

Local multiresolution order in community detection

Peter Ronhovde

Department of Physical Sciences, The University of Findlay, 1000 N. Main St., Findlay, Ohio 45840, USA

Zohar Nussinov

*Department of Physics, Washington University in St. Louis,
Campus Box 1105, 1 Brookings Drive, St. Louis, Missouri 63130, USA*

(Dated: November 3, 2018)

Community detection algorithms attempt to find the best clusters of nodes in an arbitrary complex network. Multi-scale (“multiresolution”) community detection extends the problem to identify the best network scale(s) for these clusters. The latter task is generally accomplished by analyzing community stability simultaneously for all clusters in the network. In the current work, we extend this general approach to define local multiresolution methods, which enable the extraction of well-defined local communities even if the global community structure is vaguely defined in an average sense. Toward this end, we propose measures analogous to variation of information and normalized mutual information that are used to quantitatively identify the best resolution(s) at the community level based on correlations between clusters in independently-solved systems. We demonstrate our method on two constructed networks as well as a real network and draw inferences about local community strength. Our approach is independent of the applied community detection algorithm save for the inherent requirement that the method be able to identify communities across different network scales, with appropriate changes to account for how different resolutions are evaluated or defined in a particular community detection method. It should, in principle, easily adapt to alternative community comparison measures.

PACS numbers: 89.75.Fb, 64.60.aq, 89.65.-s

I. INTRODUCTION

Applications of complex network analysis span a wide range of seemingly unrelated fields. In these networks, elements of the model system are abstracted as nodes (*i.e.*, people, atoms, etc.), and edges represent known relationships between them (*i.e.*, friendships, energies, etc.). As depicted in Fig. 1, community detection (CD) [1, 2] seeks to identify natural groups of related nodes in a network. This structure can take the form of social groups [3], clusters of atoms [4], proteins [5], and much more. Several categories of common real-world networks are characterized in Ref. [3].

Conceptually speaking, communities in a network are groups of nodes that are strongly connected inside a community but weakly connected between communities. This basic idea is well established in the literature; it seems to be easily quantifiable and perhaps even sufficient to rigorously define a community if a few small clarifications are specified. However, the amazing variety of CD algorithms as well as a limited consensus in the field contradicts this naïve assessment.

Multiresolution community detection extends the CD concepts to find the most natural resolution(s) for a network partition. It endeavors to identify the network scales that best represent the community structure of a network, effectively distinguishing between densely or sparsely connected community members. Similar to single resolution CD, multiresolution methods must quantitatively assess this description in order to obtain an objective measure of the best candidate partition and

resolution. A common approach, which is implemented in various ways, is to search for regions with stable partitions [6–10], where the community structure does not change significantly in terms of various applied measures of the candidate partitions (e.g., number of communities q , dynamic flow across the network, information, etc.). Our global multiresolution algorithm [11] (MRA) asserts that the most natural resolution for a network may be identified based on how well independently solved replicas agree on the partition as evaluated by information measures [12, 13].

Our local multiresolution algorithm (LMRA) quantitatively identifies the most natural resolution(s) for individual communities regardless of the weak or strong community correlations present in the rest of the network. That is, the LMRA method is able to select optimal CD resolution parameter(s) independently for each cluster in a graph. Our use of the term local implies that the communities are determined with respect to parameters defined “near” the individual communities or nodes (*i.e.*, community size, relations among neighboring nodes, etc.). Here, we solve the full network partition for every resolution, but the algorithm trivially adapts to CD algorithms which can identify local communities without partitioning the entire network, which is important for immense networks such as the World Wide Web.

II. BACKGROUND

One of the most popular CD methods defines a cost function that attempts to quantitatively encapsulate the

essential features for a “good” division of nodes, thus evaluating the best community structure in an objective fashion. Regardless of the specific form, the task is to optimize the function for a particular graph to determine the optimal node division(s). Newman and Girvan [14] introduced the most common approach by far with “modularity.” CD methods based on Potts model cost functions, or methods that may be cast as such [15, 16], are also common. Reichardt and Bornholdt (RB) wrote a Potts model [17] which they specialized into two main cases utilizing null models. Null models are auxiliary graphs which are selected to evaluate the quality of a candidate partition, thus implicitly selecting the “correct” scale for a graph.

The choice of a null model inherently, often implicitly, selects a pre-determined scale for a network. The most common null models by far are: the “configuration null model” which sets edge connection probabilities based on the current graph, encompassing modularity as a special case, and the Erdős-Rényi null model [18] which defines the connection probability of all edges to be equally likely based on the graph’s average edge density. Optimization of CD quality functions using these null models was shown to suffer from an inherent resolution limit [14, 17, 19–21], which cannot be resolved by varying the network scale [22, 23]. This feature hinders the proper identification of some communities in large graphs.

An important general network model is the stochastic block model (SBM) [24], which provides a descriptive and generative model of network structure. Such models can then be used by various CD techniques to identify community structure [25–27]. Decelle et. al. [28, 29] and Hu et. al. [30, 31] studied phase transitions for SBM-type networks, and Darst et. al. examined related bounds on well-defined communities [32]. Extensions have moderated the internally homogenous nature of SBM graphs to improve its performance when modeling realistic networks with more varied degree distributions [33–35]. Often, network models imply or impose an expected structure, but an adaptive method based on mixture models [26] allows for detection of unspecified types of structure in a variety of network classes.

Potts model and related CD approaches include [11, 16, 36–41], and Refs. [39, 40] generalized the RB Potts models in [17, 18], respectively. Our previous work [11, 41] advanced a local Potts model, and local models were studied in more detail in [40]. Other local methods include [5, 15, 40–45], including variants of modularity [46, 47]. Potts systems in CD can experience disorder from thermal effects [31, 48], extraneous edges (noise) [31, 41, 48–50], and system size [31, 51]. The selected model can also exacerbate disorder effects [49, 52].

Some CD methods, such as modularity, implicitly select a single “objective” scale for a candidate community division (e.g., Refs. [14, 15]), but certain networks such as hierarchical systems inherently possess multiple natural scales. Hierarchical clustering is an early multiscale method [53], but it *forces* hierarchical structure

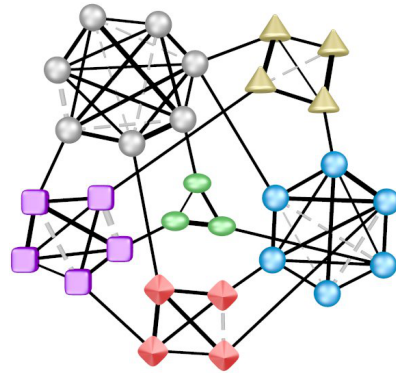


FIG. 1. (Color online) The figure illustrates a network partition where communities are represented by distinct node shapes and colors. “Friendly” or “cooperative” relations are depicted as solid, black lines which are modeled as ferromagnetic interactions with $w_{ij} > 0$ in Eq. (1). “Adversarial” or “neutral” relations (some are omitted for clarity) are depicted as gray, dashed lines. These are modeled as antiferromagnetic interactions with $u_{ij} > 0$. In both cases, the line thickness indicates a relative interaction strength. With Eq. (1), neutral interactions (unconnected or unspecified relations) are repulsive in nature since they work like adversarial relations that break up well-defined communities.

on every system without evaluating the relevance of the solved partitions. That is, it assigns but does not quantitatively evaluate whether the hierarchical structure is a good multi-scale partitioning scheme for the graph. More recent hierarchical approaches include [9, 54–58], and Ref. [59] relates the presence of hierarchical features to a scale-free-network property.

Ideally, a CD algorithm should be able to determine all relevant scales of a network. This problem is the impetus for developing quantitative multiresolution network analysis. Multiscale capable methods that utilize cost functions include [6, 8, 11, 17, 18, 43, 60]. The RB Potts model weighs the contribution of the null model [17], allowing the cost function to span different network scales. Other methods encompass varied forms of analysis [10, 61–63] to attack the problem.

Even with tunable CD cost function parameters, the question of which resolutions are the most *natural* scales for a network is not necessarily answered. Thus, multiresolution methods sought to identify the best scale(s) [6, 7, 11] for a network without imposing, or arbitrarily selecting, a preferred network scale. The most common method detects stable resolutions in terms of network and model resolution parameters [6, 11, 43]. Our multiresolution replica algorithm calculated information-based correlations [11] among independent copies of the same system to quantitatively compare the partition strength across all relevant network scales.

To our knowledge, all current multiresolution approaches analyze the network robustness in an “average” sense across all communities in a network (see Secs. VIA and VIB), but *the best local communities will not nec-*

essarily coincide at the same resolution in general. For example, communities in large networks may experience a “lost-in-a-crowd” effect which can obscure locally well-defined communities and limit the ability of global multiresolution methods (see Appendix A) to accurately isolate their structure. In some models, the effect can be exacerbated by heterogeneously-sized community structure [52, 64] depending on the network scale. Conversely, a global partition may be strongly defined for most communities, but a given cluster may still be weakly defined.

The LMRA method combines the benefits of multiresolution analysis with the local identification of community structure. While each community exists and is defined in the context of its own network neighborhood, we ideally prefer to identify strong communities independent of the global system. That is, we allow each community to stand on its own in terms of the evaluation of its community structure. Somewhat related efforts to the current work include detecting “unbalanced” communities in a network partition [65] and an efficient seed-expansion method by Havemann *et al.* [44] which could, in principle, be modified for other local cost functions.

Community detection methods utilizing quality functions generally include, or have been extended by subsequent work to include, weight parameters that serve vary the target resolution, so these methods will naturally adapt to the algorithm described in this work. Other disparate approaches include flow analysis [9, 10], spectral partitioning [66], Bayesian analysis [67], dynamics [68], network synchronization [69], k -means [70], and others.

Some of the above methods do not easily or naturally incorporate an explicit resolution parameter—relying rather on input community parameters, dynamics, or other measures of stability to identify the best resolution(s). Several of these require the number of communities q as an input parameter, or equivalently for our analysis, they may detect q based on the network (e.g., eigenvalue gaps). Then, q is passed to a clustering algorithm such as k -means. In either of these cases, q can serve as the resolution parameter for our LMRA algorithm, particularly for larger networks where small changes in q will represent correspondingly small changes in the overall community structure. For other cases, the precise implementation will be more model dependent, but the current LMRA algorithm only needs to receive the community partitions over a range of network scales, regardless of how these scales are detected or defined in a particular CD algorithm.

The remainder of the work is organized as follows: we introduce our community detection Potts model in Sec. III. Section IV A elaborates on concepts of community definitions, and Sec. IV B describes the notion of a partition resolution. We suggest a local, community-based analogy to the variation of information (VI) and normalized mutual information (NMI) measures in Sec. V which we apply in Sec. VI for our local multiresolution algorithm. Section VII illustrates the approach with three examples, and we conclude in Sec. VIII. Appendix A ex-

plains the context of local and global terminology used in this paper. Finally, Appendices B and C comment on the semi-metric property of our cluster measure as well as a couple alternative approaches to local cluster comparisons in an information-theoretic analogy.

III. POTTS MODEL HAMILTONIAN

Regardless of the underlying solution method, the ultimate goal of any community detection partitioning algorithm is a Potts-type assignment $i \rightarrow \sigma_i$ for each node i into one of q different clusters where σ_i may be regarded as a Potts-type variable. Toward this end, we focus directly on Potts variables. Some methods extend this notion to include overlapping memberships (e.g., Refs. [5, 43, 44, 71]) where nodes may be shared between, or fractionally assigned to, different communities. In these cases, the community assignment becomes a vector quantity for each node as opposed to a single integer value.

We identify community partitions by minimizing (see Sec. VIA) a general CD Potts model

$$\mathcal{H}(\{\sigma\}) = -\frac{1}{2} \sum_{i \neq j} [w_{ij} A_{ij} - \gamma u_{ij} (1 - A_{ij})] \delta(\sigma_i, \sigma_j) \quad (1)$$

which we refer to as an “absolute” Potts model (APM) since it is not defined relative to a random null model. Assuming N nodes, $\{A_{ij}\}$ is the adjacency matrix where $A_{ij} = 1$ if nodes i and j are connected and is 0 if they are not connected. The spin variable σ_i identifies the community membership of node i in the range $1 \leq \sigma_i \leq q$ where node i is a member of community k if $\sigma_i = k$. The Kronecker delta $\delta(\sigma_i, \sigma_j) = 1$ if $\sigma_i = \sigma_j$ and 0 when $\sigma_i \neq \sigma_j$. By virtue of the Kronecker delta, interactions are limited to spins in the same community, and they are ferromagnetic in nature if nodes i and j are connected and antiferromagnetic if they are not connected. The global resolution parameter γ scales the relative effects of the ferromagnetic $\{w_{ij}\}$ and antiferromagnetic $\{u_{ij}\}$ interactions, effectively allowing the model to vary the network resolution. A network resolution roughly corresponds to the typical community size, but a better characterization may be the typical community edge density (see Sec. IV B).

