{n} P_{ij} = \sum_{i,j=1}^{n} A_{ij} = 2m. \tag{2}$$

Apart from the constraint Eq. (2), there is some freedom about choosing the null model, and different null models produce variants of modularity. According to Eqs. (1) and (2), it is clear that $Q \in [-1,1]$. For a given network, the higher the Q, the better the partition $\mathcal{L}$.

B. NG modularity

A popular choice of the null model proposed by Newman and Girvan [14] is the configuration model [17], which preserves the degree sequence of the observed network. Specifically, the expected number of links between v_i and v_j in this null model is

$$P_{ij}^{\mathrm{NG}} = k_i k_j / 2m, \tag{3}$$

where $k_i = \sum_{j=1}^{n} A_{ij}$ is the degree of v_i. Replacing P_{ij} in Eq. (1) by P_{ij}^{NG} gives NG modularity (Q^{NG}), which is widely used in practice.

C. Spatial modularity

The configuration model used in NG modularity does not take attributes into account. To detect community structure beyond ρ, the null model should be redefined to simulate the effect of ρ on network topology, so that such effect can be taken out when compared to the observed network. For this reason, Expert *et al.* defined a new null model (originally for simulating the effect of space on network topology) [15]. Specifically, the expected number of links between v_i and v_j in this null model is

$$P_{ij}^{\mathrm{spa}} = h_i h_j p(d_{ij}), \tag{4}$$

where d_{ij} denotes the distance between v_i and v_j in terms of ρ, h_i indicates the importance of v_i, and $p(d)$ is the probability that two nodes are connected at a distance d and can be obtained empirically from the observed network by

$$p(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} h_i h_j}. \tag{5}$$

Replacing P_{ij} in Eq. (1) by P_{ij}^{spa} gives spatial modularity (Q^{spa}).

From Eqs. (4) and (5), we can derive that

$$\sum_{i,j|d_{ij}=d} P_{ij}^{\mathrm{spa}} = \sum_{i,j|d_{ij}=d} A_{ij}. \tag{6}$$

This indicates that the number of links between nodes at distance d in the spatial modularity null model is the same as that number in the observed network. In other words, this null model assumes that only ρ affects network topology, and thus it simulates the effect of ρ as what can be observed in network topology.

III. DISTANCE MODULARITY

In this section, we first propose a null model which allows us to freely choose any suitable function to simulate the effect of ρ on network topology, and we then present distance modularity for detecting communities beyond ρ. For simplicity, we only consider the case of undirected networks.

A. Our null model

We propose the following null model. In this model, the expected number of links between v_i and v_j is

$$P_{ij}^{\mathrm{dist}} = (\tilde{P}_{ij} + \tilde{P}_{ji})/2, \tag{7}$$

where

$$\tilde{P}_{ij} = \frac{k_i k_j f(d_{ij})}{\sum_{t=1}^{n} k_t f(d_{ti})}. \tag{8}$$

Here, $f : \mathbb{R}_{\geq 0} \to [0,1]$ is a function which can be specified freely (we will explain the significance of f later). k_i denotes the degree of v_i in the observed network. d_{ij} denotes the distance between v_i and v_j in terms of ρ. Note that d_{ij} is computed by a distance function between the attribute variables on v_i and v_j, denoted by ρ_i and ρ_j [18]. Generally, d_{ij} should satisfy the following constraints:

$$d_{ij} \geqslant 0 \text{ with equality IFF } \rho_i = \rho_j, \tag{9}$$

$$d_{ij} = d_{ji}. \tag{10}$$

Our null model has the following properties. First, from Eq. (7) we can find that

$$P_{ij}^{\mathrm{dist}} = P_{ji}^{\mathrm{dist}}. \tag{11}$$

This shows that links in our null model are undirected. Second, from Eq. (8) we can derive that

$$\sum_{i,j=1}^{n} \tilde{P}_{ij} = \sum_{i,j=1}^{n} \tilde{P}_{ji} = \sum_{i=1}^{n} k_i = 2m, \tag{12}$$

and hence $\sum_{i,j=1}^{n} P_{ij}^{\text{dist}} = 2m$. This indicates that our null model preserves the number of links of the observed network. Third, from Eqs. (7) and (8) we can derive that

$$k_i^{\text{dist}} = \sum_{j=1}^{n} P_{ij}^{\text{dist}} = \frac{k_i}{2}\left[1 + \sum_{j=1}^{n} \frac{k_j f(d_{ij})}{\sum_{t=1}^{n} k_t f(d_{tj})}\right]. \tag{13}$$

This implies that a node v_i which has high degree in the observed network tends to have high degree in our null model. In addition, note that k_i^{dist} is not necessarily equal to k_i. Thus, this may induce a slight shift in the distribution of k_i^{dist}, as compared to the distribution of k_i. Fourth, our null model can simulate the effect of ρ on network topology by specifying appropriate function f. For example, the following are true.

(1) If we specify $f(d) = e^{(-d)}$, two nodes which are similar have a higher chance of getting connected.

(2) If we specify $f(d) = \{^{1\ \text{if}\ d \leqslant \sigma}_{0\ \text{otherwise}}$, a node can only connect to those at a distance of no more than σ.

(3) If we specify $f(d) = 1$, P_{ij}^{dist} is not related to d_{ij}, and our null model recovers the configuration model used in NG modularity.

(4) If we specify $f(d) = e^{(-1/d)}$, two nodes which are dissimilar have a higher chance of getting connected—this is actually a disassortativity effect [19].

B. Distance modularity

Based on our null model, we define distance modularity as

$$Q^{\text{dist}}(\mathcal{L}) = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - P_{ij}^{\text{dist}}\right) \delta(l_i, l_j). \tag{14}$$

Note that we need to specify the function f before using distance modularity.

