Local multiresolution order in community detection

Peter Ronhovde and Zohar Nussinov
Department of Physics, Washington University in St. Louis,
Campus Box 1105, 1 Brookings Drive, St. Louis, Missouri 63130, USA
(Dated: August 28, 2012)

Multi-scale ("multiresolution") community detection attempts to identify the most relevant divisions (groups of related nodes) of an arbitrary network over a range of network scales. This task is generally accomplished by analyzing community stability in an average sense across all communities in the network. In some systems, contending partitions of the global community structure may be vague or imprecisely defined, but certain local communities may nevertheless be strongly correlated at a given network resolution. We demonstrate a general local multiresolution method where we draw inferences about local community "strength" based on correlations between clusters in independently-solved systems. We propose measures analogous to variation of information and normalized mutual information which quantitatively identify the best resolution(s) at the community level. 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. It should, in principle, easily adapt to alternate 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].

This work extends current methods of "global" multiresolution CD [6] (see Appendix A) to enable quantitative multiscale evaluation at the *local* community level [7–9], effectively "zooming" inward or outward in the network scale depending on the specific node, region, or location (e.g., image segmentation applications [10]). Our local multiresolution replica algorithm (LMRA) quantitatively identifies the most natural resolution(s) for individual *communities* regardless of the weak or strong correlations present in the full network. In essence, the LMRA method is able to select optimal values of CD resolution parameter(s) for each cluster in a graph. Here, we solve independent copies of the full system, but the approach would adapt trivially to other CD algorithms which can identify local communities within network subgraphs (*i.e.*, without the need to partition the entire network) or to other local cluster comparison measures.

One of the most popular methods of CD 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 [11] 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 [12, 13], are also common.

Reichardt and Bornholdt (RB) wrote a Potts model [14] 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 or explicitly selecting the "correct" scale for a graph. These methods were shown to suffer from an inherent "resolution limit" [11, 14–17], which is not resolved by varying the network scale [18, 19], making it difficult for them to properly identify communities in large graphs.

More Potts model and related approaches include [6–8, 13, 20–23], and Refs. [8, 23] generalized the RB Potts models in [14, 24], respectively. Our previous work [6, 7] advanced a "local" Potts model, and local models were studied in more detail in [8]. Other local methods include [5, 7–9, 12, 25, 26], including variants of modularity [27, 28]. Potts systems in CD can experience disorder from thermal effects [29, 30], extraneous edges (noise) [7, 29–32], and system size [30, 33]. The selected model can also exacerbate disorder effects [31, 34].

Some CD methods implicitly select a single "objective" scale for a candidate community division (e.g., Refs. [11, 12]), but certain networks such as hierarchical systems inherently have multiple natural scales. Hierarchical clustering is an early multiscale method [35], but it *forces* hierarchical structure on every system without evaluating the relevance of the solved partitions. More recent hierarchical approaches include [36–41], and Ref. [42] relates the presence of hierarchical features to a scale-free-network property.

A CD algorithm should be able to determine all relevant scales of a network, ideally without *ad hoc* impositions on the network structure, and this problem is the

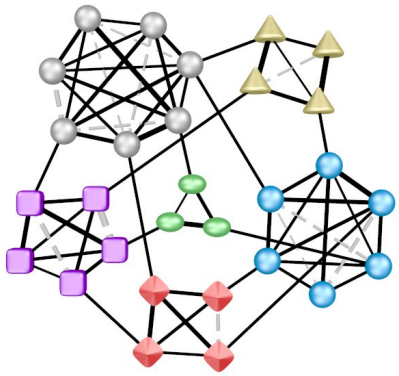


FIG. 1. (Color online) The figure illustrates a network partition where communities are represented by distinct node shapes and colors. The graph includes ferromagnetic [solid, black lines with $w_{ij} > 0$ in Eq. (1)] and antiferromagnetic interactions (gray, dashed lines with $u_{ij} > 0$), and the line thickness indicates the relative interaction strength. With Eq. (1), "neutral" interactions (unconnected or undefined relations) are repulsive in nature since they work like adversarial relations that break up well-defined communities.

impetus for developing quantitative multiresolution network analysis. Multiscale capable methods that utilize cost functions include [6, 14, 24, 25, 43–45]. The RB Potts model weighs the contribution of the null model [14], allowing the cost function to span different network scales. Other methods encompass varied forms of analysis [46–49] 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, 43, 50] 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, 25, 43]. Our multiresolution replica algorithm (MRA) calculated information-based correlations [6] 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 (see Appendices B and C) in a network, but the best local communities will not necessarily 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 [34, 51] depending on the network scale. Conversely, a global partition may be strong for most communities, but a given cluster may still be weakly defined.

We combine the benefits of multiresolution analysis with the local identification of community structure. While each community exists in the context of the surrounding network, we ideally prefer to identify strong communities independent of the global system, allowing each community to "stand on its own" in terms of the evaluation of community structure. Somewhat related efforts include detecting "unbalanced" communities in a network partition [52] and an efficient "seed-expansion" method by Havemann *et al.* [26] which could, in principle, be modified for other local cost functions.