In Eq. (1), $\{w_{ij}\}$ and $\{u_{ij}\}$ are the edge weights for “cooperative” and “neutral” or “adversarial” relations, respectively, that are defined by the graph under consideration. These weights are based on known or estimated relations between the elements as recorded or defined by the person constructing the network. We refer to edges defined by $\{w_{ij}\}$ as cooperative since these lower the community energy (*i.e.*, they reinforce the community). Relations described by $\{u_{ij}\}$ raise the energy (*i.e.*, they work to break up the community). In unweighted graphs, $u_{ij} = w_{ij} = 1$.

Both adversarial and neutral relations serve to break up community structure, so the APM [11, 41] penalizes

neutral relations much like one would expect for adversarial relations (as opposed to zero energy contributions in a purely ferromagnetic Potts model [15, 36]). This property avoids a trivial ground state solution (*i.e.*, a completely collapsed system for every graph) present in the purely ferromagnetic Potts model. In essence, the energy penalty for adversarial relations provides a “penalty function” as an alternative to how modularity resolved the trivial-ground-state problem [14] (*i.e.*, by comparing a community to an average, random distribution of edges in the graph). Ref. [39] generalized a common Potts model variant [17] to include negative link weights.

Despite the global energy sum in Eq. (1), the model is a local measure of community structure (see Appendix A) because all node assignments are made strictly by evaluating local network parameters [40, 41]. For simplicity, our current analysis will focus on undirected, static networks; but both Eq. (1) and the LMRA method in this work are suitable for general weighted, directed, and dynamic (time-dependent) networks.

IV. COMMUNITY DETECTION CONCEPTS

A precise definition of community structure in networks is still not agreed upon in the literature. Generally speaking, communities consist of nodes which are strongly connected within communities, in terms of the number or weight of edges, but nodes in different communities are more sparsely connected. When constructing the quantitative community evaluation, there is also a question as to whether the “inside” versus “outside” degree comparison is summed over *all* external communities [72, 73] or is evaluated only between *individual pairs* of communities [11, 32, 41]. Our model applies the latter case.

A. Community definitions

Communities in social networks are the prototypical CD model. People often have many more “external” relationships of varying strengths than they do within a local group in which they are a member. For example, an individual may associate with a chess club, but his network of friendships may extend to dozens or even hundreds of people beyond this local group.

In many network approximations (e.g., the ubiquitous Zachary karate club network [74]), these “extra” edges are omitted as extraneous for the reduced-size network (*i.e.*, no need to solve a large graph if we are only interested in the local club). If we were to create a more comprehensive, expanded network and re-partition the system, the additional noise induced by including these previously external relations should not intuitively disturb the original communities, provided they are still strongly defined relative to any new structure(s) in the expanded system. This intuitive concept is overlooked by some CD

methods because the quantitative evaluation of community structure in the expanded system directly changes by virtue of the size (nodes, edges) increase alone, regardless of whether the local relations in and around a given community are affected.

Ref. [72] proposed definitions for “strong” and “weak” communities: in a strong community, *all* nodes have more internal than external edges, and a weak community is one where the *sum* over all internal edges exceeds the sum of the external edges. A large social network may not have “strong” or even “weak” communities in the sense of the proposed definitions, but the communities are still well-defined empirically. Thus, these community definitions [72] neglect certain important (high noise) and intuitive [21, 65] cases.

In fact, several CD methods were compared by Lancichinetti and Fortunato [75] where most, if not all, communities were weakly-defined in the sense proposed by Ref. [72], but many of the algorithms were nevertheless able to easily identify these purported weak communities. That is, the best methods easily solved the benchmark graphs [76] well into regions where *all* nodes (on average) have more external than internal edges. This definition is apparently not characteristic enough to describe weakly-defined communities based on the capabilities of some CD algorithms. Otherwise, we would intuitively think that the detection boundary would lie somewhere near this threshold. The crux is that the proposed definition considers the sum of external community edges, but identifiable communities in many CD algorithms seem to be more aligned toward a less restrictive definition of what a weak community is.

With these examples in mind, it seems appropriate, at least in social and related networks, to favor cost functions or analysis methods that utilize *pairwise* community comparisons when evaluating node membership robustness. This assumption inherently affects the notion of well-defined partitions, communities, and individual node memberships [32, 41]. Another approach that may be fruitful is to pursue a community definition based on *edge density* as opposed to inner and outer community *edge counts*, but further quantitative analysis is beyond the scope of the current work.

B. Resolution

Intuitively, the resolution of a community partition is the typical strength of intracommunity connections. This concept can be quantified by the typical edge density p of the communities in the partition. Communities with significantly different edge densities are qualitatively different. For example, social networks may naturally display communities of “close friends” or “acquaintances.” Close friends are generally very likely to know most or all members of the same group (p is high) where acquaintances are much less likely to know each other (p is lower).

As a specific example, a community where each person

has five friendships in a group of six is a clique. That is, every node is connected to all others in the group. However, if we consider the same five friendships in a group of 100, it may not even qualify as a community of social acquaintances. These two clusters have an identical edge count, but they represent drastically different *types* of communities (*i.e.*, different network scales). As mentioned above, the inner and outer edge count is not sufficient to quantitatively describe a cluster. This distinction highlights the importance of a penalty term in various CD quality functions.

In practice, a partition will contain communities with a range of edge densities, but intuitively, the differences should not be drastic at a given resolution since the partition should manifest communities with similar levels of association. Continuing with the social network example, mixing communities of close friends and acquaintances in the same partition makes less sense than a partition that indicates close friendships in most communities. Given this argument, it is reasonable that a given γ in Eq. (1) could be applied to the whole graph and provide meaningful partition information in general, but this manuscript illustrates a method to enhance the analysis of complex networks by finding locally optimal resolutions at the community level.

We specialize the edge density analysis below to unweighted graphs for clarity, but Ref. [41] discusses weighted graphs in the same context. The edge density of community a is $p_a = \ell_a / \ell_a^{\max}$ where ℓ_a is the number of edges in the community; $\ell_a^{\max} = n_a(n_a - 1)/2$ is the maximum number of possible edges in community a with n_a nodes. The global resolution parameter γ in Eq. (1) requires a *minimum* edge density for each community in the partition,

$$p_{\min} \geq \frac{\gamma}{\gamma + 1}, \quad (2)$$

which we calculated by determining the minimum density configuration that yields an energy of zero or less. Without γ , the model can only solve a particular implicit resolution for all systems, $p_{\min}^{\gamma=1} \geq 1/2$. Other models implement similar weight parameters [6, 16–18, 39, 40, 43] which allow them to solve distinct network scales.

While Eq. (2) provides a convenient lower bound on the minimum community edge density, optimizing Eq. (1) implements the constraint implied in Eq. (2) by enforcing a stronger requirement. That is, it merges network elements (a node to a community or two communities) if the edge density between them exceeds p_{\min} . Thus, one is assured that *all sub-elements of a community are connected by at least p_{\min}* . This effectively avoids resolution-limit-type effects by acting locally [41].

C. Heuristic multiresolution method

The local and global multiresolution methods are discussed in detail later, but it is relevant to take a mo-

ment consider the basic function of our multiresolution approach and its connection to the underlying CD solver. Briefly, the global method independently solves for the community structure in a set of r replicas of the network. It then uses information-based measures (see Sec. V) to evaluate the partition similarities, asserting that the best resolutions have strongly correlated partitions among the replicas.

The exact partitions determined among the replicas depend on the efficacy of the particular CD algorithm used to determine the individual partition solutions. While Eq. (1) has a well-defined ground state of partitioned nodes that depends on the weight parameter γ , detecting transitions with the MRA global multiresolution method still depends in some sense on imperfect CD solutions provided by our robust, but nevertheless greedy, CD algorithm.

More specifically, the MRA algorithm takes advantage of the fact it is significantly more difficult for a heuristic CD algorithm to navigate the energy landscape when competing partitions have comparable energies. This condition could occur when two or more good partitions exist at a single resolution, but it appears to occur more often in transition regions between different levels or types of structure. In terms of CD model parameters, these different levels of multi-scale community structure have a minimum energy at different resolutions; but in the transition region, the minimum for each distinct division is comparable, causing different replicas of the CD algorithm to be more easily trapped in unrelated areas of the energy landscape. The MRA algorithm attempts to detect this variation in partitions and uses it to identify the best resolutions.

If we consider a perfect CD algorithm which always finds the ground state of the cost function regardless of any model parameters or the complexity of the problem, most of the variation among partition solutions in the replicas mentioned above will disappear (see comment on degeneracies below). That is, the more perfect the solver, the sharper the detected transition would become in terms of any variation in the CD model's resolution parameter(s). With a perfect solver or an easy CD problem, the transition is essentially discrete.

The primary partition similarity measures of our MRA algorithm, variation of information V and the normalized mutual information U , would only observe strong agreement among the replicas across this transition (*i.e.*, no observable extrema indicating structural shifts or preferred community structure), but changes in the secondary measures (e.g., Shannon entropy H , mutual information I , and the number of clusters q) would still mark the transition, albeit without quantitatively indicating the quality of the candidate partitions (the length of the stable region can still be a qualitative indicator of partition stability). Thus, a perfect CD algorithm illustrates that the MRA algorithm, as currently employed, relies to some extent on the imperfect solutions provided by the underlying CD solver used. A similar argument would

apply to the local multiresolution method discussed in the current work.

When using a perfect CD solver, our global multiresolution method would measure the degeneracy of the ground state energy at a given resolution. One could extend our global multiresolution method by comparing partitions at nearby resolutions to better evaluate the stability of the partitions over a range of resolutions. Under this construction, the MRA algorithm would still be able to evaluate partition stability even with a perfect CD solver at the cost of increased computational effort based on time spent comparing replicas across a range of resolution values.

V. INFORMATION MEASURES

Information measures have received broad acceptance for comparing candidate CD partitions. Commonly used measures include the variation of information [13] and normalized mutual information [12]. We leveraged the measures in Sec. V A to identify the best global network scales via a multiresolution replica method [11] (see Appendix A and Sec. VIB).

A. Partition correlations

To define VI and NMI, we select a random node from partition A and note that it has a probability n_k/N of being in community k where n_k is the number of nodes in the community and N is the total number of nodes in the system. The Shannon entropy is

$$H(A) = - \sum_{a=1}^{q_A} \frac{n_a}{N} \log \frac{n_a}{N} \quad (3)$$

where q_A is the number of communities in partition A . The mutual information $I(A, B)$ between two partitions A and B evaluates how much we learn about A if we know B . For our application, contending partitions (A, B, \dots, Z) are independently solved copies of the system.

We define a “confusion matrix” for partitions A and B which specifies how many nodes n_{ab} in community a of partition A are also in community b of partition B . Mutual information is

$$I(A, B) = \sum_{a=1}^{q_A} \sum_{b=1}^{q_B} \frac{n_{ab}}{N} \log \left(\frac{n_{ab}N}{n_a n_b} \right) \quad (4)$$

where n_a (n_b) is the number of nodes in community a (b) of partition A (B). Also, $I(A, B) = 0$ when $n_{ab} = 0$. The variation of information $V(A, B)$ metric is then

$$V(A, B) = H(A) + H(B) - 2I(A, B) \quad (5)$$

which measures the information “distance” between partitions A and B with a range of $0 \leq V(A, B) \leq \log N$. We use base 2 logarithms.

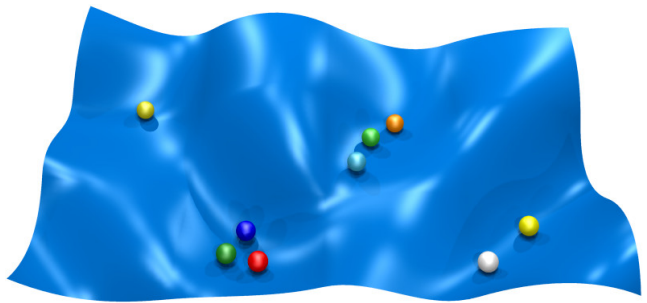


FIG. 2. (Color online) The figure schematically depicts r independent solvers (“replicas”) as spheres navigating the energy landscape of Eq. (1). Stronger agreement among the replicas, as measured by information correlations in Sec. V A, indicates a more stable, well-defined global partition (see text). In this manuscript, we demonstrate that local communities may be strongly defined even if all the communities in the global system are weakly correlated (see Fig. 3).