Like NG modularity, optimizing distance modularity is NP-hard [20]. We can use heuristics such as the Louvain algorithm [21] and the advanced modularity-specialized label propagation algorithm (LPM+) [22], which were originally developed for optimizing NG modularity. The time complexity of Louvain and LPM+ algorithms can be analyzed as follows. First, we need to compute the denominator part of Eq. (8) and keep the results for $i = 1, \ldots, n$ in memory. This operation requires a complexity of $O(n^2)$. Second, suppose $c = |\{l_i | i = 1, \ldots, n\}|$ is the number of communities. Also suppose we have enough memory to keep the $n \times c$ matrix $\bar{P}^{\text{dist}}$, whose elements are defined as $\bar{P}_{il}^{\text{dist}} = \sum_{j=1 \wedge j \neq i}^{n} P_{ij}^{\text{dist}} \delta(l_j, l)$. At the initial stage where each node forms a unique community, to save $\bar{P}^{\text{dist}}$ into memory requires a complexity of $O(n^2)$. Third, another computationally intensive step of these two algorithms is to move a node v_i to a new community that would result in the highest gain in Q^{dist}. We can derive that v_i's new community membership can be computed as [22]

$$l_i^{\text{new}} = \underset{l \in \{l_i\} \cup \{l_j | A_{ij} \neq 0\}}{\arg\max} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} A_{ij}\, \delta(l_j, l) - \bar{P}_{il}^{\text{dist}} \right). \tag{15}$$

With $\bar{P}_{il}^{\text{dist}}$ kept in real time, this computation requires a complexity of $O(k_i)$. In addition, adjusting related elements of $\bar{P}^{\text{dist}}$ (at most $2k_i$ elements) due to the movement of v_i requires another complexity of $O(k_i)$. Note that the node movement step is repeated sequentially for each node and iteratively until no gain in Q^{dist} can be attained. Suppose the number of iterations is r, which is a small number in practice. Then, the total time complexity of the two algorithms is near $O(n^2 + rm)$.

In the following, we discuss how to detect communities beyond ρ by distance modularity. A key point to this problem is to find an appropriate function f to simulate the assortativity effect of ρ in the null model, so that such effect can be taken out when compared to the observed network. However, a challenge is that we do not know to what extent the network topology is affected by ρ. Thus, it is difficult to find a function f directly.

In the framework of distance modularity, our procedure contains three steps. First, we choose a distance function [18] to compute d_{ij}. Second, we probe f using parameterized functions such as $f(d) = e^{-(d/\sigma)^2}$ and $[1 + (d/\sigma)^2]^{-1}$, where $\sigma \in (0, +\infty)$ is a parameter [23]. A benefit of these functions is that we can tune the parameter to simulate the assortativity effect at different degrees. Take the function $f(d) = e^{-(d/\sigma)^2}$ as an example. At the extreme of $\sigma \to 0+$, the null model has the strongest assortativity effect and a node v_i can only connect to its most similar nodes v_j satisfying $d_{ij} = 0$. At the extreme of $\sigma \to +\infty$, we can derive

$$\lim_{\sigma \to +\infty} \left(P_{ij}^{\text{dist}} | f(d) = e^{-(d/\sigma)^2}\right) = k_i k_j / 2m. \tag{16}$$

That is, the null model has no assortativity effect. As σ increases from zero to $+\infty$ (see Fig. 1), the assortativity effect in the null model gradually fades.

In the third step, we optimize distance modularity at various estimated values of the parameter and select one that brings the "best" partition. A possible method is the alternative parameter selection method proposed by Expert *et al.* [15], which seeks to find a consensus partition. More precisely, suppose $\mathcal{L}_{\sigma_i}$ is a partition obtained by optimizing distance modularity at $\sigma = \sigma_i (i = 1, \ldots, s)$, which are possible values of the parameter. We compute the average normalized mutual information (NMI) [5,24], denoted by I^{avg}, of each partition as follows

$$I^{\text{avg}}(\mathcal{L}_{\sigma_i}) = \sum_{\substack{j=1 \\ j \neq i}}^{s} I(\mathcal{L}_{\sigma_i}, \mathcal{L}_{\sigma_j}), \tag{17}$$

where I is the function of NMI. The partition with the highest I^{avg} score, which is the consensus partition and the closest to the others, is our final partition for communities beyond ρ.

The time complexity of the above procedure is as follows. First, optimizing distance modularity at $\sigma = \sigma_1, \ldots, \sigma_s$ requires a complexity of $O[s(n^2 + rm)]$. Second, computing the I^{avg} score for s partitions requires a complexity of $O[s^2(n + c^2)]$. Finally, the total complexity of detecting communities beyond ρ by distance modularity is

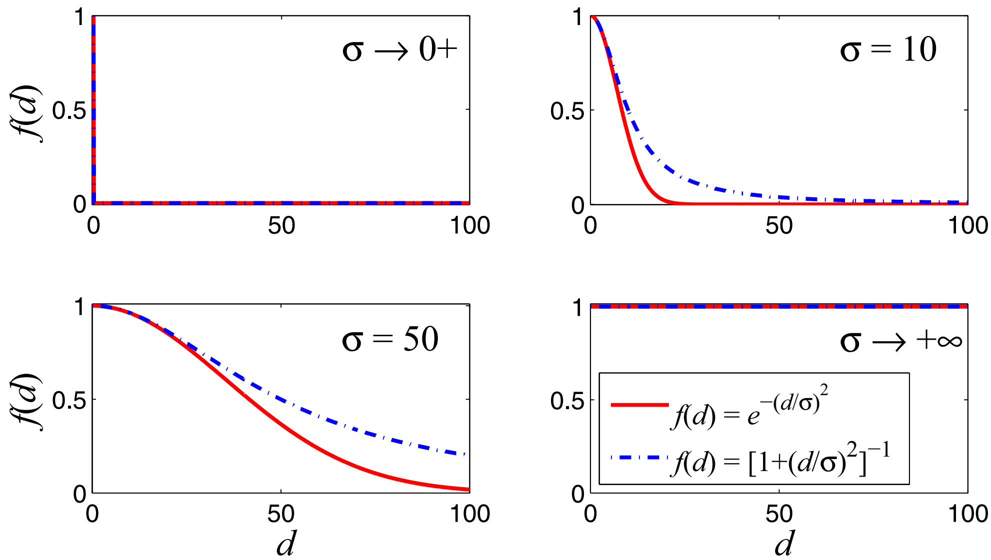


FIG. 1. (Color online) The plot of function $f(d) = e^{-(d/\sigma)^2}$ and $f(d) = [1 + (d/\sigma)^2]^{-1}$ at different values of σ.