The remainder of the work is organized as follows: we introduce our community detection Potts model in Sec. II. Section III A elaborates on concepts of community definitions, and Sec. III B describes the notion of a partition *resolution*. We suggest a local, community-based analogy to the variation of information (abbr., VI) and normalized mutual information (NMI) measures in Sec. IV which we apply in Sec. V for our local multiresolution algorithm. Section VI illustrates the approach with two examples, and we conclude in Sec. VII. Appendix A explains the context of local and global terminology used in this paper. Appendices B and C elaborate on our community detection and global multiresolution algorithms which form the basis of the local analysis presented in the current work. Finally, Appendices D and E comment the semi-metric property of our cluster measure and alternative approaches to local cluster comparisons in an information-theoretic analogy.

II. POTTS MODEL HAMILTONIAN

Regardless of the underlying solution method, the ultimate goal of any community detection partitioning algorithm is a Potts type assignment $i \to \sigma_i$ 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, 25, 26, 53]) 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 Appendix B) a general CD Potts model

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

which we refer as an "absolute" Potts model (APM) since it is not defined relative to a 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. As mentioned above, 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.

In Eq. (1), $\{w_{ij}\}$ and $\{u_{ij}\}$ are the edge weights for "cooperative" and "neutral" or "adversarial" relations, respectively. In unweighted graphs, $a_{ij} = b_{ij} = 1$. Both adversarial and neutral relations serve to break up community structure, so the APM [6, 7] penalizes neutral relations much like one would expect for adversarial relations (as opposed to zero energy contributions as in a purely ferromagnetic Potts model [12, 20]). This property avoids a trivial ground state solution (*i.e.*, a completely collapsed system) present in the purely ferromagnetic Potts model, providing an alternative "penalty function" to how modularity resolved the problem [11]. Ref. [23] generalized a common Potts model variant [14] to include "negative" link weights. A network *resolution* roughly corresponds to the typical community size, but it is better characterized by a typical community edge density (see Sec. III B). The global resolution parameter γ in Eq. (1) scales the relative effects of the ferromagnetic $\{w_{ij}\}$ and antiferromagnetic $\{u_{ij}\}$ interactions, effectively allowing the model to vary the network scale,

Despite the global energy sum, the model is a *local* measure of community structure (see Appendix A) because all node assignments are made strictly by evaluating local network parameters [7, 8]. 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.

III. 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 internally, in terms of the number or weight of edges, but those between communities are more sparsely connected. There is a question as to whether the "inner" versus "outer" degree comparison is summed across *all* external communities [54, 55] or is evaluated between *individual pairs* of communities [6, 7, 56].

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 their local group where 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 their local group. In many network approximations (e.g., the ubiquitous Zachary karate club network [57]), these "extra" edges are omitted as extraneous in a reduced-size network, but the additional "noise" induced by including these relations in a more comprehensive network should not intuitively disturb the natural communities provided they are strongly defined relative to any structure in the expanded system.

Ref. [54] 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 edge edges exceeds the sum of the external edges. A large social network, such as that mentioned above, 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 [54] neglect certain important (high noise) and intuitive [17, 52] cases.

Further, several CD methods compared by Lancichinetti and Fortunato [58] demonstrated that even weak communities as defined are not restrictive or characteristic of the capabilities of some CD algorithms. That is, the best methods easily solved the benchmark graphs [59] into regions where *all* nodes (on average) have more external than internal edges. 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 [7, 56]. With this in mind, it may be fruitful to pursue a community definition based on *edge density* as opposed to inner and outer community edge *counts*, but a 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 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 "perfect 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. [7] 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,

$$p_{\min} \geq \frac{\gamma}{\gamma+1}, \tag{2}$$

which we calculate 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 [8, 13, 14, 23–25, 43] which allow the models 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 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 avoids situations where a minimal number of connecting edges merge internally dense sub-graphs in order to arbitrarily satisfy the cost function. It also avoids resolution-limit-type effects by acting locally [7].

IV. INFORMATION MEASURES

Information measures have received broad acceptance for comparing candidate CD partitions. Commonly used measures include the variation of information [60] and normalized mutual information [61]. We leveraged the measures in Sec. IV A to identify the best global network scales via a multiresolution replica method [6] (see Appendices A and C).

A. Partition correlations

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

$$H(A) = -\sum_{k=1}^{q_A} \frac{n_k}{N} \log \frac{n_k}{N} \tag{3}$$

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. In practice for our application, contending partitions (A, B, ..., X) are defined as independent 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

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

where n_a (n_b) is the number of nodes in community a (b) of partition A (B). The variation of information $V(A,B)$ metric is then

$$V(A,B) = H(A) + H(B) - 2I(A,B) \tag{5}$$

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.

Some analysts prefer a normalized information measure [61] for partition similarity

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

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 D) on partitions A and B in part because $U(A,A) = 1$ not 0.

B. Local information analogies

In defining a cluster comparison measure, we wish to maintain consistency with the trend in CD towards information-theoretic partition evaluations. If we were to compare larger (multi-cluster) sub-graphs, a natural approach is to cut the subgraph from the whole network and compare the reduced-size partition. This breaks down at the cluster level because there is no partition-of-unity associated with an individual cluster as is used to define NMI or VI for CD.