A normalized information measure [12] of partition similarity is

$$U(A, B) = \frac{2I(A, B)}{H(A) + H(B)}. \quad (6)$$

NMI and VI are closely related, $U(A, B) = 1 - V(A, B)/[H(A) + H(B)]$. While NMI is a valuable measure of partition similarity, it is not a formal metric (see Appendix B) on partitions A and B in part because $U(A, A) = 1$ not 0.

Some researchers prefer NMI as a normalized measure where it maps uncorrelated partitions to 0 and perfectly correlated partitions to 1. Others prefer the stricter metric properties of VI, but VI can also be normalized, if desired. In most cases, the two measures provide the same information, but occasionally distinctions between them can be observed. For example, we previously showed [41] that maxima in VI and minima in NMI mark structural transitions in partitions, that provides information about the average intercommunity edge density, but only VI clearly shows an extremum before the system collapses into disjoint communities as γ is lowered in Eq. (1).

B. Local information analogies

When defining a cluster comparison measure, we wish to maintain consistency with the trend in CD towards information-theoretic partition evaluations. Toward this end, we consider the cluster embedded in the full system of N nodes (see also later comments), where N is the number of nodes in the network. This comparison gives our measure a context for the resulting cluster-level entropy or information calculations based on associated partition-of-unity probabilities.

From Eq. (3), the entropy contribution of community

a in partition A is

$$H_a(A) \equiv -\frac{n_a}{N} \log \left(\frac{n_a}{N} \right) \quad (7)$$

where n_a is the number of nodes in community a . Similarly, Eq. (4) indicates the mutual information contribution when comparing cluster a in partition A , (a, A) , to cluster b in partition B , (b, B) ,

$$I_{ab}(A, B) \equiv \frac{n_{ab}}{N} \log \left(\frac{n_{ab}N}{n_a n_b} \right). \quad (8)$$

In analogy with Eq. (5), we introduce the *cluster variation of information* (CVI) $v(a, b)$

$$v(a, b) \equiv H_a(A) + H_b(B) - 2u(a, b). \quad (9)$$

CVI exhibits appealing “distance-like” properties of a semi-metric for comparing clusters (a, A) and (b, B) (see Appendix B for a trivial proof). Summing over all pairs of clusters a and b , VI is related to CVI by

$$V(A, B) = \sum_a^{q_A} \sum_b^{q_B} v(a, b) - (q_B - 1)H(A) - (q_A - 1)H(B). \quad (10)$$

Appendix C provides additional remarks.

From Eq. (6), we introduce the natural *cluster normalized mutual information* (CNMI) analogy

$$u(a, b) \equiv \frac{2n_{ab} \log \left(\frac{n_{ab}N}{n_a n_b} \right)}{n_a \log \left(\frac{N}{n_a} \right) + n_b \log \left(\frac{N}{n_b} \right)}. \quad (11)$$

While CNMI is not a metric [in part because $u(a, a) = 1$ not 0], it has the same intuitive property of cluster similarity that makes NMI attractive for partition comparisons. Equation (11) is essentially a normalized variant of CVI, $u(a, b) = 1 - v(a, b) / [H_a + H_b]$. On smaller networks, CVI provides a clearer picture of transitions with its distance-like semi-metric properties, but CNMI is more easily evaluated for larger networks because variations in CVI become small as N becomes large.

If we were to compare larger (multi-cluster) subgraphs, a natural approach is to cut the subgraph from the whole network and compare the reduced-size partitions. This breaks down at the cluster level because there is no partition-of-unity associated with an individual cluster as used to define NMI [12] or VI [13] for community detection. Implementing an arbitrary measure for clusters is difficult, so we chose to consider the cluster comparisons in the context of a larger network of nodes. Strictly speaking we do not need to use the true size of the network for our cluster comparisons. Rather, we could use some other $N' \neq N$, but it is conceptually appealing to evaluate a cluster in the context of the full network.

VI. LOCAL MULTIREOLUTION ALGORITHM

Our local multiresolution algorithm identifies relevant local multiresolution order, meaning well-defined local communities. We invoke $v(a, b)$ in Eq. (9) and $u(a, b)$ in Eq. (11) to compare local clusters a and b across r “replicas” (independent solutions). Figure 2 depicts solutions with the global MRA [11] algorithm given in Sec. VIB. The LMRA method depicted in Fig. 3 extends the MRA method by incorporating comparisons between specific clusters.

A. Community detection algorithm

We begin by introducing our greedy CD algorithm which dynamically moves nodes into the community that best lowers the local energy according to Eq. (1) given the current state of the system $\{\sigma_i\}$. The process iterates through the nodes until no further nodes are available. Typically, $O(10)$ iteration cycles through all N nodes are required except in rare instances that lie in or near the “hard” (or “glassy”) phase [31, 41, 48].

The CD steps are:

(0) *Initialize the system.* Initialize the connection matrix A_{ij} and edge weights w_{ij} and u_{ij} . Determine the number of optimization trials t .

(1) *Initialize the clusters.* The initial partition is usually a “symmetric” state wherein each node is the lone member of its own community (*i.e.*, $q_0 = N$).

(2) *Optimize the node memberships.* Sequentially select each node, traverse its neighbor list, and calculate the energy change that would result if it were moved into each connected cluster (or an empty cluster). Immediately move it to the community which best lowers the energy (optionally allowing zero energy changes).

(3) *Iterate until convergence.* Repeat step (2) until a (perhaps local) energy minimum is reached where no nodes can move.

(4) *Test for a local energy minimum.* Merge any connected communities if the combination lowers the summed community energies. If any merges occur, return to step (2) and attempt additional node-level refinements.

(5) *Repeat for several trials.* Repeat steps (1)–(4) for t independent trials and select the lowest energy result as the best solution. By a trial, we refer to a copy of the network in which the initial system is randomized in a symmetric state with a different node order.

The optimal q is usually dynamically determined by the lowest energy state although the algorithm can also fix q during the dynamics. Empirically, the computational effort scales as $O(tL^{1.3} \log k)$ where k is the average node degree and $\log k$ is from a binary search implemented on large sparse matrix systems. This greedy variant can accurately scale to at least $O(10^9)$ edges [41].

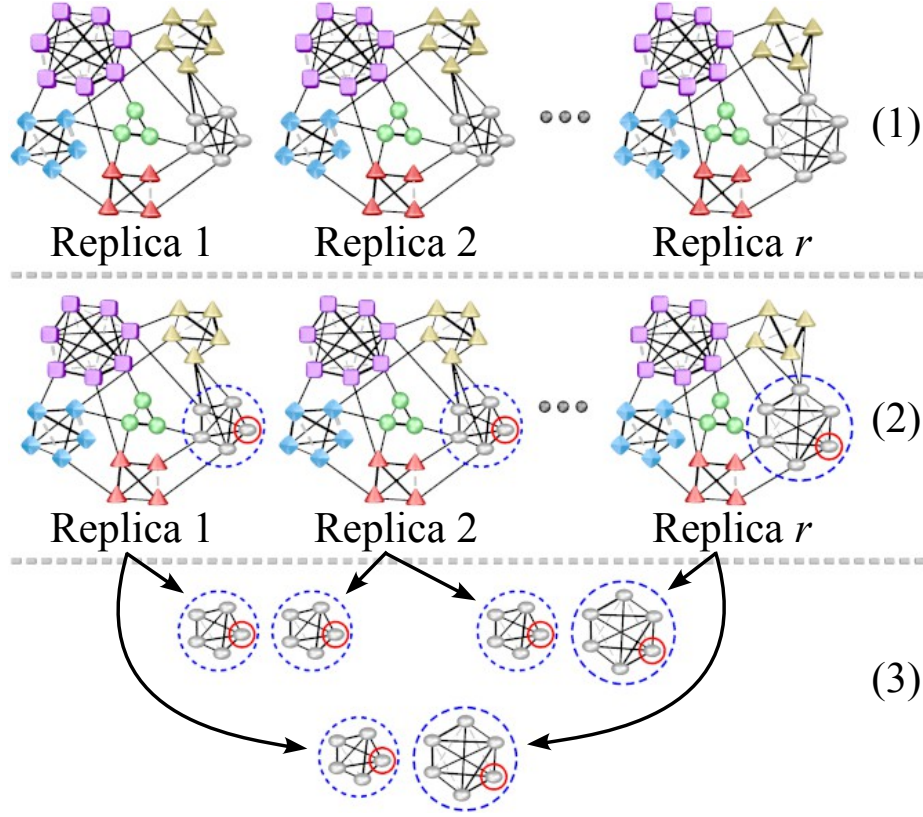


FIG. 3. (Color online) The figure illustrates our local multiresolution algorithm discussed in detail in Sec. VI. The graphs include ferromagnetic [“cooperative” with $w_{ij} > 0$ in Eq. (1)] relations depicted by solid, black lines and antiferromagnetic (“neutral” or “adversarial” with $u_{ij} > 0$) interactions depicted by gray, dashed lines. The line thickness indicates the relative interaction strength, and we omit intercommunity adversarial and neutral relations for clarity. The following steps are completed for each γ used in Eq. (1): In step (1), we independently solve a series of r “replicas” of the community detection problem (although we could, in general, improve the efficiency by solving only the local communities embedded in the network). Step (2) identifies the target node(s) of interest (solid red circles) and their corresponding parent clusters (blue dashed circles). Step (3) uses Eqs. (9) and (11) to calculate correlations among all pairs of parent clusters in order to determine the community robustness at the current resolution specified by γ in Eq. (1).

We can extend it with a stochastic heat bath [48] solver or a simulated annealing algorithm [17] at the cost of significantly increased computational effort, but the greedy variant performs exceptionally well on many systems.

B. Global multiresolution algorithm

In order to set the stage for introducing our local multiresolution algorithm, we first discuss the global algorithm and some of its features. As depicted in Fig. 2, our multiresolution algorithm iteratively applies the CD algorithm in Sec. VI A to quantitatively evaluate the most stable community partitions over a range of network scales. After convergence, these replicas sample the local energy minima of the energy landscape, giving an estimate of its associated complexity.

In its basic form, the global MRA algorithm iteratively solves the CD problem for a graph over a range of γ in Eq. (1) and evaluates the average strength of the parti-

tion correlations to find more stable partitions. This process quantitatively estimates the robustness of the identified partitions by sampling the complexity of the energy landscape. Previous work by the authors [11] on the Lancichinetti-Fortunato-Kertész [76] as well as other synthetic and real benchmarks [41] show that these strongly correlated regions regularly correspond to the known, accurate partitions.

Generally speaking, poorer correlations occur when there are contending partitions of comparable strength [*i.e.*, the energy difference of the applied cost function is near zero], the resolution is inside a glassy phase (extraneous intercommunity edges obscure the dynamic process of locating the best solution), or the graph is more random in nature. In the case of contending partitions, local multiresolution methods, such as the one presented in the current work, may be able to reliably extract the well-defined communities.

We quantify the partition correlations using information theoretic (or other appropriate) measures (see Sec.

V A). If most or all solvers (replicas) agree on the best solution, then we rate the partition as strongly correlated, but if the partitions have large variations, we say the solution is weak. In either case, we select the lowest energy replica solution to represent the best answer at a given resolution γ_i , but one could also construct a consensus partition [15, 77, 78], particularly in the latter case of weak solutions [79].

As a function of the resolution parameter γ in Eq. (1) (or any relevant CD scale parameter for another model [6, 17]), the best resolutions may be identified by peaks or plateaus in NMI [11], minima or plateaus in VI [8, 11], and/or plateaus in the number of clusters q [6] or other measures [8, 11]. Plateaus in these measures (*i.e.*, NMI, VI, H , q , etc.) as a function of γ imply more stable features of the network, although caution must be exercised when interpreting some measures [11]. Sharper peaks in NMI or narrow troughs in VI indicate strongly defined but more transient features. Significant peaks in VI or troughs in NMI generally indicate transitions between dominant structures. More generally, we can further extract pertinent details of the network from other extrema in NMI and VI (e.g., Ref. [80] also analyzed peaks in VI to perform image segmentation using CD concepts).