$O[s(n^2 + rm)] + O[s^2(n + c^2)]$. Suppose $s,c,r \ll n$, and the network is sparse such that $O(m) = O(n)$. This complexity can be simplified to $O(sn^2)$.

IV. EXPERIMENTS

Modularity is widely used for detecting communities. In general, one takes modularity as an objective function and finds the "best" community partition by an optimization algorithm [16,25–31]. In this section, we use three examples to demonstrate that distance modularity is practically useful in detecting communities beyond assortativity-related attributes. We also compare distance modularity with NG modularity and spatial modularity. All results are obtained using the LPM+ optimization algorithm, because it can find higher modularity scores than the Louvain algorithm, without much additional running time [22].

A. Cerina's synthetic spatial networks

The first example is that of the synthetic spatial networks proposed by Cerina *et al.* [13] for testing whether a modularity can detect communities beyond the space attribute. Our scheme is as follows.

(1) We generate a set of synthetic spatial networks with known community structure (the true partition). The network topology is generated based on both space and community membership—two nodes which are spatially closer have a higher chance of getting connected, and two nodes which have the same community membership also have a higher chance of getting connected.

(2) We apply NG modularity, spatial modularity, and distance modularity to these networks to detect the communities beyond space.

(3) We compute NMI between the true partition and partitions by the three modularities. The higher the NMI score, the better the corresponding modularity.

The network generation procedure contains the following three steps.

(1) Generate 100 nodes in a (x,y) two-dimensional space. The first 50 nodes are around the north center (0,1) and fall in the north area $\{(x,y)| -1 < x < 1, 0 < y < 2\}$. The second 50 nodes are around the south center $(0,-1)$ and fall in the south area $\{(x,y)| -1 < x < 1, -2 < y < 0\}$. We generate the coordinates (x_i,y_i) of a node v_i according to probability $p_{\text{coord}}(x_i,y_i) \propto e^{(-d_{ic})}$, where d_{ic} is the Euclidean distance between v_i and its corresponding center.

(2) Arrange nodes into communities C_{+1} and C_{-1}. We assign the community membership l_i of node v_i as

$$l_i = \begin{cases} -\text{sgn}(y_i) & \text{with probability } \epsilon \\ +\text{sgn}(y_i) & \text{with probability } 1-\epsilon \end{cases}, \tag{18}$$

where sgn denotes the sign function and $\epsilon \in [0.1,0.5]$ is a parameter representing the correlation between space and community membership. In the case $\epsilon = 0.1$, space and community membership are highly correlated, such that 90% of the north nodes are assigned to community C_{+1} and 90% of the south nodes are assigned to community C_{-1}. In the case $\epsilon = 0.5$, space and community membership are totally uncorrelated, and nodes are assigned to either communities with probability 0.5.

(3) Generate links. We generate a link between v_i and v_j according to probability $p_{\text{link}}(v_i,v_j) \propto e^{\beta l_i l_j - d_{ij}}$, where d_{ij} is the Euclidean distance between v_i and v_j. We can see that $p_{\text{link}}(v_i,v_j)$ is positively related to $l_i l_j$ and d_{ij}. Thus, the existence of a link is affected by both space and community membership. Here $\beta \in [0.3,1.0]$ is a parameter determining space, or community membership, which has the leading effect in network topology. In the case $\beta = 0.3$, space has the leading effect and links are essentially between spatially close nodes. In the case $\beta = 1.0$, community membership has the leading effect and links are essentially between nodes of the same community.

By tuning parameters ϵ and β, we can create various cases reflecting the interplay between space and community membership. In particular, we illustrate four extreme cases of this series of networks in Table I and Fig. 2. These networks

TABLE I. The four extreme cases in Cerina's synthetic spatial networks.

	$\beta = 0.3$	$\beta = 1.0$
$\epsilon = 0.1$	(1) Space has the leading effect	(1) Community membership has the leading effect
	(2) Space and community membership are highly correlated	(2) Space and community membership are highly correlated
	(3) Please see Fig. 2(a) for visualization	(3) Please see Fig. 2(b) for visualization
$\epsilon = 0.5$	(1) Space has the leading effect	(1) Community membership has the leading effect
	(2) Space and community membership are uncorrelated	(2) Space and community membership are uncorrelated
	(3) Please see Fig. 2(c) for visualization	(3) Please see Fig. 2(d) for visualization

enable us to systematically study the performance of different modularities.