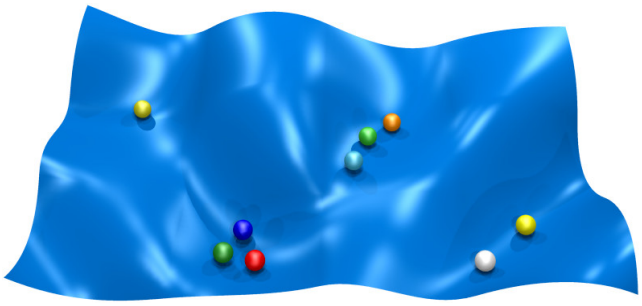


FIG. 2. (Color online) The figure schematically depicts r independent solvers ("replicas") as spheres navigating the energy landscape of Eq. (1). Stronger agreement among the replicas, as measured by information correlations in Sec. IV A, indicates a more accurate global solution. In this manuscript,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

Nevertheless, we can envision comparing *any* pair of clusters independent of the global system, but implementing an arbitrary measure is difficult in this context. Therefore, we consider the cluster embedded in the full system of N nodes, giving it a context for the resulting cluster-level entropy or information content based on the associated partition-of-unity probabilities. As will be evident below, strictly speaking we need not actually use the true size of the network for our cluster comparisons. That is, we could use some other $N' \neq N$, but it is conceptually appealing to evaluate a cluster in the context of the full network.

From Eq. (3), the entropy *contribution* of community a in partition A is

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

where n_a (n_b) 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)

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

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

$$v(a, b) \equiv H_a(A) + H_b(B) - 2I_{ab}. \tag{9}$$

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

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

Appendix E provides additional remarks.

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

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

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.

V. LOCAL MULTIRESOLUTION ALGORITHM

Our local multiresolution algorithm isolates relevant local multiresolution order (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 the basic MRA [6] algorithm given in Appendix C. The LMRA method depicted in Fig. 3 extends the MRA method by incorporating comparisons between specific clusters.

A. LMRA replica method

In general, clusters naturally change as the resolution is varied, so *how* do we identify the appropriate target clusters for comparison? Two natural approaches include: compare clusters for "nearby" resolutions as specified by a particular γ_i in Eq. (1) or compare targeted ("parent") clusters for specific node(s) of interest across the replicas. In the latter case, the node may be selected *a priori* based on a particular identity, or it may be randomly selected. One may also first analyze the global system and "work backwards" to identify relevant nodes as members of communities with interesting features.

In the first case, 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.

The latter option is used in the current work where we select a *node* of interest (e.g., a specific terrorist as in Sec. VI B), and trace the parent clusters among the replicas across a range of network scales [*i.e.*, different γ_i's in Eq. (1)]. 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.

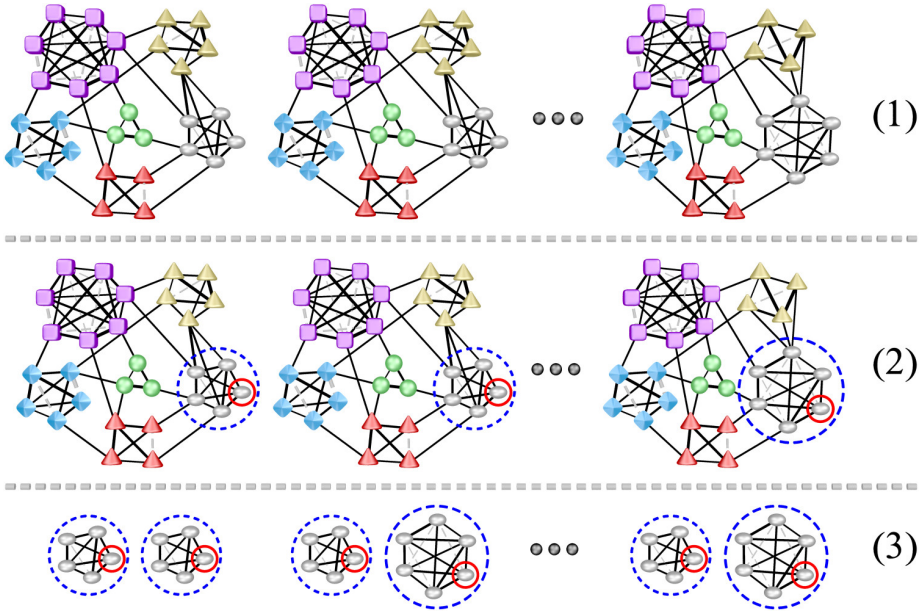


FIG. 3. (Color online) The figure illustrates our local multiresolution algorithm discussed in detail in Sec. V. The graphs include ferromagnetic ["cooperative" with $w_{ij} > 0$ in Eq. (1)] relations depicted by solid, black lines and antiferromagnetic ("neutral" or "adversarial" with $u_{ij} > 0$) interactions depicted by gray, dashed lines. The line thickness indicates the relative interaction strength, and we omit intercommunity adversarial and neutral relations for clarity. 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). Depending on the application, we could alternately calculate the correlations among *all* pairs of communities and determine whether the individual clusters are strongly or weakly defined. 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).

That is, at a given γ_i, we only need to know what cluster to which 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 $\overline{v}(a,b)$ or $\overline{u}(a,b)$ measures over the replica pairs can be compared *across* different γ_i's to evaluate the relative strength of the parent communities.