Correlations among the replicas evaluate the level of agreement on stable, low-energy solutions. If a problem has two equally viable partitions (*i.e.*, with the same energy) that are located by two replicas, the one with the highest overlap among all other replicas would be preferred. This corresponds to the volume of a configuration space basin associated with this preferred partition where the other partition with the same energy has an associated smaller basin size (the number of states or exponentiated entropy). This is like an “order by disorder” effect [81–84] present in various systems (e.g., entropic contributions to the free energy in finite temperature systems) which lifts the degeneracy between equally viable partitions and favors one partition (or a subset of partitions) over the others.

The MRA algorithm is:

(0) *Initialize the algorithm.* Select the number of independent replicas r . Identify the set of resolutions $\{\gamma_i\}$ to analyze using Eq. (1) along with a starting γ_0 . It is often convenient to begin at high gamma and step downward, stopping if the system completely collapses.

(1) *Initialize the system.* For the current γ_i , initialize each replica with a unique set of N spin indices (*i.e.*, $q_0^{(j)} = N$ for each replica j).

(2) *Solve each replica.* Independently solve each replica according to the CD algorithm in Sec. VIA.

(3) *Compare all replicas.* Calculate the Shannon entropy for every replica and compare all pairs of replicas using the mutual information $I(A, B)$, normalized mutual information $U(A, B)$, and variation of information $V(A, B)$ measures in Sec. VA.

(4) *Iterate to the next resolution.* Increment to the next resolution γ_{i+1} . A geometric step size $\Delta\gamma = 10^{1/s}$

is often convenient where $s \approx O(10)$ is an integer number of γ_i ’s per decade of γ . Repeat steps (1)–(3) until the system is fully collapsed (if stepping down in γ_i) or no γ_i ’s remain.

The information correlations in steps (3) and (4) allow the determination of the best global network scale(s) [11] (see Appendix A) based upon regions of γ with high NMI or low VI. Plateaus in H , I , and q may also provide supplemental information regarding partition stability. The solution cost scales linearly in r with the CD algorithm in `secrapp:CDalgorithm`, $O(rtL^{1.3} \log k)$. We have solved systems with $O(10^7)$ edges on a single processor [11] in a few hours.

The algorithm may detect, but does not impose, a hierarchical community structure. That is, as shown in Sec. VII A, the MRA algorithm will show strongly correlated regions at the well-defined hierarchical levels, but it is also able to analyze non-hierarchical multiresolution structure. This approach is preferable to *forcing* a hierarchical structure on every analyzed network [53] since some networks may not naturally possess this type of organization. Once the preferred resolutions are identified, the specific hierarchical nature can be analyzed and evaluated by other means [85, 86].

C. LMRA replica method

Clusters naturally change as the resolution is varied, so how do we identify the appropriate target clusters for comparison? Two immediate approaches include: (i) compare clusters for “nearby” resolutions as specified by a particular γ in Eq. (1) or (ii) compare targeted parent clusters for specific node(s) of interest across the replicas. Another natural approach is to mix the above methods: (iii) select a node of interest but further compare the target clusters at neighboring resolutions. This is an extension of case (ii), so we focus on former cases, leaving more complicated implementations to future work.

In case (i), if one deviates too far from γ_i , the cluster will change substantially and the evaluation will be less useful. That is, at some point, the cluster changes enough that it is no longer the “same” community. We could quantitatively define this comparison based on the relevant CVI values, but the practical question of identifying the appropriate community for comparison across all partitions becomes increasingly difficult.

In case (ii), the node may be selected *a priori* based on a known identity in the real network, or it may be randomly selected. (One may also first analyze the global system and use any communities with interesting features to identify important nodes.) This option has two advantages: it is simpler to implement, but more importantly, the studied clusters are always well-defined, enabling comparisons of community robustness across all relevant resolutions. That is, at a given γ_i , we only need to know to which cluster node i belongs, regardless of

any structural changes in its network neighborhood as γ is varied. Cluster correlations are quantitatively evaluated at a given γ_i , but the average $\bar{v}(a, b)$ or $\bar{u}(a, b)$ values over the replica pairs can still be compared across different γ_i 's to evaluate the relative strength of the parent communities.

Option (ii) is used in the current work. We select a *node* of interest (e.g., a person in a terrorist network, see Sec. VIIB), and trace the parent clusters among the replicas across a range of network scales [*i.e.*, different γ_i 's in Eq. (1)]. As depicted in Fig. 3, the LMRA algorithm is:

(0) *Initialize the algorithm.* Select the number of replicas r and the number of independent optimization trials t per replica (see Secs. VIA and VIB). Select a set of nodes $\{\eta\}$ to track based on problem parameters (e.g., a person of interest). Identify the set of resolutions $\{\gamma_i\}$ to analyze (often selected to sample all relevant network scales, see step 4 in Sec. VIB) by minimizing Eq. (1). Select a starting γ_0 .

(1) *Solve r independent replicas.* For the current γ_i in Eq. (1), apply steps (1)–(3) of the global MRA algorithm in Sec. VIB.

(2) *Identify parent clusters.* Identify the parent cluster a_{ij} corresponding to each target node η at the current γ_i in each replica j .

(3) *Compare clusters.* For each parent cluster a_{ij} , calculate CVI $v(a_{ij}, a_{ik})$ in Eq. (9) and CNMI $u(a_{ij}, a_{ik})$ in Eq. (11) with the corresponding parent cluster a_{ik} in replica k . Calculate the average of measure S_i [generically referring to measures $v(a, b)$, $u(a, b)$, etc. in Sec. VB] over all replica pairs at γ_i by

$$\bar{S}_i(a, b) = \frac{2}{r(r-1)} \sum_{k>j} S_{ijk}(a, b), \quad (12)$$

where i refers to a particular resolution parameter index for γ_i in Eq. (1), and j and k refer to replica summations.

(4) *Identify the best resolutions.* For each parent cluster a_{ij} , find the lowest CVI values $v(a, b)$ or the highest CNMI values $u(a, b)$ and their corresponding resolution(s) $\{\gamma_i^{\text{Best}}\} \subset \{\gamma_i\}$. These are the best resolutions for each cluster a_{ij} .

Step (4) is the final result for the algorithm, indicating which resolutions and candidate partitions are mostly likely to be useful. As with the global MRA approach in Sec. VIB, we are interested in extrema or plateaus in the pertinent measures in Sec. V. Empirically, $r \approx O(10)$ or less appears to be sufficient for most problems. We estimate the cost to be $O(Lr^2)$ which is comparable to the base MRA algorithm cost in Sec. VIB.

D. Alternative implementations

In the current work, we contrast local, community-level analysis with global multiresolution correlations. Thus,

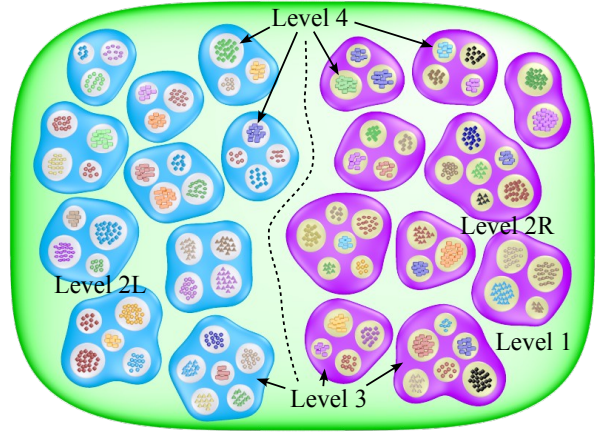


FIG. 4. (Color online) The figure depicts a constructed $N = 1024$ node four-level hierarchy. Level 1 is the complete network with two sides of supercommunities that are randomly connected at a low edge density between them. Level 2 consists of two roughly equal sized branches ($N_L = 502$ and $N_R = 522$) which we denote by left (L, blue or darker tone) and right (R, purple or medium tone) as the figure indicates. Level 3 is the set of supercommunities, and level 4 is the set of smallest communities strictly contained within the supercommunities. At levels 3 and 4, elements of the left branch are connected at higher internal and intercommunity edge densities than the corresponding right branch elements. See the text for a more detailed description of the network. This construction results in a more “blurred” global multiresolution signature in Fig. 5(a) where level 4L is lost in the global MRA plot at feature (iv). The corresponding LMRA plot for node 951 in Fig. 6(c) is nevertheless able to clearly identify level 4L as a strongly defined resolution.

in this algorithm, we solve for all communities in the full system and then select the appropriate parent clusters for the community-level analysis. Since the only global parameter that we need to evaluate CVI or CNMI is the system size N , a more efficient approach could take advantage of our local cost function in Eq. (1) (see also Ref. [44] for a more efficient method applied a different fitness function [43]). Specifically, we would solve for the target communities around a particular node of interest a_i by examining community membership opportunities strictly for the neighbors of nodes in or connected to a_i 's local neighborhood. The remainder of the graph partition need not be specified in detail to apply Eqs. (9) and (11).

A more comprehensive alternative to step (3) could be useful if there are no *a priori* nodes of interest to study. That is, we could compare all pairs of clusters and identify the best matching cluster b_{ik} for a_{ij} based on the minimum $v^{(jk)}(a, b)$ at the current γ_i . Then we would average CVI over all cluster matches for each best-cluster pair. In this scenario, we could further pursue the relative cluster comparisons among the replicas by evaluating whether the best clusters match among themselves. That is, we would determine if b_{ik} of partition A also

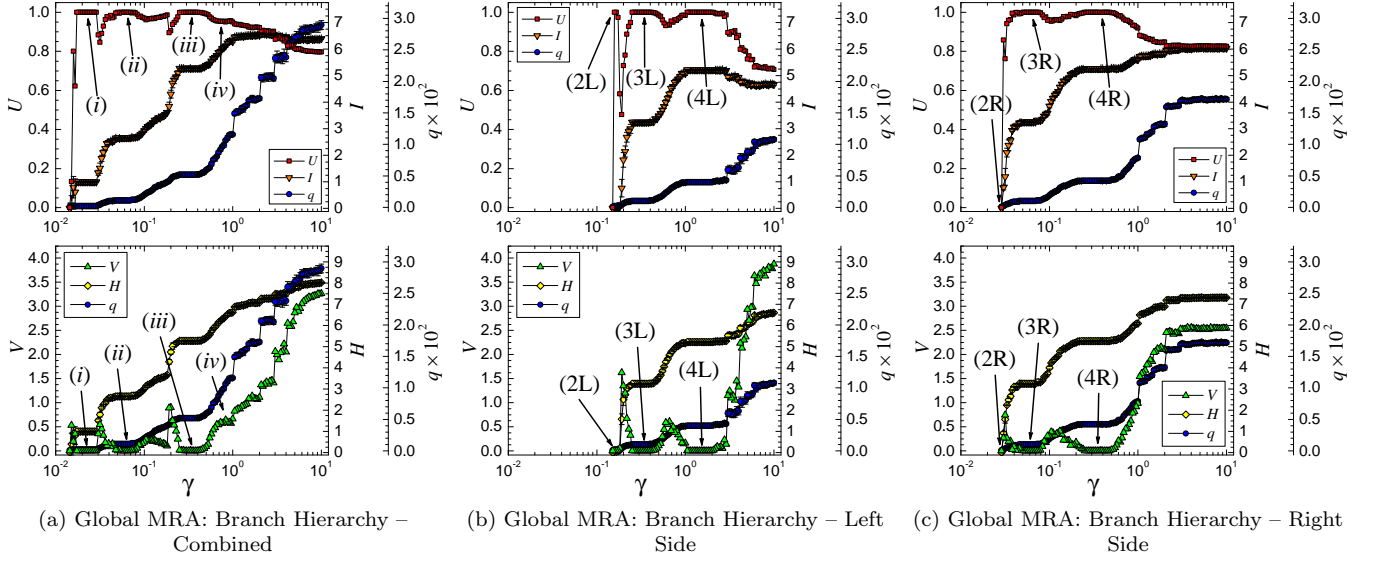


FIG. 5. (Color online) In panels (a), we apply our global multiresolution algorithm (MRA, see Appendix A and Sec. VIB) to the $N = 1024$ node, four-level, branched hierarchy depicted in Fig. 4. Panels (b) and (c) show the MRA method applied separately to the left and right level 2 hierarchy branches, respectively. In the top sub-panels (a-c), we compare replica *partitions* using normalized mutual information U (left axes, see Sec. VA) and mutual information I (right axes). In the corresponding bottom sub-panels, we plot variation of information V (left axes) and the Shannon entropy H (right axes). We also plot the average number of communities q (offset right axes) in top and bottom sub-panels. Features (i)–(iii) demonstrate that the global MRA algorithm can detect network-wide stable partitions [11]. Feature (iv) in panel (a) shows that the level 4 community structure on the left side, known to be present at feature (4L) in panel (b), is *almost completely obscured* because the right branch is significantly more random at the same network scale [*i.e.*, value of γ in Eq. (1), see also Sec. IVB]. In Fig. 6, we compare parent communities using the local multiresolution algorithm in Sec. VI where we demonstrate that the method can accurately extract level 4L for the targeted nodes.