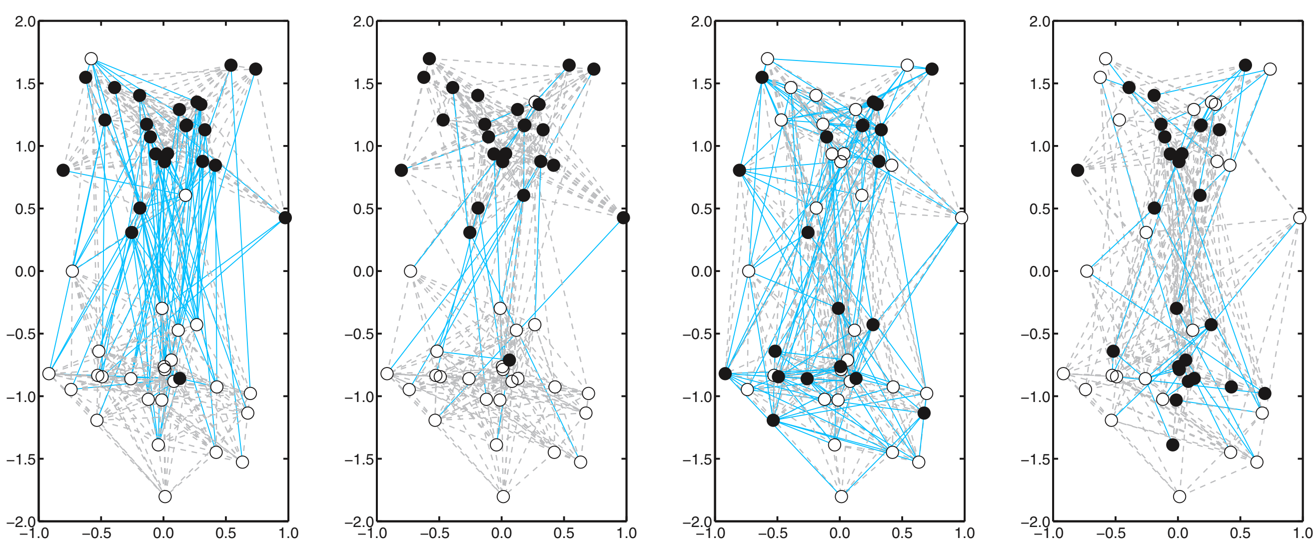
We adopted the following procedure to apply distance modularity. First, we used the Euclidean distance between (x_i, y_i) and (x_j, y_j) to compute d_{ij}. Second, considering that we did not know to what extent the network topology is affected by space, we specified $f(d) = e^{-(d/\sigma)^2}$, so that we can tune σ to simulate the space effect at different degrees. Third, suppose $\bar{d} = \sum_{i,j=1}^{n} d_{ij}/n^2$ is the average distance of all node pairs. We conducted the alternative parameter selection for $\sigma \in [0.1\bar{d}, 2.0\bar{d}]$, with a step length of $0.1\bar{d}$. Then, we reported the NMI score of the consensus partition. In addition, we reported the highest NMI score obtained in this σ interval.

As for spatial modularity, Expert *et al.* suggested a binning strategy to smoothen the probability function $p(d)$, which is expressed in Eq. (5) [15]. More precisely, they introduced a bin size parameter τ, which can influence the form of $p(d)$. For various possible values of τ, they conducted the alternative parameter selection. We followed their suggestions in applying spatial modularity and reported the NMI score of the consensus partition. We also reported the highest NMI score obtained by various values of τ.

Figure 3 shows the NMI scores by NG modularity, spatial modularity, and distance modularity, for $\beta \in [0.3, 1.0]$ and $\epsilon =$ 0.1, 0.3, and 0.5. Overall, the NMI scores follow an upward trend as β increases from 0.3 to 1.0. This is because as β increases the community membership has greater effect in network topology, and there are more links between nodes of the same community. As a result, it becomes easier to detect the communities.

By comparison, distance modularity performs the best, followed by spatial modularity. The reason for spatial modularity's inferiority is that it failed to accurately simulate the space effect in its null model. To see this, let us look back at Eq. (6), the foundation of the spatial modularity null model. This null model tries to simulate the space effect, so that the number of links between nodes at distance d is the same as that number in the observed network. However, since network topology is affected by both space and community membership, the pure effect caused by space itself is not what we observed in the network topology [32]. In particular, when there is a strong correlation between space and community membership, the effect of space is quite different from what we observed in the network topology. For this reason, spatial modularity does not perform well when $\epsilon = 0.1$.

On the other hand, NG modularity cannot detect the true communities with 100% accuracy, even when β is large. This is because NG modularity does not take attributes into account and thus cannot take out the space effect.



(a) (b) (c) (d)

FIG. 2. (Color online) Visualization of Cerina's synthetic spatial networks at the four extreme cases: (a) $\epsilon = 0.1$, $\beta = 0.3$; (b) $\epsilon = 0.1$, $\beta = 1.0$; (c) $\epsilon = 0.5$, $\beta = 0.3$; and (d) $\epsilon = 0.5$, $\beta = 1.0$. For the sake of clarity, only 50 out of the 100 nodes are displayed here. The nodes in community C_{+1} and C_{-1} are painted in black and white, respectively. The between-community and within-community links are plotted by solid and dashed lines, respectively.

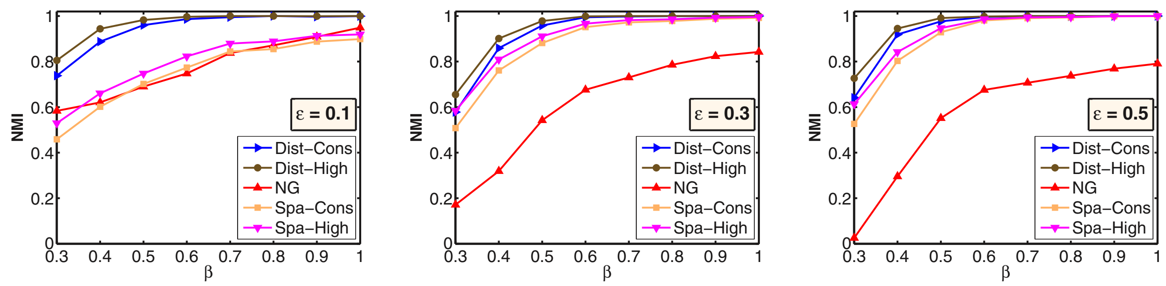


FIG. 3. (Color online) The NMI scores by NG modularity, spatial modularity, and distance modularity in Cerina's synthetic spatial networks, for $\beta \in [0.3,1.0]$ and $\epsilon = 0.1$, 0.3, and 0.5. Each score is based on an average of 100 implementations. NG stands for scores by NG modularity. Spa-High and Spa-Cons stand for the highest scores and the scores of the consensus partitions by spatial modularity, respectively. Distance-High and Distance-Cons stand for the highest scores and the scores of the consensus partitions by distance modularity, respectively.

