Local resolution-limit-free Potts model for 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: April 19, 2010)

We report on an exceptionally accurate spin-glass-type Potts model for community detection. With a simple algorithm, we find that our approach is at least as accurate as the best currently available algorithms and robust to the effects of noise. It is also competitive with the best currently available algorithms in terms of speed and size of solvable systems. We find that the computational demand often exhibits superlinear scaling $O(L^{1.3})$ where L is the number of edges in the system, and we have applied the algorithm to synthetic systems as large as 40×10^6 nodes and over 1×10^9 edges. A previous stumbling block encountered by popular community detection methods is the so-called "resolution limit." Being a "local" measure of community structure, our Potts model is free from this resolution-limit effect, and it further remains a local measure on weighted and directed graphs. We also address the mitigation of resolution-limit effects for two other popular Potts models.

PACS numbers: 89.75.Fb, 64.60.Cn, 89.65.–s

I. INTRODUCTION

"Community detection" describes the problem of finding closely related sub-groups within a network. Applications of the problem are extremely broad finding realizations in the World Wide Web [1, 2], food webs [3], social networks [3], protein interactions [4], consumer purchasing patterns [5], mobile phone networks [6], criminal networks [7], epidemiology [8], biological networks [3, 9], and other areas. A recent review of community detection appears in Ref. [10].

In this paper, we present an improvement to the Potts model as applied to community detection, and we demonstrate that it is extremely accurate, robust to noise, and competitive with the best available methods in terms of computational speed and the size of solvable systems. Our approach also corrects known problems encountered by other popular measures that do not properly resolve small communities when large sets of data are examined.

The data in networks can often be cast as a graph consisting of members represented by nodes with pair-wise relationships between the nodes represented by edges. In general, these relationships can be specified in one direction along an edge, and they can be weighted or unweighted. The goal of community detection is to efficiently separate clusters of closely related nodes from each other. Each cluster will have a proportionally higher number of internal edges compared its external connections to each other community in the partition.

The most popular quantitative community measure is that of "modularity" which was originally introduced by Newman and Girvan [11]. This measure constituted a work that transformed the field. Modularity measures the deviation of a proposed community structure compared to what is expected from an "average" case based on a particular random distribution (a "null model").

Our approach is a physics-inspired method that casts community detection as a Potts model spin glass. Communities correspond to Potts model spin states, and the associated system energy indicates the quality of a candidate partition. Some earlier approaches utilizing Potts models are in [12] and [13]. Our particular model was originally inspired by a minimal cut method by Djidjev [14] which is equivalent to modularity. The resulting generalized Hamiltonian was previously presented by Reichardt and Bornholdt (RB) [15]. In their specific implementation, RB generalized null model based approaches to community detection, including modularity as a special case, and elaborated on the connection between physics and community detection. Two other Potts model approaches are by Hastings [16] and Ispolatov *et al.* [17].

Modularity and the RB Potts model (RBPM) utilize a random null model selected to evaluate the strength of a proposed community partition. Larger deviations (more intra-community links and fewer inter-community links) from the random case indicate better community structure. A null model is usually based on parameters of the graph being examined which allows the measure to "scale" to arbitrary graphs in an objective manner (see Sec. VI A). Typical null models for the RBPM are: (*i*) an Erdős-Rényi null model (RBER) in which all edges are equally likely to be connected and (*ii*) the configuration null model (RBCM) in which edge connection probabilities are based on the current graph's degree distribution. For modularity, the dependence on the null model is inherent to the definition of the measure. Within the RB scheme, the dependence on a null model is introduced by design.