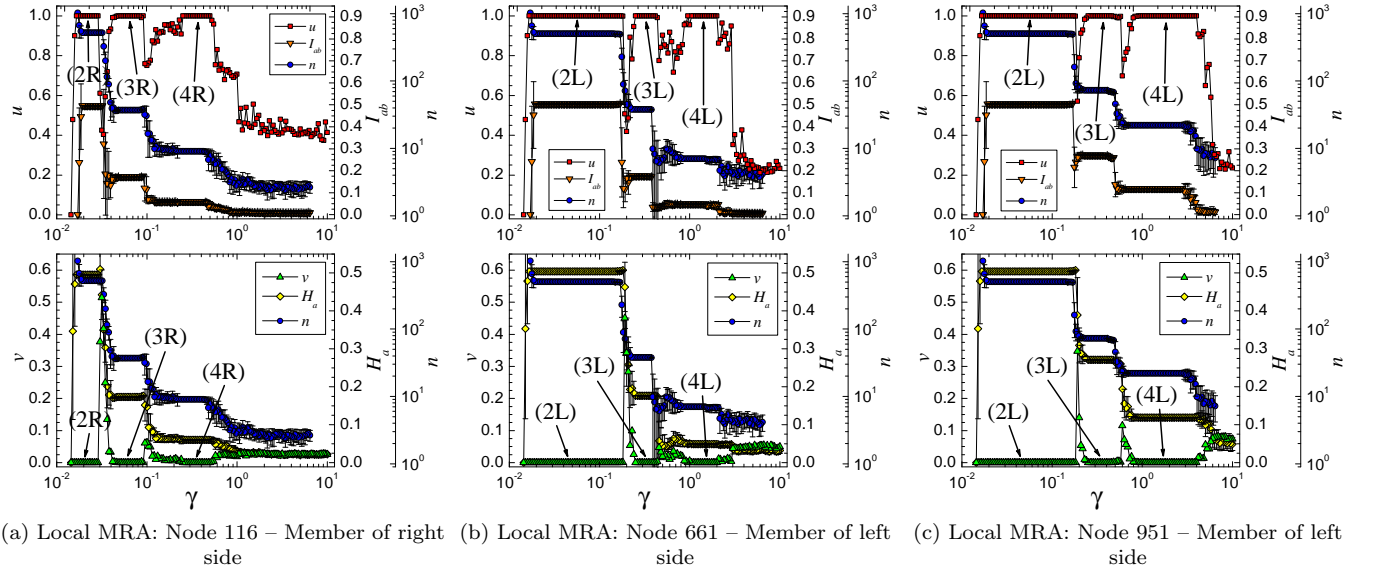


FIG. 6. (Color online) In panels (a-c), we apply our local multiresolution algorithm (LMRA) in Sec. VI to randomly selected nodes of the branched hierarchy depicted in Fig. 4. The top sub-panels compare targeted *communities* in the solved replicas (independent solutions) using the cluster normalized mutual information $u(a,b)$ (left axes, see Sec. VB) and the mutual information contribution I_{ab} . The corresponding bottom sub-panels plot the cluster variation of information $v(a,b)$ (left axes) and the Shannon entropy contribution H_a (right axes). Both top and bottom sub-panels also plot the average number of nodes n in the respective parent communities on the offset right axes. The LRMA method is easily able to extract the relevant levels 3 and 4 for the target nodes as evidenced by regions of low CVI (or high CNMI) even though level 4L of the hierarchy is almost completely obscured at feature (iv) in the combined global MRA plot in Fig. 5(a).

matches the parent cluster d_{il} in partition B , repeating the process to the desired depth.

With this alternative to step (3), individual community matches among the r replicas (see Fig. 3) are not necessarily symmetric. That is, while Eq. (9) is symmetric in (a, A) and (b, B) , this does not require that the best matching clusters in the respective partitions necessarily agree. Consequently, it would provide an additional measure of community robustness based on the level of mutual agreement (number of agreed matches compared to the total possible matches among all replicas).

VII. EXAMPLES

As discussed in Sec. VIB, we calculate the global MRA algorithm for the network and concurrently apply the LMRA algorithm in Sec. VI to targeted nodes by tracking the respective parent clusters across a full range of network scales. Comparing explicit values of VI and CVI is difficult, so we evaluate relative values of VI or CVI for a given network. We demonstrate the LMRA method with a constructed network example and a small, real terror network.

A. Branched hierarchy

We construct a branched, strict hierarchy as depicted in Fig. 4 which we use to test the LMRA method of Sec. VI. Level 1 is the full system of $N = 1024$ nodes; level 2 is the two-part branch split (groups of superclusters) with $N_L = 502$ and $N_R = 522$ nodes for the left (L) and right (R) sides, respectively; level 3 is the set of superclusters; level 4 is the set of innermost clusters.

Level 1 was defined by connecting nodes in the left and right branches (levels 2L and 2R) with an *intercommunity* density $p_1 = 0.015$. The approximate *intracommunity* edge densities at level 4 were $p_{4L} = 0.9$ and $p_{4R} = 0.6$ assigned randomly with a normal distribution of $\sigma_p = 0.02$. We connected nodes *between* the respective communities in the intermediate levels 2 and 3 with probabilities: $p_{3L} = 0.37$, $p_{3R} = 0.10$, $p_{2L} = 0.16$, and $p_{2R} = 0.03$. These values were selected in order to demonstrate a somewhat “blurred” multiresolution signature in a controlled example where the underlying local structure is nevertheless strongly defined.

In Fig. 5(a), we show the global MRA algorithm from Ref. [11] (summarized in Sec. VIB) applied to the full $N = 1024$ node network using $r = 20$ replicas and $t = 10$ optimization trials per replica. A more thorough discussion follows, but briefly, feature (iv) illustrates how poorly-correlated communities almost completely obscure the well-defined level 4L structure. Nevertheless, the local MRA algorithm in Sec. VI can *fully extract this hidden section of the hierarchy*.

In Fig. 5, the left axes plot NMI, U , and VI, V , from Sec. VA in the top and bottom sub-panels, respectively,

averaged over all replica pairs. On the right axes, we plot the average mutual information I and the Shannon entropy H for the top and bottom sub-panels, respectively. The right offset axes in both sub-panels plot the average number of communities q . Panels (b) and (c) show the MRA results applied to the separate left and right branches of the hierarchy, respectively, using the same r and t as in panel (a).

For completeness, features (i)–(iii) in panel (a) illustrate how the global MRA signature can identify preferred or stable resolutions by low VI or high NMI correlations (or plateaus in H , I , and q in this example) averaged over the independently-solved replica partitions. Specifically, feature (i) corresponds to level the 2 partition with $q_i = 2$, and feature (ii) concurrently identifies levels 2L and 3R with $q_{ii} = 11$ because the respective community edge densities are similar (see Sec. IV B). Likewise, feature (iii) solves levels 3L and 4R with $q_{iii} = 52$. These specific partitions consist of combinations of well-resolved sub-graphs at different levels of the branched hierarchy, but it is the loss of level 4L in the global MRA plot that is the main topic of this example.

At feature (iv) in panel (a), the poor correlations show that *the global analysis of the full system misses level 4L*. This occurs because the well-defined local clusters conflict with more random partitions for the right-side subgraph in Fig. 4. In contrast, panels (b) and (c) show that the MRA method applied to the separate left and right branches are perfectly defined with $V = 0$ and $U = 1$ [marked by (2L), (3L), ..., (4R), respectively]. That is, the structure clearly exists locally, but the global MRA method in panel (a) cannot resolve level 4L.

In Fig. 6(a–c), we plot the results of the new LMRA method from Sec. VB for the parent clusters of three randomly selected nodes 116, 661, and 951, respectively, as identified within the full $N = 1024$ node system. On the left axes, we plot CNMI $u(a, b)$ in Eq. (11) and CVI $v(a, b)$ in Eq. (9), respectively, averaged over all community pairs in the respective replicas. On the right axes, we plot the mutual information contribution I_{ab} in Eq. (8) and the Shannon entropy contribution H_a in Eq. (7) averaged over all pairs of target communities in the replicas or all target communities, respectively. The offset right axes plot the average number of nodes n over all targeted communities.

Despite being buried within the full $N = 1024$ node system, the parent cluster of node 951 corresponding to level 4L is clearly present in the LMRA analysis in Fig. 6(b,c). This illustrates how *our LMRA algorithm can resolve well-defined local structure even when the global signature is obscured*. In principle, we could further apply the LMRA algorithm to all clusters in the partitions and unambiguously identify the entire set of well-defined level 4L communities.

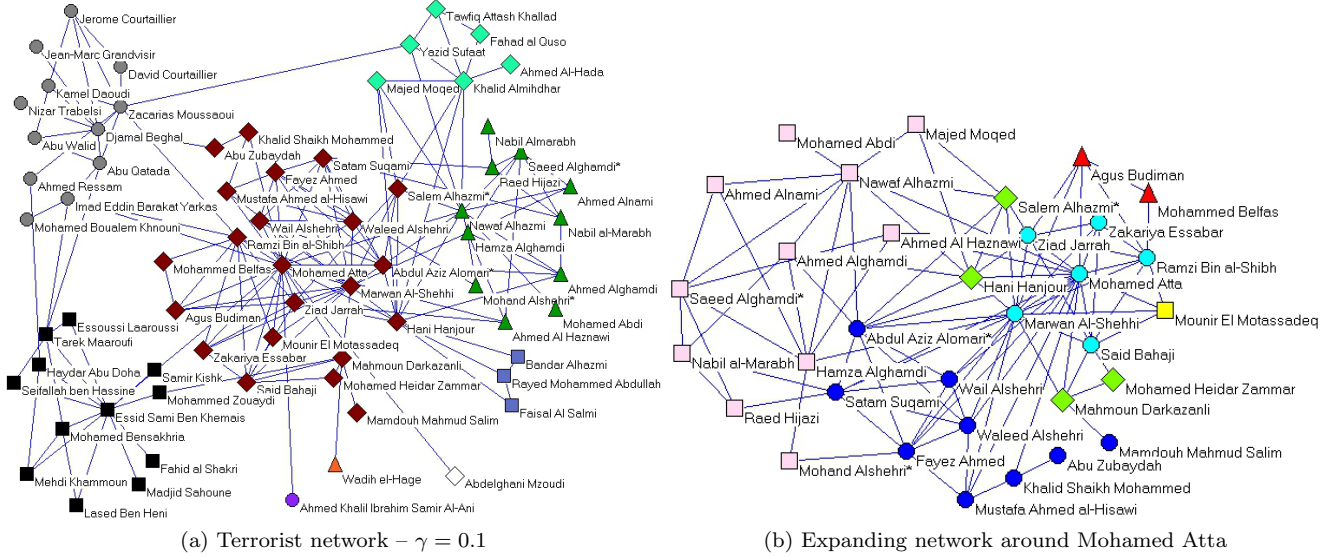


FIG. 7. (Color online) The figure depicts a small terrorist network collected from publicly available data [87]. Panel (a) shows the overall network at $\gamma = 0.1$ in Eq. (1) where distinct node shapes indicate separate communities. Panel (b) shows an “expanding” community around Mohamed Atta where his “local” cluster grows roughly outward in the diagram. Here, new node categories (shapes and colors) indicate nodes *added* to the parent cluster (as opposed to new communities) as γ is lowered to particular well-defined resolutions (see text). In this network, our local multiresolution algorithm indicates that these communities are strongly defined on an individual basis with CVI $v(a, b) = 0$ in Fig. 9(b) even at resolutions where the overall system structure is more vaguely defined in Fig. 8. This illustrates the main benefit of our local multiresolution approach.