Furthermore, we can find that the gap between the highest NMI scores and scores of the consensus partitions by distance modularity is not large. This indicates that the alternative parameter selection can help us find a good partition which is close to the best possible one. In real-world applications, we do not know the true partition. Thus, this gap also implies that there is still a potential room for improvement—if we can reduce the search interval of σ based on some background knowledge of the network, our results can be even better.

B. Lazega's partner advice network

The second example is based on a data set collected by Lazega [33] and Snijders *et al.* [34] on relations between partners in a New England law firm. From the data set, we constructed a symmetrized network, where nodes represent 36 partners in the firm and links represent 395 advisee-adviser relations (we ignore the direction of links). We weight a link by one if one partner has ever sought professional advice from the other and weight a link by two if both partners have sought advice from each other.

Moreover, various partners' attributes are also part of the data set. For example, we have information about age, gender, office location (Hartford, Providence, and Boston), and practice (litigation or corporate law) of each partner. In Fig. 4, we use red (dark gray), white, and blue (light gray) colors to differentiate partners in terms of office location, and we use square and round symbols to differentiate partners in terms of practice.

We can expect that the attributes of office location and practice have significant assortativity effects on network topology, because the advisee-adviser relationship is more likely between partners working in the same office and those with the same practice. Figure 4(a) shows the three-community partition by NG modularity. Note that this partition successfully separates red nodes (office = Hartford) from white and blue nodes (office = Providence and office = Boston), and it also separates square nodes (practice = litigation) from round nodes (practice = corporate law) to some degree. Thus, without taking attributes into account, NG modularity brings a compromise between the partition based on office location and the partition based on practice. An interesting problem is whether we can find the communities based on practice beyond the attribute of office location.

To solve this problem by distance modularity, we used the discrete distance between office locations of v_i and v_j

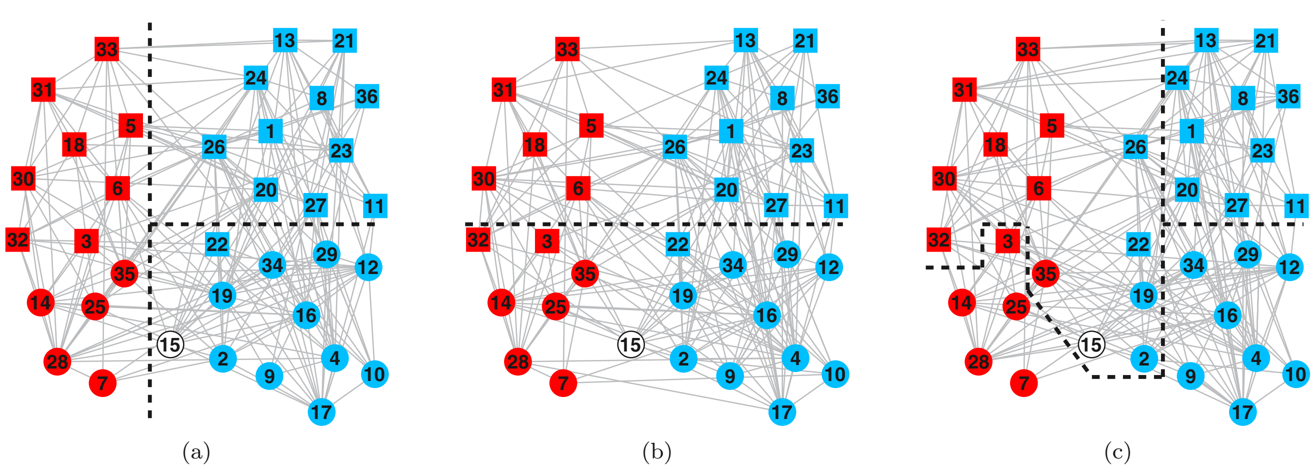
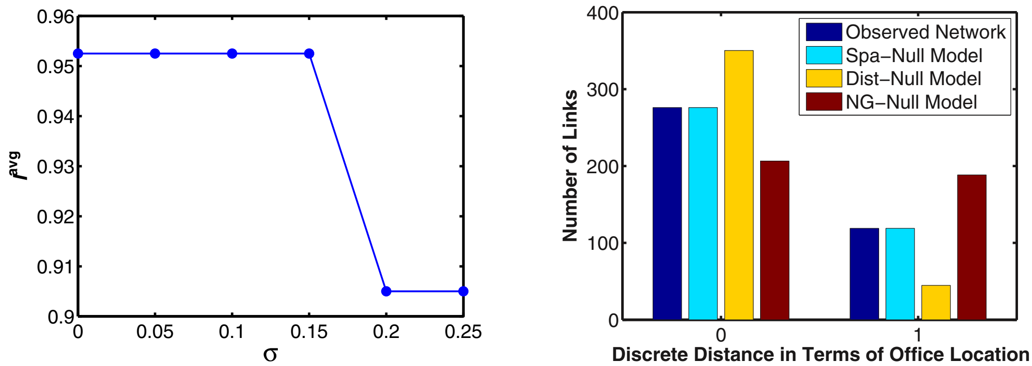


FIG. 4. (Color online) Visualization of Lazega's partner advice network. Partners whose practice is litigation and corporate law are symbolized as square and round nodes, respectively. Partners whose office location is Hartford, Providence, and Boston are painted in red (dark gray), white, and blue (light gray), respectively. (a) The partition by NG modularity. (b) The partition by distance modularity. (c) The partition by spatial modularity.