Fortunato and Barthélemy [18] later determined that modularity optimization can result in incorrect community divisions due to a *resolution limit*. The resolution limit is an inherent scaling in the expected number of communities q which roughly scales as $\sqrt{L}$ where L is the total number of edges in the graph. The RBPM model is also subject to a resolution limit [19] due to how it is cast by design, analogous to modularity, in terms of an arbitrary null model comparison. The number of commu-

nities roughly scales as $\sqrt{\gamma_{RB} L}$, where γ_{RB} is a weight applied to the null model comparison. Optimizing either measure (maximizing modularity or minimizing the Potts model energy) tends to merge small clusters in large systems, or it may incorrectly partition large communities. Although the RBPM allows for an arbitrary choice of null model, the resolution limit was shown to persist [19] regardless of the null model that is used.

Our approach avoids a null model comparison [20]. Instead, it penalizes for missing edges directly in the energy sum [21]. In effect, a community is defined by its edge density as opposed to allowing each graph to independently define a community through the use of a null model. One consequence is that it removes the ability of the model to automatically scale the solution based on global properties of a graph (see Sec. VI A), but the change results in a robust model with significant improvements to several desirable properties.

In this paper, we demonstrate a simple but effective implementation of the q-state Potts model to community detection. In Sec. II, we discuss our Potts model and some of its properties along with the RBPM and its main variants. We also explain the concept of the *resolution* of a partition. In Sec. III, we present our algorithm, and Sec. IV illustrates its accuracy compared to several other approaches. Our Potts model and the RBCM model are directly compared in Sec. V. Issues regarding local and global measures and the resolution limit for general graphs are addressed in Sec. VI. We solve two examples in Sec. VII and conclude in Sec. VIII. Appendix A argues that the *unweighted* variant of the RBER model can be strengthened to eliminate the resolution limit. In Appendix B, we explain the variation of information (VI) metric which we use in Sec. V. Appendices C and D elaborate on details related to the benchmark in Sec. V B. Appendix E explains a *community detection transition* observed in the same benchmark.

II. POTTS MODEL HAMILTONIANS

Quality functions in community detection evaluate the "best" partitions based on at least two criteria: Edges inside a community strengthen the community. In order to consistently avoid a trivial solution (a single community), a quality function must also apply a "penalty function." The most common method compares community degree distributions to an "expected" distribution based on a null model, but other approaches are possible.

A. Absolute Potts model

Our Potts model directly penalizes for missing edges within a community. The result is a robust model that is highly accurate, a local model for general graphs (weighted, unweighted, and directed), and *free of the resolution limit*. We also connect the introduced model weight γ to the *resolution* of a system and relate the interaction energies to the stability of communities.

1. Hamiltonian

We construct the Potts model with the following considerations. Edges *inside* communities and missing edges *outside* communities are both favorable for a well-defined community structure, so the energy of the system is lowered by these arrangements. The opposite holds for edges outside communities and missing edges inside communities. This generalized Potts Hamiltonian is [15]

$$\mathcal{H}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\big(a_{ij}A_{ij} - b_{ij}J_{ij}\big)\big[2\delta(\sigma_i,\sigma_j) - 1\big] \quad (1)$$

where $\{A_{ij}\}$ is the set of adjacency matrix elements: $A_{ij} = 1$ if nodes i and j are connected and is 0 if they are unconnected, and $J_{ij} \equiv (1 - A_{ij})$. The edge weights ($\{a_{ij}\}$ and $\{b_{ij}\}$) and connection matrices ($\{A_{ij}\}$ and $\{J_{ij}\}$) are defined by the system. The Potts spin variable σ_i takes an integer value in the range $1 \leq \sigma_i \leq q$ which designates the community membership of node i (node i is in community k if $\sigma_i = k$). The number of communities q can be set as a constraint, or it can be determined from the lowest energy configuration. The Kroneker delta $\delta(\sigma_i, \sigma_j) = 1$ if $\sigma_i = \sigma_j$ and 0 if $\sigma_i \neq \sigma_j$.