B. Small terrorist network

Even small networks can possess strongly-defined local clusters within a more indistinct global partition. We apply the LMRA method to a small terrorist network related to the terrorism attacks of September 11, 2001, as constructed from publicly available data [87]. Given that the highest quality intelligence would be classified, our purpose here is to demonstrate the practical application of the LMRA on real data as opposed to setting forth a rigorous study of the terrorist network. Toward this end, we select Mohamed Atta for study as the leader of the operation. We further consider Hani Hanjour, who was another pilot, and Zacarias Moussaoui, who was the only terrorist of the 20 prevented from participating in the attack.

Figure 7(a) depicts the network at $\gamma = 0.1$ in Eq. (1) corresponding to the minimum VI at feature (i) in Fig. 8 with $V \simeq 0$ (see below). Here, distinct node shapes indicate separate *communities*. The community partitions with $V = 0$ at the lowest γ 's are disjoint clusters where all nodes are completely collapsed into their respective connected groups, resulting in a trivial partition. The left axes plot U and V (see Sec. V A) for top and bottom sub-panels, respectively, averaged over all replica pairs. On the right axes, we plot I and H for the top and bottom sub-panels, respectively, and the offset axes in both sub-panels plot the average number of communities q .

Figure 7(b) shows the expanding network core centered

on Mohamed Atta at several strongly-defined resolutions as determined from Fig. 9(b) where $v(a, b) = 0$. In this panel, distinct node shapes and colors indicate *added nodes* [as opposed to new communities in panel (a)], roughly spreading outward as γ is lowered. Specifically, the fixed resolutions correspond to $\gamma = 10$ (smallest, innermost cyan circles), $\gamma = 3$ (yellow square), $\gamma = 0.6$ (green diamonds), $\gamma = 0.3$ (red triangles), $\gamma = 0.125$ (dark blue circles), and $\gamma = 0.05$ (largest, pink squares) with a few other small fluctuations not depicted.

On the left axes in Fig. 9(a-c), we plot CNMI $u(a, b)$ in Eq. (11) and CVI $v(a, b)$ in Eq. (9), respectively, averaged over all pairs of parent communities in the respective replicas. Similarly, the right axes plot the mutual information contribution I_{ab} in Eq. (8) and the Shannon entropy contribution H_a in Eq. (7) averaged over all pairs of parent communities or all parent communities, respectively. The right offset axes display the average number of nodes n over the parent communities.

Each panel shows distinct, but different, regions of γ where the parent clusters are strongly defined, but the cluster correlations in the full network in Fig. 8 are more poorly defined at most resolutions. Hani Hanjour has a LMRA signature distinct from Mohamed Atta for $\gamma \gtrsim 1$, but they match at lower γ because they are mutual members of the same communities.

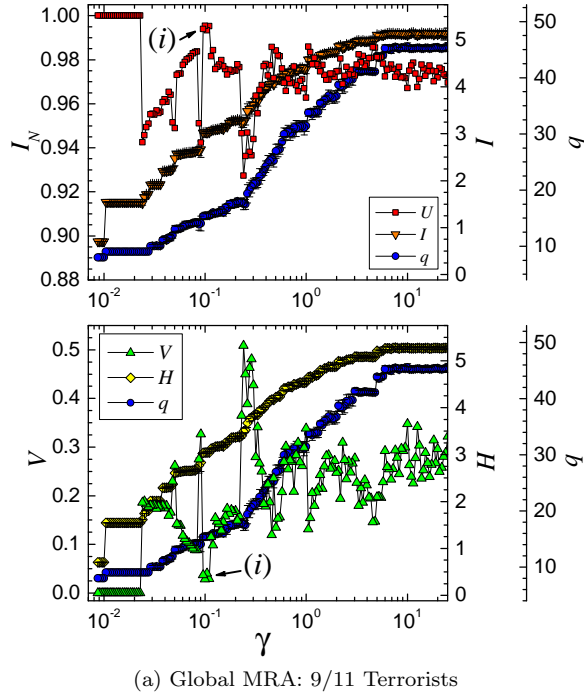


FIG. 8. (Color online) We apply our multiresolution algorithm (see Appendix A and Sec. VIB) to a small terrorist network [87]. Although the plot shows a best resolution at $\gamma \simeq 0.1$ (depicted in Fig. 7) as indicated by $V \simeq 0$, the remainder of the plot has a largely blurred multiresolution signature (high VI or low NMI). The $V = 0$ region on the far left is an essentially trivial partition into nearly disjoint clusters. In Fig. 9, we show results from the local multiresolution algorithm in Sec. VI to three selected terrorists where we track the respective parent clusters over a range of resolutions [*i.e.*, values of γ in Eq. (1)] and calculate the cluster correlations using the CVI and CNMI in Sec. VB.

C. LFR Benchmark

We also tested the LMRA method on a common CD benchmark by Lancichinetti Fortunato and Radicchi [76], which was designed to emulate a series of strongly defined networks with realistic distributions of community sizes and edge assignments. Specifically, it defines a power-law distribution of community sizes specified by an exponent β , minimum size n_{\min} , and maximum size n_{\max} . It adds random unweighted, undirected edges to the network, defining both the communities and any intercommunity noise (extraneous edges outside of the well-defined communities), according to a power-law distribution of node degrees given by an exponent α , average power-law degree $\langle k \rangle$ (or minimum degree k_{\min}), and maximum degree k_{\max} .

In the current tests, we solve for each system using the algorithm in Sec. VIC using 20 replicas to ensure that we have a good sample of possible partitions and $t = 4$ trials per CD solution attempt. We used $N = 10000$, $\langle k \rangle = 35$, $n_{\min} = 10$, $n_{\max} = 50$, $\alpha = -2$, $\beta = -1$, and $\mu = 0.1$

or 0.5 as indicated in Fig. 10. The mixing parameter μ controls the level of intercommunity noise. The results of the global MRA method of Sec. VIB are shown in Fig. 10, and the corresponding local LMRA data for three randomly selected nodes are in Fig. 11. As with previously demonstrated examples [11, 75], the global MRA algorithm correctly identifies the constructed global partition with exceptional accuracy in the presence low and high noise in panels (a) and (b). The extreme noise case in panel c was included for comparison purposes, since mixing parameters higher than $\mu \simeq 0.7$ present exceptional challenges for all tested CD algorithms in [75]. The planted partition may even exceed limits of well-defined communities (see [28, 32, 48] for some general discussion).

In Fig. 11(a) where $\mu = 0.1$, the LRMA method identifies the proper community of node 7603, as indicated by the arrows at feature (i). Panel (b) of Fig. 11 for node 1213 shows a similar for $\mu = 0.5$, where the correct cluster is again identified perfectly. In both cases, the cluster structures appear to have reasonably well-defined parent communities with v being roughly near zero for $\gamma \gtrsim 4$ and 1, respectively, but the CNMI measure clearly shows $u < 1$ indicating poorly defined communities. Thus, the proper interpretation of parameters v and u requires considering the relative values over the studied range of γ . Further, one should consider all of the measures together when identifying relevant community structure(s).

Feature (i) in panel (c) for node 2502 indicates two γ values which display perfect correlations in CNMI and CVI, respectively. At $\gamma = 0.12$, the parent cluster of the target node forms a trivial split into a size $n = 2$ community due to the excessive noise. The result at $\gamma = 0.095$ shows a partial-identification of the intended cluster with $u = 0.88$ when compared with the intended community by construction. This correlation represents a transient, partially-false-positive identification since $u = 1$ among the algorithm replicas. Additional helps mitigate such identifications. The large error bars to the left of feature (i) indicate wide disagreement between the different candidate clusters among the replicas.

For other problems with very low noise (not depicted), both CVI and CNMI may indicate a strong partition with $u \simeq 1$ or $v \simeq 0$ across a wide range of γ 's without clearly marking the transitions by changes in v or u . This may imply that the communities in which the node resides are not blurred by significant noise so that many of the structural transitions as γ is varied are sharply defined. The supplemental measures of H_a and I_{ab} for communities a and b or the number of communities q can help distinguish between these different structures, but in these cases, the base LMRA method may not be conclusive.

VIII. CONCLUSION

Multiresolution network analysis extends the basic notions of community detection to select the best resolution(s) for a given network over a range of network scales.

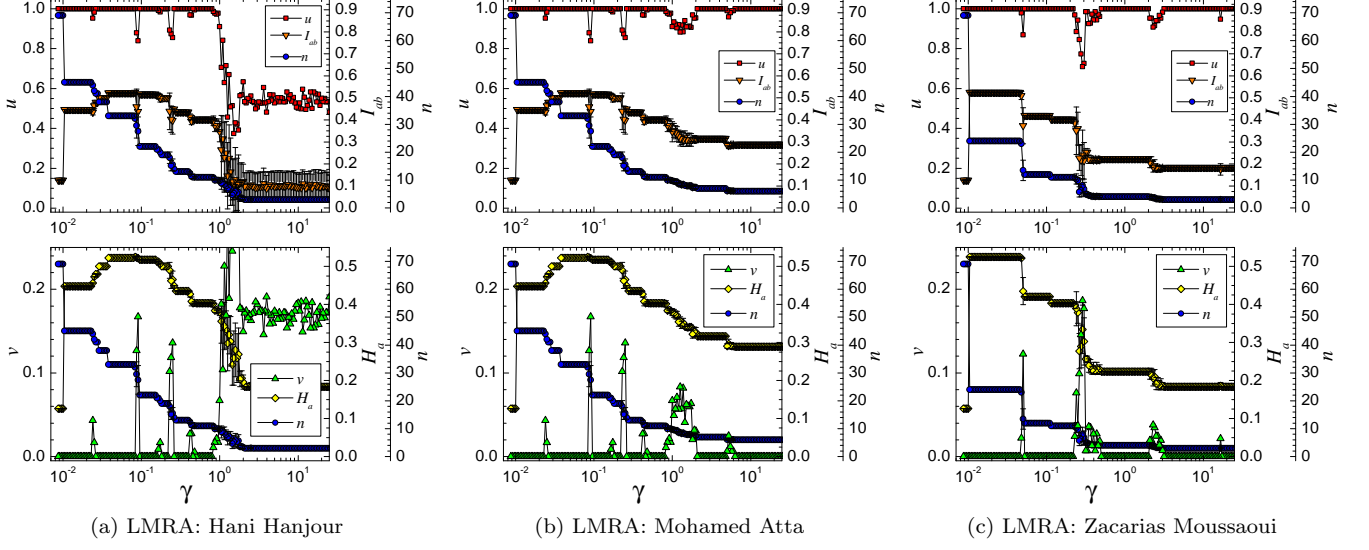


FIG. 9. (Color online) In each panel, we apply our local multiresolution algorithm (LRMA, see Sec. VI) to a small terrorist network [87]. We analyze three selected terrorists (see text) by tracking the respective parent clusters over a range of resolutions [*i.e.*, values of γ in Eq. (1)]. We then calculate the cluster correlations using the community comparison measures in Sec. V B. Note that the individual nodes possess certain strongly preferred resolutions with $v(a, b) = 0$ for their parent clusters whereas the global system in Fig. 8 is less well-defined for most values of γ .