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. Select a set of nodes $\{a\}$ to track based on problem parameters (e.g., a person of interest in a terror network in Sec. VI B). Identify the set of resolutions $\{\gamma_i\}$ to analyze (often selected to sample all relevant network scales, see step 4 in Appendix C) 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 Appendix C.

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

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

$$\overline{S_i}(a,b) = \frac{2}{r(r-1)} \sum_{k>j} S_{ijk}(a,b) \tag{12}$$

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}.

As with the global MRA approach in Appendix C, we are interested in extrema or plateaus in the pertinent measures in Sec. IV. 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 Appendix C.

B. Alternative implementations

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

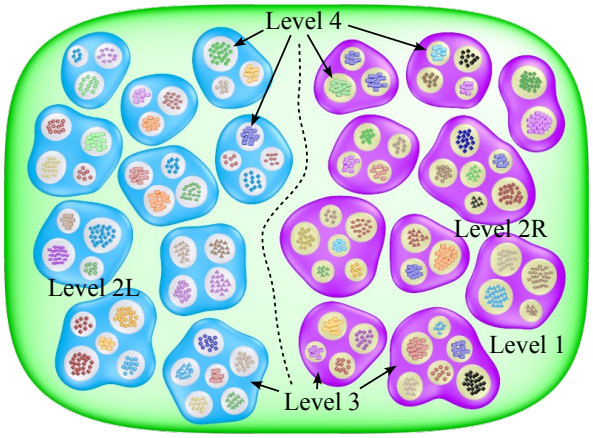


FIG. 4. (Color online) The figure depicts a constructed $N = 1024$ node four-level hierarchy. Level 1 is the complete network with two "sides" of supercommunities that are randomly connected at a low edge density between them. Level 2 consists of two roughly equal sized branches ($N_L = 502$ and $N_R = 522$) which we denote by "left" (L, blue or darker tone) and "right" (R, silver or medium tone) as the picture 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 the full system and 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. [26] for a more efficient method applied a different fitness function [25]). 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 in step (3*) is useful if there are no *a priori* nodes of interest to study. 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 matches the parent cluster d_{il} in partition B, repeating the process to the desired depth.

With this alternate 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).

VI. EXAMPLES

As discussed in Appendix C, we calculate the global MRA algorithm for the network and concurrently apply the LMRA algorithm in Sec. V to targeted nodes by tracking the respective parent clusters across a full range of relevant 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.

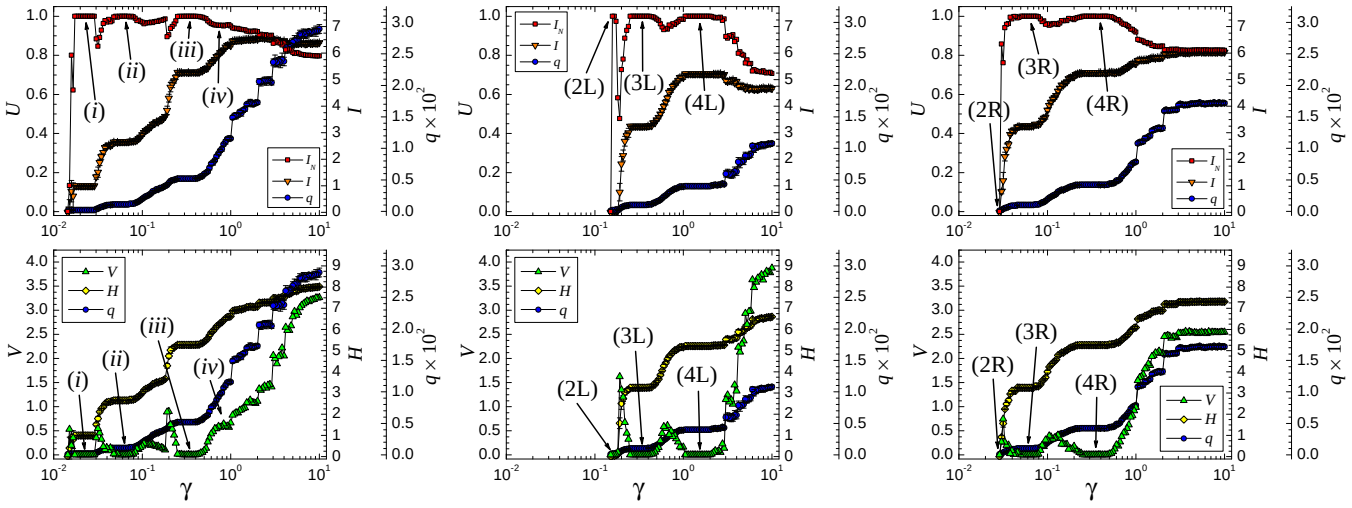
A. Branched hierarchy

We construct a branched, strict hierarchy as depicted in Fig. 4 which we use to test the LMRA method of Sec. V. 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 *inter*community density $p_1 = 0.015$. The approximate *intra*community edge densities at level 4 were $p_{4\mathrm{L}} = 0.9$ and $p_{4\mathrm{R}} = 0.6$ 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: $p_{3\mathrm{L}} = 0.37$, $p_{3\mathrm{R}} = 0.10$, $p_{2\mathrm{L}} = 0.16$, and $p_{2\mathrm{R}} = 0.03$. 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. [6] (summarized in Appendix C) 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. V can *fully extract this hidden section of the hierarchy.*