(a) (b)

FIG. 5. (Color online) Lazega's partner advice network. (a) The averaged normalized mutual information (I^{avg}) as a function of σ. (b) The effects of office location in the observed network, the spatial modularity null model, the distance modularity null model ($\sigma = 0.1$), and the NG modularity null model.

to compute d_{ij}. That is,

$$d_{ij} = \begin{cases} 0 & \text{if } v_i \text{ and } v_j \text{ has the same office location} \\ 1 & \text{otherwise} \end{cases}. \quad (19)$$

Considering that the link probability is much different for partners working in the same and different offices, we specified function f as

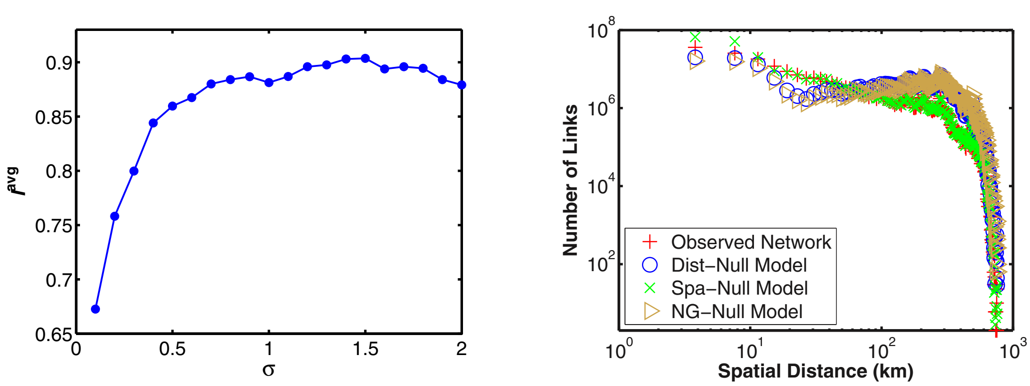
$$f(d) = \begin{cases} 1 & \text{if } d = 0 \\ \sigma & \text{otherwise} \end{cases}, \quad (20)$$

where σ is a parameter representing the probability of a link that exists between two nodes with different office locations in the distance modularity null model. Then, we conducted the alternative parameter selection for $\sigma \in [0, 0.25]$, with a step length of 0.05. The consensus partition was obtained at $\sigma \in [0, 0.15]$, as shown in Fig. 5(a). Figure 4(b) illustrates this two-community partition. We can find that it is almost the same as the partition based on practice, with only three nodes (nos. 3, 22, and 32) classified differently. Compared to the three-community partition by spatial modularity, as shown in Fig. 4(c), our partition is much closer to the partition based on practice.

Figure 5(b) shows the assortativity effects of office location in the observed network, the spatial modularity null model, the NG modularity null model, and the distance modularity null model. We can find that the spatial modularity null model assumes the same effect as what we observed in the network topology, while the distance modularity null model assumes an even greater effect. Which is correct? According to Pearson's chi-squared test, the p value for the null hypothesis that office location and practice are independent from each other is as high as 0.4848. Thus, under the pure effect of practice itself, links should not have a significant tendency to exist between nodes with the same office location. As a result, a reasonable explanation is that the actual effect of office location itself is greater than what we observed, and such effect is weakened due to the additional effect of practice. This can help us interpret why spatial modularity failed to detect the communities based on practice.

C. D4D antenna network

The third example is based on a data set of anonymous records of cellular phone calls between 5×10^6 Orange customers in the Ivory Coast between 1 December 2011 and 28 April 2012 (Orange is the key brand of France Telecom, one of the world's leading telecommunications operators). This data



(a) (b)

FIG. 6. (Color online) D4D antenna network. (a) The averaged normalized mutual information (I^{avg}) as a function of σ. (b) The space effect in the observed network, the spatial modularity null model, the distance modularity null model, and the NG modularity null model.