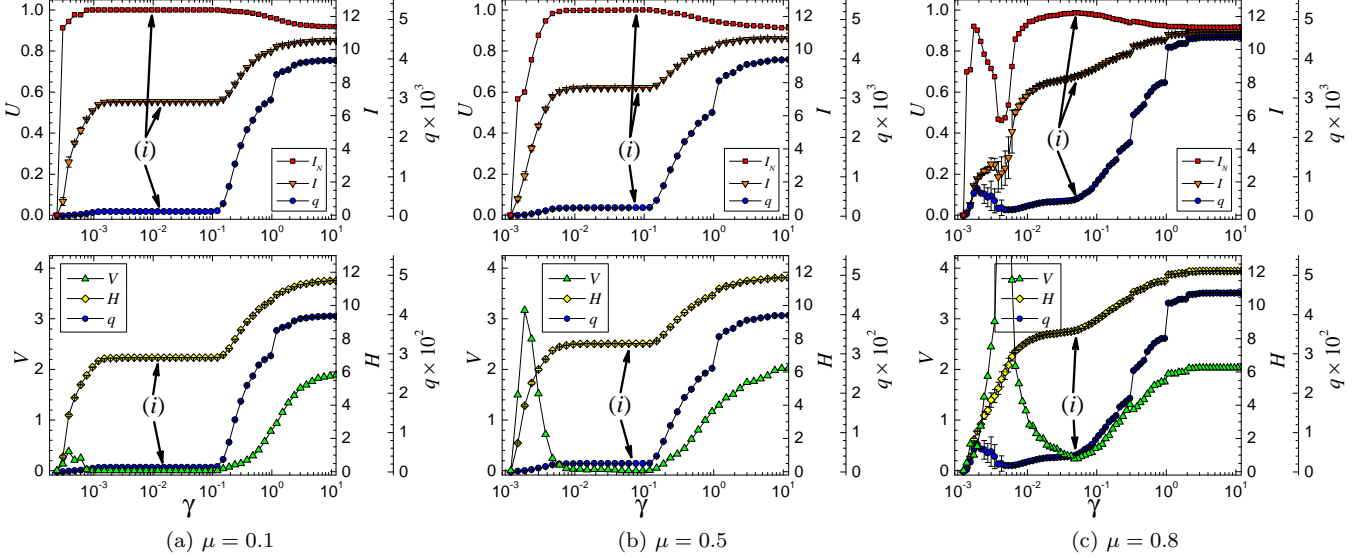


FIG. 10. (Color online) We apply our multiresolution algorithm from Sec. VIB [11, 75] to the LFR benchmark [76] with mixing parameters $\mu = 0.1, 0.5$, and 0.8 . These values correspond to a low (panel a), moderately high (b), and extremely high (c) levels of network noise, respectively. See the text for other benchmark parameters. In panels (a) and (b), the constructed partitions are correctly identified at feature (i) by low VI or high NMI. In panel (c), a preferred partition resolution is implied by similar extrema at feature (i), but the exact identity of the intended partition is probably beyond the capability of our algorithm to extract. Further, based on overall results in [75], the intended partition likely lies beyond the detection capability of current CD algorithms. Other works [28, 32, 48] discuss maximum detectability limits and transitions, which may apply here. In Fig. 11, we show corresponding results from the local multiresolution algorithm in Sec. VI for randomly selected nodes where we track the respective parent clusters over a range of resolutions [*i.e.*, values of γ in Eq. (1)] and calculate the cluster correlations using the CVI and CNMI in Sec. V B.

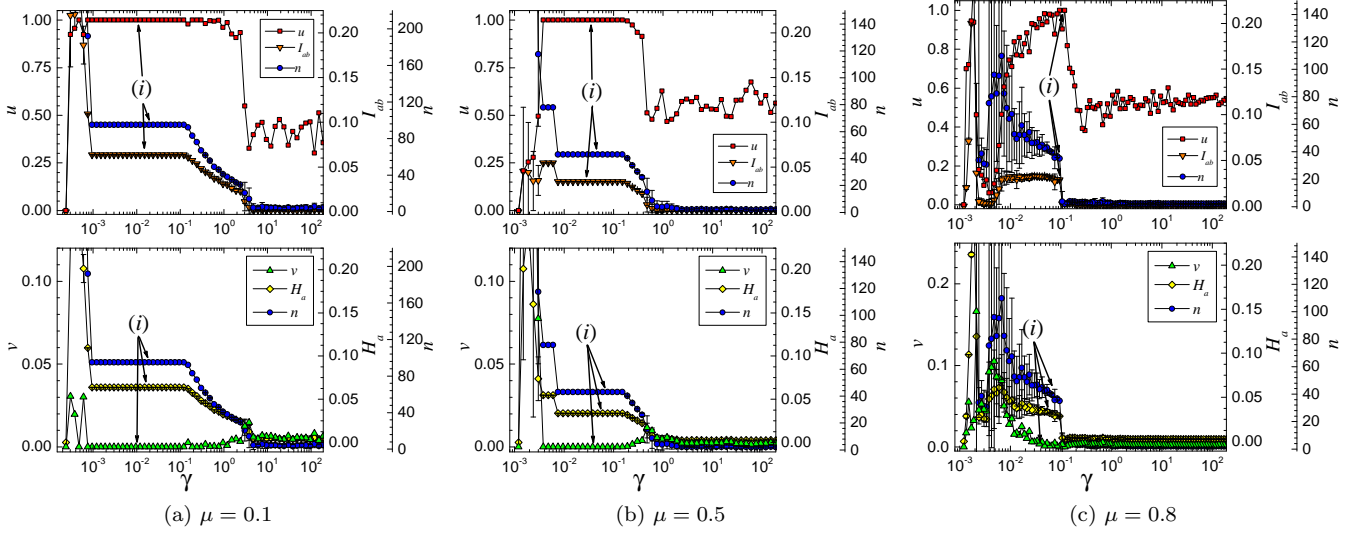


FIG. 11. (Color online) In each panel, we apply our local multiresolution algorithm of Sec. VI to an implementation of the LFR benchmark [76]. Mixing parameters are $\mu = 0.1$, 0.5 , and 0.8 which correspond to a low (panel a), moderately high (b), and extremely high (c) levels of network noise, respectively. See the text for other benchmark parameters. We analyze the parent communities for randomly selected nodes over a range of resolutions [*i.e.*, values of γ in Eq. (1)]. Other nodes display comparable multiresolution signatures. We then calculate the cluster correlations using the community comparison measures in Sec. V B. In panels (a) and (b), the defined communities are correctly identified as feature (i) with $v = 0$ and $u = 1$ for their parent clusters, corresponding to the correct solution for the global system in Fig. 10. The correct benchmark communities are clearly detected as stable regions in each of the various cluster measures. In panel (b), note that v is near zero for this plot beyond $\gamma \gtrsim 4$, but the CNMI measure clearly shows relatively low u values, indicating that we should focus on the correct partition at feature (i). Feature (i) in panel (c) has two perfect correlations in CNMI and CVI at $\gamma = 0.12$ and 0.095 , respectively. The former is a trivial split of the target node into a size $n = 2$ community due to the excessive noise, and the latter is a partial-identification of the intended cluster ($u = 0.88$ with the intended construction), which represents a transient, partially-false-positive identification since $u = 1$ among the algorithm replicas. However, the exact intended partition likely lies beyond the detection capability of current CD algorithms (see [75] and also [28, 32, 48] which discuss maximum detectability limits and transitions), and there is some concern as to whether the intended partition is even well-defined.

Certain networks may present situations where local clusters experience a lost-in-a-crowd effect. Despite being strongly defined, the local structure may be lost among a collection of more poorly defined communities at a given resolution. This may occur due to the sheer size of a network or because most clusters do not coalesce in their strongest state(s) at the same scale(s).

We presented an extension of an existing global multiresolution method [11] to detect and quantitatively assess local multiresolution structure. We proposed cluster-level analogies to variation of information and normalized mutual information which evaluate the strength of local communities in the context of a pair of network partitions. After applying these measures to evaluate correlations among individual parent communities in multiple independent solutions (replicas), we demonstrated that the proposed local multiresolution algorithm is able to identify the best resolutions and extract the local structure despite a blurred global multiresolution signature. In addition, we demonstrated that the algorithm correctly identifies individual clusters in the single-layered structure of the LFR benchmark. Our approach only requires output communities from a multiresolution capable CD algorithm, so it is independent of the search

algorithm or CD model, making it suitable for use with any CD method that can identify partitions across different network scales.

ACKNOWLEDGMENTS

This work was supported by NSF grant DMR-1106293 (ZN). We wish to thank S. Chakrabarty, R. K. Darst, P. Johnson, and D. Hu for discussions and ongoing work. ZN also thanks the Aspen Center for Physics and NSF Grant #1066293 for hospitality during the final stages of this work.

Appendix A: Local and global terminology

The meaning of the terms local and global depends on the context. For our purposes, global *cost functions* are those that *require* network wide (global) parameters (e.g., number of edges L , number of communities q , overall graph density p , etc.) in the quantitative evaluation of community structure [14, 17]. Global *multiresolution*

methods are those for which the best partition is simultaneously determined for the entire system, effectively averaging the partition robustness over all communities. This is true regardless of whether the cost function is itself local or global in nature.

Local *cost functions* [11, 40, 41] or algorithms [15] utilize parameters only in the neighborhood of a community or node (e.g., size of community a , edges of node i , etc.) to evaluate the best community structure. These can be subdivided into weak and strong local cost functions [41] where weakly-local cost functions may depend on the details of the community structure. Briefly, strongly local cost functions determine community membership for a given node based only on the node's own relations with candidate communities. Local *multiresolution methods*, such as the current work, seek to identify the best communities based on their own robustness at a given resolution. That is, each community determines whether it is strongly defined regardless of the community structure present in the remainder of the network. The evaluation of the best resolution is not effectively averaged over all the communities in the graph. Further, each community may be strongly resolved at different network scales (often described in terms of certain model weight parameters).

Appendix B: Semi-metric property of CVI

A semi-metric possesses intuitive “distance-like” properties for comparing cluster similarity. The proof that CVI is a semi-metric is trivial. A measure $S(a, b)$ on a set X with two variables a and b in X is a semi-metric if and only if it satisfies the following conditions:

- Non-negativity – $S(a, b) \geq 0$ for all a and b .
- Zero only for equality – $S(a, b) = 0$ if and only if $a = b$.
- Symmetry – $S(a, b) = S(b, a)$ for all a and b .

$S(a, b)$ is a metric if it additionally satisfies the triangle inequality $S(a, c) \leq S(a, b) + S(b, c)$ for three variables a , b , and c in X .

Theorem 1. *CVI in Eq. (9) is a semi-metric between two clusters a and b in partitions A and B of size $|A| = |B| = N$ in the space of possible partitions of the N nodes: (1) It is non-negative and equal to zero if and only if $a = b$. (2) It is symmetric with respect to clusters (a, A) and (b, B) , $v(a, b) = v(b, a)$.*

Proof.

- (1) It is non-negative and strictly equal to zero if and

only if $a = b$. From Eq. (9)

$$\begin{aligned} v(a, b) &= -\frac{n_a}{N} \log\left(\frac{n_a}{N}\right) - \frac{n_b}{N} \log\left(\frac{n_b}{N}\right) \\ &\quad - 2\frac{n_{ab}}{N} \log\left(\frac{n_{ab}N}{n_a n_b}\right) \\ &= \frac{n_a - n_{ab}}{N} \log\left(\frac{N}{n_a}\right) + \frac{n_b - n_{ab}}{N} \log\left(\frac{N}{n_b}\right) \\ &\quad + \frac{n_{ab}}{N} \log\left(\frac{n_a}{n_{ab}}\right) + \frac{n_{ab}}{N} \log\left(\frac{n_b}{n_{ab}}\right) \\ v(a, b) &\geq 0 \end{aligned} \tag{B1}$$

since $n_a > 0$, $n_b > 0$, $n_{ab} \geq 0$, $n_a \geq n_{ab}$, $n_b \geq n_{ab}$, $N \geq n_a$, and $N \geq n_b$.

Furthermore, $a = b$ implies that $n_a = n_b = n_{ab}$, which trivially yields

$$\begin{aligned} v(a, a) &= -\frac{n_a}{N} \log\left(\frac{n_a}{N}\right) - \frac{n_a}{N} \log\left(\frac{n_a}{N}\right) \\ &\quad - 2\frac{n_a}{N} \log\left(\frac{n_a N}{n_a n_a}\right) \\ &= -2\frac{n_a}{N} \log\left(\frac{n_a}{N}\right) + 2\frac{n_a}{N} \log\left(\frac{n_a}{N}\right) \\ v(a, a) &= 0. \end{aligned} \tag{B2}$$

Now, if $v(a, b) = 0$, this implies

$$\begin{aligned} v(a, b) &= -\frac{n_a}{N} \log\left(\frac{n_a}{N}\right) - \frac{n_b}{N} \log\left(\frac{n_b}{N}\right) \\ &\quad - 2\frac{n_{ab}}{N} \log\left(\frac{n_{ab}N}{n_a n_b}\right) \\ 0 &= \frac{n_a - n_{ab}}{N} \log\left(\frac{N}{n_a}\right) + \frac{n_b - n_{ab}}{N} \log\left(\frac{N}{n_b}\right) \\ &\quad + \frac{n_{ab}}{N} \log\left(\frac{n_a}{n_{ab}}\right) + \frac{n_{ab}}{N} \log\left(\frac{n_b}{n_{ab}}\right) \end{aligned} \tag{B3}$$

Since $n_{ab} \geq 0$, $n_a \geq n_{ab}$, $n_b \geq n_{ab}$, $N \geq n_a$, and $N \geq n_b$, all terms are non-negative; so they cannot cancel each other. Each term must be individually zero.