In Fig. 5, the left axes plot NMI, U, and VI, V, from Sec. IV A in the top and bottom sub-panels, respectively, averaged over all replica pairs. On the right axes, we

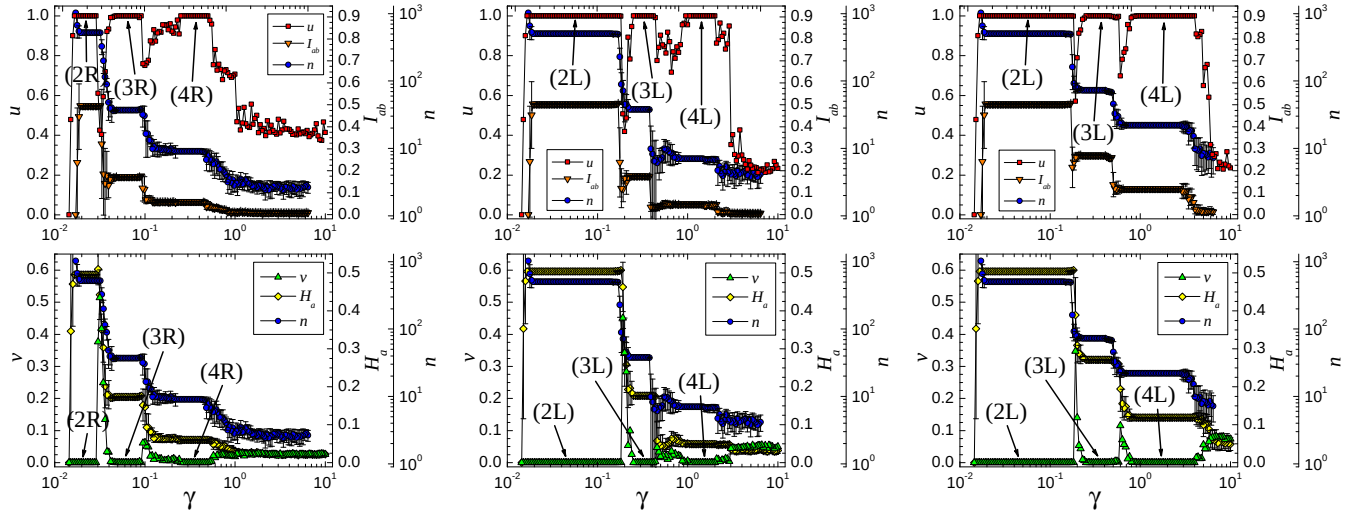


(a) Global MRA: Branch Hierarchy – Combined

(b) Global MRA: Branch Hierarchy – Left Side

(c) Global MRA: Branch Hierarchy – Right Side

FIG. 5. (Color online) In panels (a), we apply our global multiresolution algorithm (MRA, see Appendices A and C) 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. IV A) 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 [6]. 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. III B]. In Fig. 6, we compare parent *communities* using the local multiresolution algorithm in Sec. V where we demonstrate that the method can accurately extract level 4L for the targeted nodes.

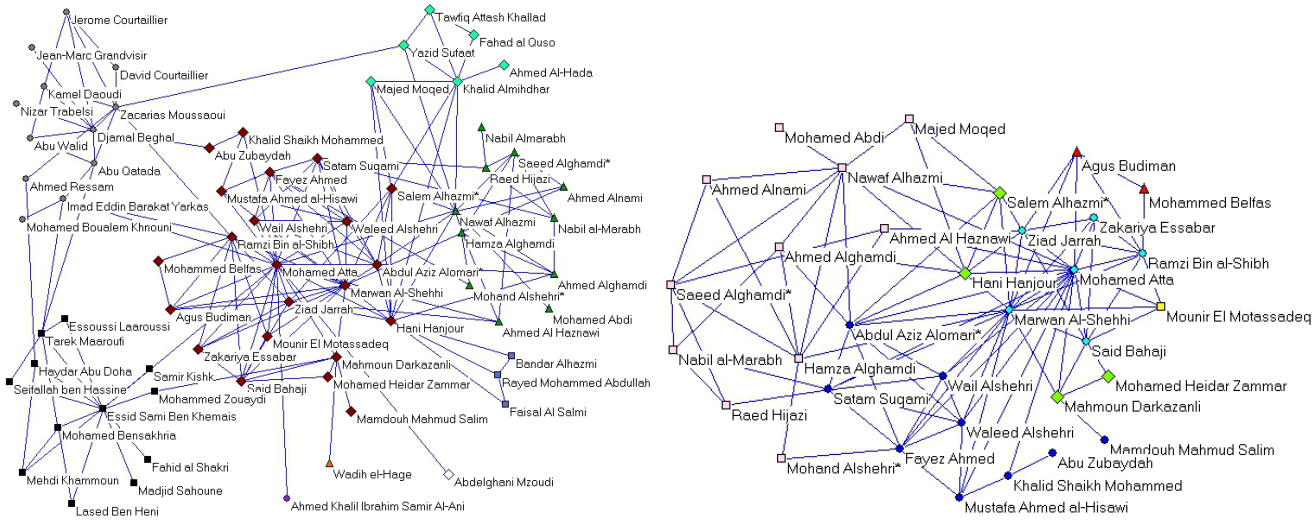


(a) Local MRA: Node 116 – Member of right side

(b) Local MRA: Node 661 – Member of right side

(c) Local MRA: Node 951 – Member of left side

FIG. 6. (Color online) In panels (a–c), we apply our local multiresolution algorithm (LMRA) in Sec. V to targeted nodes of the the $N = 1024$ node, four-level, "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. IV B) 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).