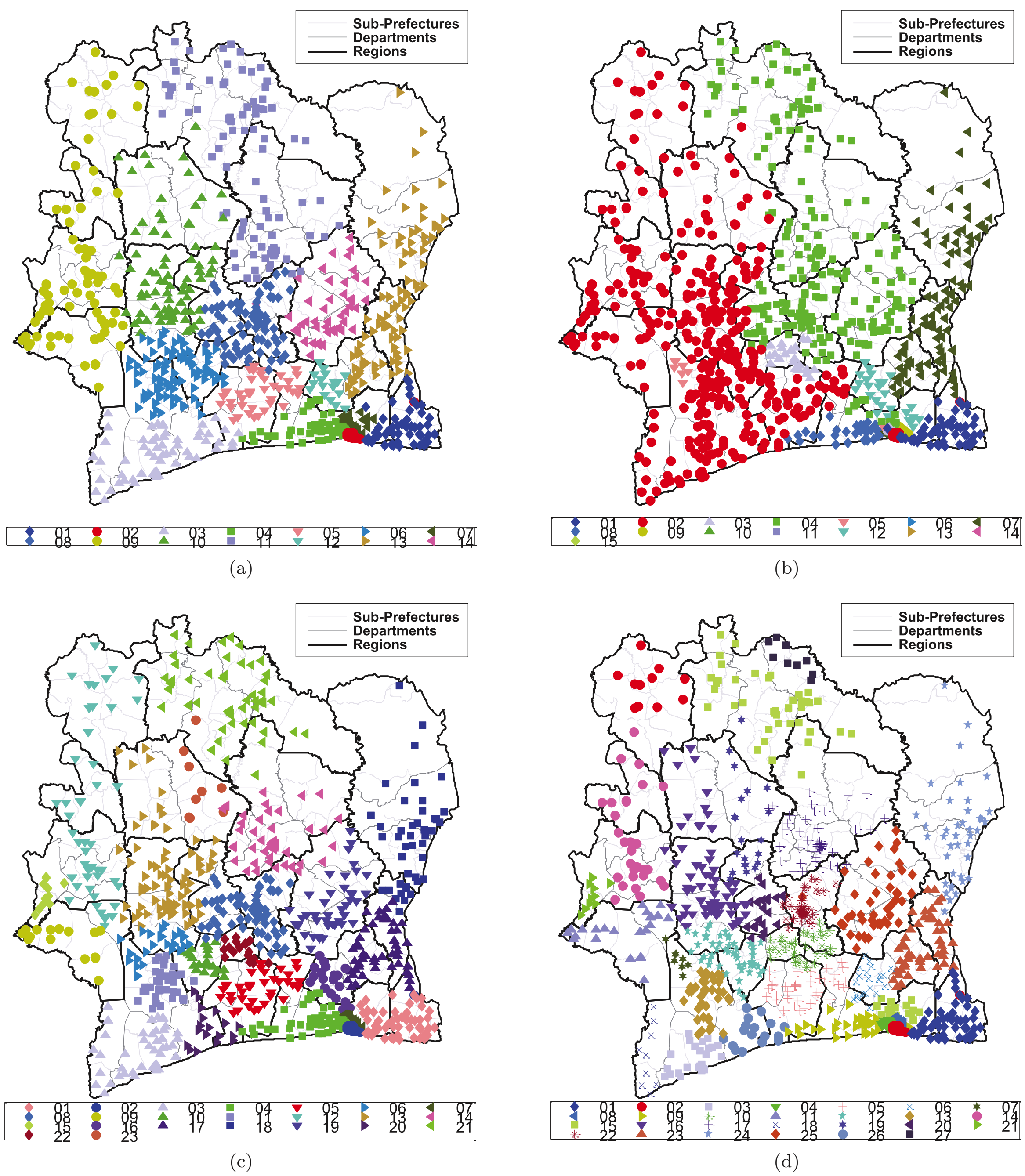


FIG. 7. (Color online) The partitions of the D4D antenna network by (a) NG modularity, (b) spatial modularity, (c) distance modularity at $\sigma = 1.5\bar{d}$ (the consensus partition based on the alternative parameter selection), and (d) distance modularity at $\sigma = 0.7\bar{d}$.

set was provided through the Data for Development (D4D) Challenge [35]. From the data set, we constructed an antenna-antenna network, which contains 1216 nodes representing cellular tower antennas of the country and 689 909 links representing communications between antennas, with weight indicating the number of calls. Additionally, we have coordinate information about geographical locations of antennas.

Many studies showed that space has an assortativity effect on network topology due to high cost associated with spatially distant links [36]. Thus, we aim to detect communities beyond space by distance modularity. The specific procedure is as follows. First, we used the great-circle distance to compute d_{ij}. Second, we specified $f(d) = [1 + (d/\sigma)^2]^{-1}$, since this reciprocal function is reminiscent of the gravity models which have long been used to model space related interactions [37,38]. Note that this reciprocal function allows us to simulate the assortativity effect of space at different degrees by tuning σ—the assortativity effect in the null model gradually fades as σ increases from zero to $+\infty$ (see Fig. 1). Third, we conducted the alternative parameter selection for $\sigma \in [0.1\bar{d}, 2.0\bar{d}]$, with a step length of $0.1\bar{d}$, and finally arrived at the consensus partition for $\sigma = 1.5\bar{d}$ [see Fig. 6(a)].

Figures 7(a)–7(c) visualize the partitions by NG modularity, spatial modularity, and distance modularity. We can find that

TABLE II. The NMI scores between the partitions by the three modularities and the partitions based on the three-level administrative divisions.

	Level 1	Level 2	Level 3
NG modularity	0.7269	0.7133	0.6586
Spatial modularity	0.4096	0.3970	0.4196
Distance modularity	0.7385	0.7556	0.7217

the communities by distance modularity and NG modularity are spatially compact. In contrast, the communities by spatial modularity are spatially wide. For example, community 02 in Fig. 7(b) covers almost the whole west part of the country. Figure 6(b) compares the space effect in the observed network, the spatial modularity null model, the NG modularity null model, and the distance modularity null model. We can see that the spatial modularity null model assumes a space effect slightly different from what we observed in the network topology [the difference is due to the binning strategy for smoothing $p(d)$]. On the other hand, the distance modularity null model assumes that the space has a milder assortativity effect. This can explain why distance modularity brings spatially more compact communities than spatial modularity.

As an example, in Fig. 7(d) we visualize the partition by distance modularity at $\sigma = 0.7\bar{d}$. A difference from the consensus partition at $\sigma = 1.5\bar{d}$ is that some communities such as community 15 are composed of several spatially distant groups of nodes which are themselves spatially compact. Note that the distance modularity null model at $\sigma = 0.7\bar{d}$ assumes a greater assortativity effect of space than that at $\sigma = 1.5\bar{d}$. Thus, bringing together some spatially distant groups would contribute to a higher score of Q^{dist} at $\sigma = 0.7\bar{d}$, and this results in the partition in Fig. 7(d). However, we can find that the key components of each community are spatially compact

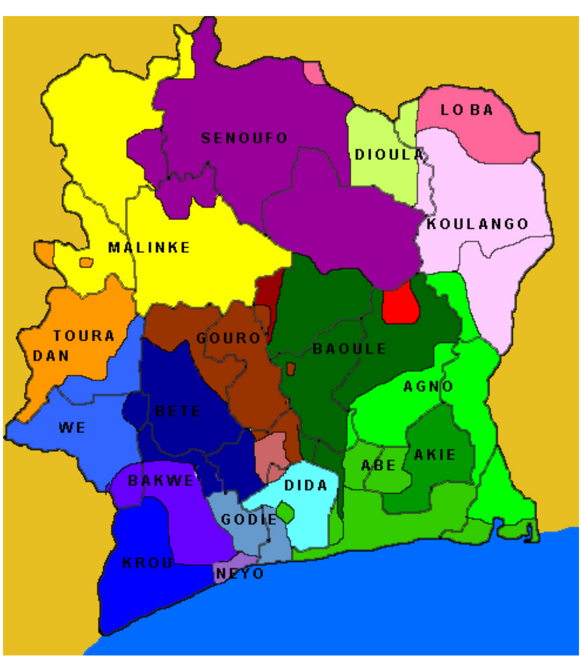


FIG. 8. (Color online) The ethnic distribution in the Ivory Coast (image reprinted from http://fr.wikipedia.org/).

nodes. This indicates that the most critical factors that drive the network are highly correlated with space. Therefore, our partition based on the alternative parameter selection is reasonable.