First, if $n_{ab} = 0$ the last two terms in Eq. (B3) are zero. Then since $n_a > 0$ and $n_b > 0$, $N = n_a = n_b$, but this result requires that $n_{ab} = N$ which contradicts the assumption that $n_{ab} = 0$. Thus, $n_{ab} > 0$.

Now, the third term in Eq. (B3) vanishes only if $n_a = n_{ab}$, and fourth term is zero only if $n_b = n_{ab}$. Thus, $n_a = n_b = n_{ab}$ which means that $a = b$.

Thus, $v(a, b) = 0$ if and only if $a = b$.

(2) It is symmetric with clusters (a, A) and (b, B) , $v(a, b) = v(b, a)$.

Since n_{ab} is necessarily equal to n_{ba} , $I_{ab}(A, B)$ is symmetric in clusters (a, A) and (b, B) . The symmetry of $v(a, b)$ is then immediately obvious.

Thus, CVI is a semi-metric. \square

We have not proved the triangle inequality for CVI, making it a metric, but the triangle inequality appears to be violated rarely, if at all.

Appendix C: Alternative cluster measures

A tempting alternate measure for CVI might be defined based on the individual terms of

$$V(A, B) = H(A|B) + H(B|A) = \sum_{a,b} \left[\frac{n_{ab}}{N} \log \frac{n_b}{n_{ab}} + \frac{n_{ab}}{N} \log \frac{n_a}{n_{ab}} \right]. \quad (\text{C1})$$

where A and B are two partitions of a graph. From this equivalent definition of VI, the natural CVI definition for two clusters a in A and b in B would be

$$v'_{ab}(A, B) = \frac{n_{ab}}{N} \log \frac{n_a}{n_{ab}} + \frac{n_{ab}}{N} \log \frac{n_b}{n_{ab}}. \quad (\text{C2})$$

Unlike CVI in Eq. (9), Eq. (C2) has the nice property that the individual cluster contributions sum to $V(A, B)$, $V(A, B) = \sum_a^{q_A} \sum_b^{q_B} v'_{ab}(A, B)$.

Unfortunately, this particular measure does not work for cluster comparisons. While $v'_{aa}(A, A) = 0$ as desired, it is also the case that $v'_{ab}(A, B) = 0$ if $n_{ab} = 0$. That is, it is zero if *no overlap* exists between a and b which violates the notion of a “distance” as well as one of the requirements for being a (semi)metric. VI is a metric on partitions A and B because it *sums* over all a and b in A and B , respectively.

We could also consider an alternate *ad hoc* definition by redefining the CVI entropy terms in Eq. (10) according to $v''_{ab}(A, B) = H_a(A)/q_B + H_b(B)/q_A - 2I_{ab}(A, B)$. This variant would again yield the desirable property $V(A, B) = \sum_a^{q_A} \sum_b^{q_B} v''_{ab}(A, B)$, but the measure loses the semi-metric requirements $v''_{ab}(A, B) \geq 0$ and $v''_{aa}(A, A) = 0$.

-
- [1] A. E. Motter and R. Albert, *Physics Today* **65**, 43 (2012).
 - [2] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
 - [3] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, *PLoS ONE* **5**, e11976 (2010).
 - [4] P. Ronhovde, S. Chakrabarty, D. Hu, M. Sahu, K. K. Sahu, K. F. Kelton, and Z. Nussinov, *Euro. Phys. J. E* **34**, 105 (2011).
 - [5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Nature (London)* **435**, 814 (2005).
 - [6] A. Arenas, A. Fernández, and S. Gómez, *New J. Phys.* **10**, 053039 (2008).
 - [7] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, *Fluct. Noise Lett.* **7**, L209 (2007).
 - [8] D. J. Fenn, M. A. Porter, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones, *Chaos* **19**, 033119 (2009).
 - [9] M. Rosvall and C. T. Bergstrom, *PLoS ONE* **6**, e18209 (2011).
 - [10] X.-Q. Cheng and H.-W. Shen, *J. Stat. Mech.* **04**, P04024 (2010).
 - [11] P. Ronhovde and Z. Nussinov, *Phys. Rev. E* **80**, 016109 (2009).
 - [12] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, *J. Stat. Mech.* **09**, P09008 (2005).
 - [13] M. Meilä, *J. Multivariate Anal.* **98**, 873 (2007).
 - [14] M. E. J. Newman and M. Girvan, *Phys. Rev. E* **69**, 026113 (2004).
 - [15] U. N. Raghavan, R. Albert, and S. Kumara, *Phys. Rev. E* **76**, 036106 (2007).
 - [16] M. J. Barber and J. W. Clark, *Phys. Rev. E* **80**, 026129 (2009).
 - [17] J. Reichardt and S. Bornholdt, *Phys. Rev. E* **74**, 016110 (2006).
 - [18] J. Reichardt and S. Bornholdt, *Phys. Rev. Lett.* **93**, 218701 (2004).
 - [19] S. Fortunato and M. Barthélemy, *Proc. Natl. Aca. Sci. U.S.A.* **104**, 36 (2007).
 - [20] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, *Euro. Phys. J. B* **56**, 41 (2007).
 - [21] X. S. Zhang, R. S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang, and L. Chen, *Europhys. Lett.* **87**, 38002 (2009).
 - [22] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **84**, 066122 (2011).
 - [23] J. Xiang and K. Hu, *Physica A* **391**, 4995 (2012).
 - [24] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Social networks* **5**, 109 (1983).
 - [25] T. A. Snijders and K. Nowicki, *J. classification* **14**, 75 (1997).
 - [26] M. E. Newman and E. A. Leicht, *PNAS* **104**, 9564 (2007).
 - [27] P. Latouche, E. Birmele, and C. Ambroise, *Statistical Modelling* **12**, 93 (2012).
 - [28] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011).
 - [29] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Phys. Rev. E* **84**, 066106 (2011).
 - [30] D. Hu, P. Ronhovde, and Z. Nussinov, *Philosophical Magazine* **92**, 406 (2012).
 - [31] D. Hu, P. Ronhovde, and Z. Nussinov, *Phys. Rev. E* **86**, 066106 (2012).
 - [32] R. K. Darst, D. R. Reichman, P. Ronhovde, and Z. Nussinov, *J. Complex Networks* **2** (2014), 10.1093/comnet/cnu042.
 - [33] B. Karrer and M. E. J. Newman, *Phys. Rev. E* **83**, 016107 (2011).
 - [34] Y. Zhu, X. Yan, and C. Moore, *J. Complex Networks* **2**, 1 (2014).
 - [35] X. Yan, C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang, and Y. Zhu, *J. Stat. Mech.: Theory and Exp.* **2014**, P05007 (2014).
 - [36] M. Blatt, S. Wiseman, and E. Domany, *Phys. Rev. Lett.* **76**, 3251 (1996).
 - [37] I. Ispolatov, I. Mazo, and A. Yuryev, *J. Stat. Mech.* **09**, P09014 (2006).
 - [38] M. B. Hastings, *Phys. Rev. E* **74**, 035102 (2006).
 - [39] V. A. Traag and J. Bruggeman, *Phys. Rev. E* **80**, 036115 (2009).
 - [40] V. A. Traag, P. Van Dooren, and Y. Nesterov, *Phys. Rev. E* **84**, 016114 (2011).
 - [41] P. Ronhovde and Z. Nussinov, *Phys. Rev. E* **81**, 046114 (2010).
 - [42] J. P. Bagrow and E. M. Boltt, *Phys. Rev. E* **72**, 046108 (2005).

- (2005).
- [43] A. Lancichinetti, S. Fortunato, and J. Kertész, *New J. Phys.* **11**, 033015 (2009).
 - [44] F. Havemann, M. Heinz, A. Struck, and J. Gläser, *J. Stat. Mech.* **01**, P01023 (2011).
 - [45] Y. Zhao, E. Levina, and J. Zhu, *PNAS* **108**, 7321 (2011).
 - [46] A. Clauset, *Phys. Rev. E* **72**, 026132 (2005).
 - [47] S. Muff, F. Rao, and A. Cafilisch, *Phys. Rev. E* **72**, 056107 (2005).
 - [48] D. Hu, P. Ronhovde, and Z. Nussinov, *Phil. Mag.* **92**, 406 (2012).
 - [49] B. H. Good, Y.-A. de Montjoye, and A. Clauset, *Phys. Rev. E* **81**, 046106 (2010).
 - [50] R. R. Nadakuditi and M. E. J. Newman, *Phys. Rev. Lett.* **108**, 188701 (2012).
 - [51] P. Ronhovde, D. Hu, and Z. Nussinov, *EPL* **99**, 38006 (2012).
 - [52] L. Danon, A. Díaz-Guilera, and A. Arenas, *J. Stat. Mech.* **11**, P11010 (2006).
 - [53] B. Everitt, S. Landau, and M. Leese, *Cluster analysis* (2001).
 - [54] A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453**, 98 (2008).
 - [55] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, *PNAS* **104**, 15224 (2007).
 - [56] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *J. Stat. Mech.* **10**, P10008 (2008).
 - [57] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, *Physica A* **388**, 1706 (2009).
 - [58] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Nature (London)* **466**, 761764 (2010).
 - [59] E. Ravasz and A.-L. Barabási, *Phys. Rev. E* **67**, 026112 (2003).
 - [60] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnella, *Science* **328**, 876 (2010).
 - [61] J. Zhang, K. Zhang, X. ke Xu, C. K. Tse, and M. Small, *New J. Phys.* **11**, 113003 (2009).
 - [62] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, *PLoS ONE* **6**, e18961 (2011).
 - [63] H.-W. Shen, X.-Q. Cheng, and B.-X. Fang, *Phys. Rev. E* **82**, 016114 (2010).
 - [64] H.-W. Shen, X.-Q. Cheng, and B.-X. Fang, *Phys. Rev. E* **82**, 016114 (2010).
 - [65] S. Zhang and H. Zhao, *Phys. Rev. E* **85**, 066114 (2012).
 - [66] H.-W. Shen and X.-Q. Cheng, *J. Stat. Mech.* **10**, P10020 (2010).
 - [67] J. M. Hofman and C. H. Wiggins, *Phys. Rev. Lett.* **100**, 258701 (2008).
 - [68] V. Gudkov, V. Montealegre, S. Nussinov, and Z. Nussinov, *Phys. Rev. E* **78**, 016113 (2008).
 - [69] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, *Phys. Rev. E* **75**, 045102(R) (2007).
 - [70] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**, 881 (2002).
 - [71] B. Ball, B. Karrer, and M. E. J. Newman, *Phys. Rev. E* **84**, 036103 (2011).
 - [72] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2658 (2004).
 - [73] S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and A. Costa, *Phys. Rev. E* **85**, 046113 (2012).
 - [74] W. W. Zachary, *J. Anthropol. Res.* **33**, 452 (1977).
 - [75] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **80**, 056117 (2009).
 - [76] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008).
 - [77] A. L. N. Fred and A. K. Jain, in *Proceedings of the IEEE Computer Society Conference on Computer Vision Pattern Recognition*, Vol. 2 (IEEE Computer Society, 2003) pp. 128–133.
 - [78] A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. Fred, in *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference* (IEEE Computer Society, 2004) pp. 225–232.
 - [79] A. P. Topchy, A. K. Jain, and W. Punch, in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference* (IEEE Computer Society, 2003) pp. 331–338.
 - [80] D. Hu, P. Ronhovde, and Z. Nussinov, *Phys. Rev. E* **85**, 016101 (2012).
 - [81] J. Villain, R. Bidaux, J.-P. Carton, and R. Conte, *J. Phys. France* **41**, 1263 (1980).
 - [82] E. F. Shender, *Sov. Phys.* **56**, 178 (1982).
 - [83] C. L. Henley, *Phys. Rev. Lett.* **62**, 2056 (1989).
 - [84] Z. Nussinov, M. Biskup, L. Chayes, and J. van den Brink, *Europhys. Lett.* **67**, 990 (2004).
 - [85] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen, *Phys. Rev. Lett.* **92**, 178702 (2004).
 - [86] E. Mones, L. Vicsek, and T. Vicsek, *PLoS ONE* **7**, e33799 (2012).
 - [87] V. E. Krebs, *Connections* **24**, 43 (2002).