(a) Terrorist network – $\gamma = 0.1$

(b) Expanding network around Mohamed Atta

FIG. 7. (Color online) The figure depicts a small terrorist network collected from publicly available data [62]. 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.

plot the average mutual information I and the Shannon entropy H for 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).

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 between the independently-solved replica partitions. Specifically, feature (i) corresponds to level the 2 partition with $q_i = 2$, and feature (ii) identifies levels 2L and 3R with $q_{ii} = 11$ concurrently because of the respective community edge densities (see Sec. III B). Similarly, feature (iii) solves levels 3L and 4R with $q_{iii} = 52$. These particular 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.

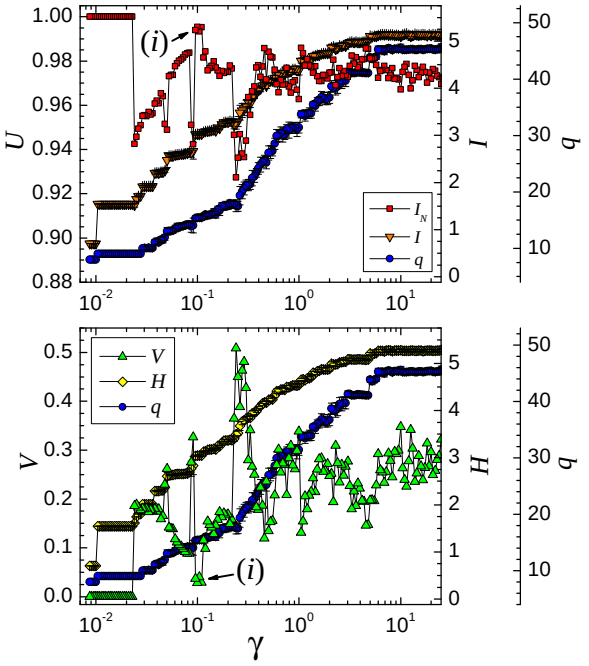
At feature (iv) in panel (a), the poor correlations show that *the global analysis of the full system misses level* 4L. This occurs because the well-defined local clusters conflict with more random partitions for the right-side subgraph in Fig. 4. In contrast, panels (b) and (c) show that the MRA method 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. IV B for the parent clusters of 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 relevant 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.

B. Small terrorist network

Even small networks can experience strongly-defined local clusters among indistinct global resolutions. We apply the LMRA method to a small terrorist network constructed from publicly available data [62]. Given that



(a) Global MRA: 9/11 Terrorists

FIG. 8. (Color online) We apply our multiresolution algorithm (see Appendices A and C) to a small terrorist network [62]. 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. V 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. IV B.

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 terror network.

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 settings are unimportant disjoint collapsed clusters. The left axes plot U and V (see Sec. IV A) for top and bottom sub-panels, respectively, averaged over all replica pairs. On the right axes, we plot I and H for 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 in Fig. 9(b) with $v(a, b) = 0$. 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.