According to the 1998 Census, the Ivory Coast has 19 divisions at the region level, 50 at the department level, and 185 at the subprefecture level. The three-level administrative divisions are depicted by lines of different colors and widths in Fig. 7. It is interesting to find that the partition by distance modularity coincides with the three-level divisions to a great extent. Indeed, as listed in Table II, distance modularity has the highest NMI scores between its partition and the three-level divisions. Note that the administrative divisions are highly correlated with the ethnic distribution of the country, as shown in Fig. 8. Thus, the partition by distance modularity is a good predictor of ethnic groups.

V. CONCLUSION

In this paper, we focus on the problem of community detection beyond assortativity-related attributes ρ. A challenge of this problem is that we do not know to what extent the network topology is affected by ρ, and thus it is difficult to accurately simulate the effect of ρ in the null model. We proposed distance modularity, which allows us to freely choose a function f to simulate the effect of ρ. To apply distance modularity to the problem, the key points are to probe f using parameterized functions and conduct the alternative parameter selection to find a consensus partition. The success of our method lies in choosing the right form of function f and giving a good estimation of the parameter interval. Thus, having a background knowledge about the network under study would be of help. We used three examples to demonstrate the effectiveness of our method.

Our method has significant practical applications. In particular, detecting terrorist communities beyond space may assist in tracking higher-level organizations, such as a logistics group that provides support to the terrorist cells. Shakarian *et al.* (U.S. Military Academy) are working with the agencies in the U.S. Department of Defense and developing a software based on distance modularity for this emerging application [39].

One issue with our method is the scalability. State of the art community detection and graph partitioning techniques which consider only network topology may scale to more than 100×10^6 nodes [40–42]. However, our method, which considers both network topology and node attribute information, requires $O(sn^2)$ time complexity. This limits applications to small and medium-sized networks. How to speed up the computation by high-performance computing resources, such as multicores, graphics processing units, and clusters, is an important direction. This is left for our future work.

ACKNOWLEDGMENTS

The authors are grateful to Prof. Alessandro Chessa (University of Cagliari) for providing the code for generating the synthetic spatial networks, to Prof. Tom A. B. Snijders (University of Oxford) for providing the data set of partners' relationships in the New England law firm, and to France Telecom and Orange Cote d'Ivoire for providing the cell phone call data set. This work was partly supported by the Japan

Science and Technology Agency (JST), the Core Research of Evolutionary Science and Technology (CREST) research project, and the National Natural Science Foundation of China (Grant No. 61203154).

[1] A. L. Barabási, *Linked: The New Science of Networks* (Perseus, Cambridge, 2002).
[2] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[3] S. Fortunato, Physics Reports **486**, 75 (2010).
[4] M. E. J. Newman, Nature Physics **8**, 25 (2011).
[5] L. Danon, A. D.-Guilera, J. Duch, and A. Arenas, J. Stat. Mech. (2005) P09008.
[6] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
[7] G. K. Orman, V. Labatut, and H. Cherifi, J. Stat. Mech. (2012) P08001.
[8] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[9] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[10] M. McPherson, L. S.-Lovin, and J. M. Cook, Annual Review of Sociology **27**, 415 (2001).
[11] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, Nature (London) **489**, 537 (2012).
[12] F. Bavaud, Journal of Geographical Systems **15**, 233 (2013).
[13] F. Cerina, V. D. Leo, M. Barthelemy, and A. Chessa, PloS One **7**, e37507 (2012).
[14] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[15] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proc. Natl. Acad. Sci. USA **108**, 7663 (2011).
[16] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[17] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[18] M. Levandowsky and D. Winter, Nature (London) **234**, 34 (1971).
[19] S. Johnson, J. J. Torres, J. Marro, and M. A. Munoz, Phys. Rev. Lett. **104**, 108702 (2010).
[20] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikolski, and D. Wagner, Technical Report No. 2006-19, 2006 (unpublished).
[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. (2008) P10008.
[22] X. Liu and T. Murata, Physica A **389**, 1493 (2010).
[23] Note that there are many candidate functions. Choosing the right function form should depend on the background knowledge of the problem we are dealing with. Please see Sec. IV for more information.
[24] A. L. N. Fred and A. K. Jain, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003* (IEEE, Piscataway, NJ, 2003), pp. 128–133.
[25] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[26] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[27] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[28] A. Medus, G. Acuna, and C. O. Dorso, Physica A **358**, 593 (2005).
[29] K. Wakita and T. Tsurumi, in *Proceedings of the 16th International Conference on the World Wide Web, WWW'07* (ACM, New York, 2007), pp. 1275–1276.
[30] P. Schuetz and A. Caflisch, Phys. Rev. E **78**, 026112 (2008).
[31] M. J. Barber and J. W. Clark, Phys. Rev. E **80**, 026129 (2009).
[32] Actually, under the space effect itself, the number of links between nodes at distance d can be larger or smaller than that number in the observed network. This depends on two factors: (1) the proportion of effects by space and community membership and (2) the correlation between space and community membership.
[33] E. Lazega, *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership* (Oxford University, New York, 2001).
[34] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, Sociological Methodology **36**, 99 (2006).
[35] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, arXiv:1210.0137.
[36] M. Barthélemy, Physics Reports **499**, 1 (2011).
[37] G. A. Carrothers, Journal of the American Institute of Planners **22**, 94 (1956).
[38] A. G. Wilson, Transportation Research **1**, 253 (1967).
[39] P. Shakarian, P. Roos, D. Callahan, and C. Kirk, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'13, Chicago, IL, USA* (ACM, New York, 2013), pp. 1402–1409.
[40] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).
[41] Y. Dourisboure, F. Geraci, and M. Pellegrini, in *Proceedings of the 16th International Conference on the World Wide Web, 2007* (ACM, New York, 2007), pp. 461–470.
[42] J. Ugander and L. Backstrom, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 2013* (ACM, New York, 2013), pp. 507–516.