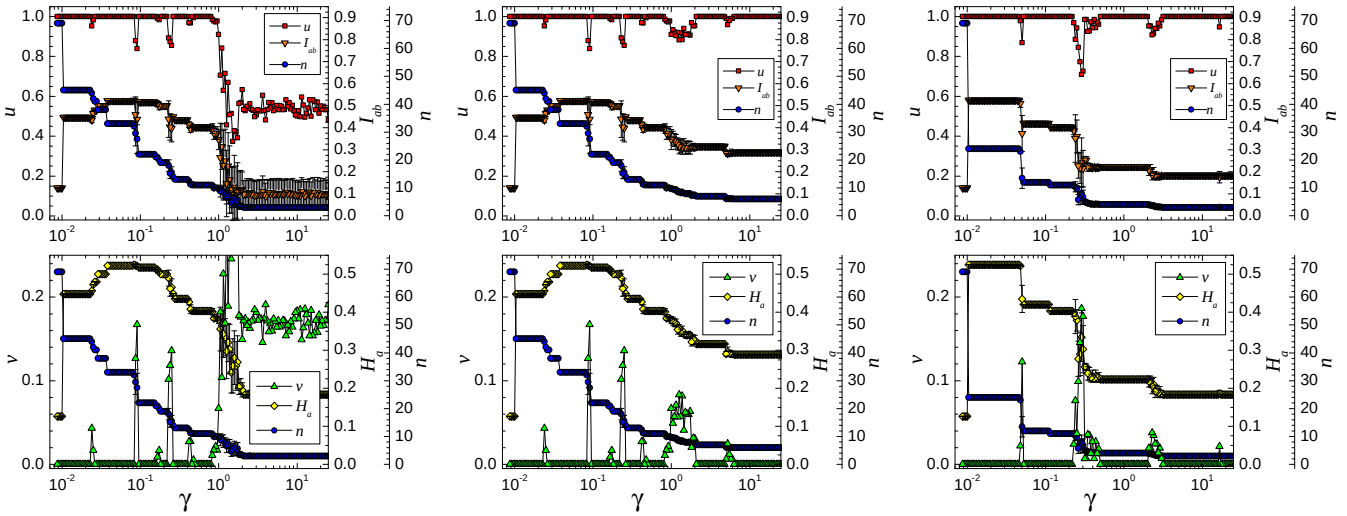
VII. CONCLUSION

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. 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 [6] to detect and quantitatively assess local multiresolution order. 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. We applied these measures to evaluate correlations among individual parent communities in multiple independent solutions (replicas), and we demonstrated that the proposed local multiresolution algorithm is able to extract local structure despite a blurred global multiresolution signature. Our approach is independent of the search algorithm or community detection model *making it suitable for use with any community detection 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



(a) LMRA: Hani Hanjour (b) LMRA: Mohamed Atta (c) LMRA: Zacarias Moussaoui

FIG. 9. (Color online) In each panel, we apply our *local* multiresolution algorithm (LRMA, see Sec. V) to a small terrorist network [62]. We analyze three selected terrorists 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. IV 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.

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 [11, 14]. 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* [6–8] or algorithms [12] 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 [7] where weakly-local cost functions may depend on the details of the community structure. Local *multiresolution methods*, such as the current work, seek to identify the best communities based on their strength at a given resolution. That is, the evaluation of the best resolution is not effectively "averaged" over all the communities in the graph, and each community may be strongly resolved at different network scales (often described in terms of distinct model weighting parameters).

Appendix B: Community detection algorithm

Our greedy CD algorithm 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 [7, 29, 30].

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 [7]. We can extend it with a stochastic heat bath [29] solver or a simulated annealing algorithm [14] at the cost of significantly increased computational effort.

Appendix C: *Global* multiresolution algorithm

As depicted in Fig. 2, our multiresolution algorithm iteratively applies the CD algorithm in Appendix B to quantitatively evaluate the best community partitions over a range of network scales. In its basic form, we independently solve the CD problem for a given graph over a range of γ in Eq. (1) and evaluate the average strength of the partition correlations. This process quantitatively estimates the robustness of the best solution(s) by sampling the complexity of the energy landscape.

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. IV 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 [12, 63, 64], particularly in the latter case of weak solutions [65].

As a function of the resolution parameter γ in Eq. (1) (or any relevant CD scale parameter for another model [14, 43]), the best resolutions may be identified by peaks or plateaus in NMI [6], minima or plateuas in VI [6, 44], and/or plateaus in the number of clusters q [43] or other measures [6, 44]. 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 [6]. 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. [10] also analyzed *peaks* in VI to perform image segmentation using CD concepts).

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 Appendix B.

(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. IV A.

(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) [6] (see Appendix A) based upon regions of γ with high NMI or low VI. Plateaus in I and q may also provide supplemental information regarding partition stability. The solution cost scales linearly in r with the CD algorithm in Appendix B, $O(rtL^{1.3}\log k)$. We have solved systems with $O(10^7)$ edges on a single processor [6] in a few hours.

The algorithm may detect, but *does not impose*, a strictly *hierarchical* community structure. That is, as shown in Sec. VI 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 somewhat preferable over *forcing* a hierarchical structure on every analyzed network [35] 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 [66, 67].

Appendix D: 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 – $S(a,b) = 0$ 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 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 only if $a = b$. From Eq. (9)

$$\begin{aligned} v(a,b) &= -\frac{n_a}{N}\log\left(\frac{n_a}{N}\right) - \frac{n_b}{N}\log\left(\frac{n_b}{N}\right) \\ &\quad -2\frac{n_{ab}}{N}\log\left(\frac{n_{ab}N}{n_a n_b}\right) \\ &= \frac{n_a - n_{ab}}{N}\log\left(\frac{N}{n_a}\right) + \frac{n_b - n_{ab}}{N}\log\left(\frac{N}{n_b}\right) \\ &\quad +\frac{n_{ab}}{N}\log\left(\frac{n_a}{n_{ab}}\right) + \frac{n_{ab}}{N}\log\left(\frac{n_b}{n_{ab}}\right) \\ v(a,b) &\geq 0 \end{aligned} \tag{D1}$$

since $n_a > 0$, $n_b > 0$, $n_{ab} \geq 0$, $n_a \geq n_{ab}$, and $n_b \geq n_{ab}$. Furthermore, it is zero only when, $n_a = n_b = n_{ab}$. That is, it is zero when $a = b$.

(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). Symmetry of $v(a,b)$ is then immediately obvious.

Thus, CVI is a semi-metric. □

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 E: Alternate cluster measures

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

$$\begin{aligned} V(A,B) &= H(A|B) + H(B|A) \\ &= \sum_{a,b}\left[\frac{n_{ab}}{N}\log\frac{n_b}{n_{ab}} + \frac{n_{ab}}{N}\log\frac{n_a}{n_{ab}}\right]. \end{aligned} \tag{E1}$$

From this equivalent variant of VI, the natural CVI definition would be

$$v'_{ab}(A,B) = \frac{n_{ab}}{N}\log\frac{n_a}{n_{ab}} + \frac{n_{ab}}{N}\log\frac{n_b}{n_{ab}}. \tag{E2}$$

Unlike CVI in Eq. (9), Eq. (E2) has the nice property that the individual cluster contributions sum to VI, $V(A,B) = \sum_a^{q_A}\sum_b^{q_B} v(a,b)'$.

Unfortunately, this particular launching point does not work for cluster comparisons. While $v'_{aa}(A,A) = 0$ as desired, it is also the case that $v'_{ab} = 0$ 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 $v(a,b)'' = H_a(A)/q_B + H_b(B)/q_A - 2I_{ab}(A,B)$. This variant would again yield the desirable property $V(A,B) = \sum_a^{q_A}\sum_b^{q_B} v(a,b)''$, but the measure loses the semi-metric requirements $v(a,b)'' \geq 0$ and $v(a,a)'' = 0$.

[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] P. Ronhovde and Z. Nussinov, Phys. Rev. E **80**, 016109 (2009).
[7] P. Ronhovde and Z. Nussinov, Phys. Rev. E **81**, 046114 (2010).
[8] V. A. Traag, P. Van Dooren, and Y. Nesterov, Phys. Rev. E **84**, 016114 (2011).
[9] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).
[10] D. Hu, P. Ronhovde, and Z. Nussinov, Phys. Rev. E **85**, 016101 (2012).
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[12] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).
[13] M. J. Barber and J. W. Clark, Phys. Rev. E **80**, 026129 (2009).
[14] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[15] S. Fortunato and M. Barthélemy, Proc. Natl. Aca. Sci.

U.S.A. **104**, 36 (2007).
[16] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Euro. Phys. J. B **56**, 41 (2007).
[17] X. S. Zhang, R. S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang, and L. Chen, Europhys. Lett. **87**, 38002 (2009).
[18] A. Lancichinetti and S. Fortunato, Phys. Rev. E **84**, 066122 (2011).
[19] J. Xiang and K. Hu, Physica A **391**, 4995 (2012).
[20] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[21] I. Ispolatov, I. Mazo, and A. Yuryev, J. Stat. Mech. **09**, P09014 (2006).
[22] M. B. Hastings, Phys. Rev. E **74**, 035102 (2006).
[23] V. A. Traag and J. Bruggeman, Phys. Rev. E **80**, 036115 (2009).
[24] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[25] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. **11**, 033015 (2009).
[26] F. Havemann, M. Heinz, A. Struck, and J. Gläser, J. Stat. Mech. **01**, P01023 (2011).
[27] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[28] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[29] D. Hu, P. Ronhovde, and Z. Nussinov, Phil. Mag. **92**, 406 (2012).
[30] D. Hu, P. Ronhovde, and Z. Nussinov, e-print arXiv:1204.4167 (2012).
[31] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
[32] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[33] P. Ronhovde, D. Hu, and Z. Nussinov, EPL **99**, 38006 (2012).
[34] L. Danon, A. Díaz-Guilera, and A. Arenas, J. Stat. Mech. **11**, P11010 (2006).
[35] B. Everitt, S. Landau, and M. Leese, *Cluster analysis* (2001).
[36] A. Clauset, C. Moore, and M. E. J. Newman, Nature **453**, 98 (2008).
[37] M. Rosvall and C. T. Bergstrom, PLoS ONE **6**, e18209 (2011).
[38] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Aca. Sci. U.S.A. **104**, 15224 (2007).
[39] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. **10**, P10008 (2008).
[40] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, Physica A **388**, 1706 (2009).
[41] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761764 (2010).
[42] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[43] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. **10**, 053039 (2008).
[44] D. J. Fenn, M. A. Porter, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones, Chaos **19**, 033119 (2009).
[45] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnella, Science **328**, 876 (2010).
[46] J. Zhang, K. Zhang, X. ke Xu, C. K. Tse, and M. Small, New J. Phys. **11**, 113003 (2009).
[47] X.-Q. Cheng and H.-W. Shen, J. Stat. Mech. **04**, P04024 (2010).
[48] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, PLoS ONE **6**, e18961 (2011).
[49] H.-W. Shen, X.-Q. Cheng, and B.-X. Fang, Phys. Rev. E **82**, 016114 (2010).
[50] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Fluct. Noise Lett. **7**, L209 (2007).
[51] H.-W. Shen, X.-Q. Cheng, and B.-X. Fang, Phys. Rev. E **82**, 016114 (2010).
[52] S. Zhang and H. Zhao, Phys. Rev. E **85**, 066114 (2012).
[53] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 036103 (2011).
[54] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[55] S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and A. Costa, Phys. Rev. E **85**, 046113 (2012).
[56] R. K. Darst, P. Ronhovde, and Z. Nussinov, (in preparation) (2012).
[57] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[58] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
[59] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[60] M. Meilă, J. Multivariate Anal. **98**, 873 (2007).
[61] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech. **09**, P09008 (2005).
[62] V. E. Krebs, Connections **24**, 43 (2002).
[63] 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.
[64] 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.
[65] 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.
[66] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen, Phys. Rev. Lett. **92**, 178702 (2004).
[67] E. Mones, L. Vicsek, and T. Vicsek, e-print arXiv:1202.0191 (2012).