The spin glass type Potts model of Eq. (1) can be reduced, up to an additive constant, to a form that greatly simplifies implementation. For comparison, we introduce a form similar in appearance to the notation used by RB

$$\mathcal{H}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\big(a_{ij}A_{ij} - \gamma b_{ij}J_{ij}\big)\delta(\sigma_i,\sigma_j). \quad (2)$$

Spins interact only with other spins in the same community ($\sigma_i = \sigma_j$). The introduced factor of γ allows the model to scale the relative strengths of the connected and missing edge weights. The generality of the weights ($\{a_{ij}\}$ and $\{b_{ij}\}$) [20, 22] enables the study of directed graphs, weighted graphs, and graphs with missing link weights (*i.e.*, levels of "repulsion"). Traag and Bruggeman [23] also presented a generalization of the RBCM that similarly allows for "negative" link weights. Unweighted graphs use edge weights of $a_{ij} = b_{ij} = 1$.

The Hamiltonian of Eq. (2) describes a system wherein spins in the same community interact ferromagnetically if they are connected and antiferromagnetically if they are not connected. We identify communities by minimizing Eq. (2), and despite a global energy sum, our model is a *local* measure of community structure (see Sec. VI). We refer to Eq. (2) as an *"absolute" Potts model* (APM) as it is not defined relative to a null model. Although our analysis in this paper will focus on the static APM, it is defined for both *static* systems and *dynamic* networks with time-dependent weights and adjacency matrices.

2. Resolution

Intuitively, the *resolution* of a community partition is set, on average, by the strength of intra-community connections. That is, the resolution of the partition may be specified by the typical edge density of the communities within the partition. Communities with substantially different edge densities have different qualitative features.

In social networks for example, a partition intending to convey the "close friends" within a network would intuitively have a higher typical edge density than a partition that includes all "acquaintances" since the disparate acquaintances are much less likely to know each other. Ideally, a partition should contain communities that convey similar qualitative information (*i.e.*, similar "levels" of association). In practice, it will contain communities with different edge densities, but intuitively the differences would not be drastic for a given resolution.

For unweighted graphs, the edge density p_s of community s is $p_s = \ell_s/\ell_s^{\max}$ where ℓ_s is the number of edges in the community. $\ell_s^{\max} = n_s(n_s-1)/2$ where n_s is the number of nodes. The model weight γ in Eq. (2) is related to the *minimum* edge density of each community,

$$p_{\min} \geq \frac{\gamma}{\gamma+1}, \tag{3}$$

which is determined by calculating the minimum community density that gives an energy of zero or less. Alternately, we can use an inductive argument based on the maximum intercommunity edge density that causes two arbitrary communities to merge. For weighted graphs, we define a "weight density" $p_s \equiv w_s/w_s^{\max}$ where w_s is the sum of all weighted edges in community s and the "maximum weight" $w_s^{\max} \equiv \overline{w}_s \ell_s^{\max}$ where $\overline{w}_s$ is the average edge weight. The minimum density is $p_{\min} \geq \gamma/(\gamma+\overline{w}_s/\overline{u}_s)$ where $\overline{u}_s$ is the average weight of the missing links. Without γ, the model is restricted to solving one particular resolution of a system. This relation between γ and the community density is distinctly different from a resolution limit because the communities are determined through only *local* constraints (see Sec. VI).

3. Community and node stability

From Eq. (2), the interaction energy E_{rs} between communities r and s is

$$E_{rs} = -w_{rs} + \gamma u_{rs} \tag{4}$$

where w_{rs} is the energy sum over all edges and u_{rs} is the energy sum over all *missing* links strictly *between* the two communities. $E_{ss} \equiv E_s$ is the internal energy of community s where the energy sum is over all *internal* edges and missing links. When $E_s \simeq 0$, the assignment of community s is more sensitive to local perturbations.

Similarly, the interaction energy E_{ri} of node i with community r is given by Eq. (4). If $E_{si} - E_{ri} \simeq 0$ for node i in community s, then the node is susceptible to displacement by system perturbations. When a node contributes a large fraction of the energy E_s of its own community, the community is susceptible to disruption if the node is moved. Equation (4) indicates the strong local behavior of the APM (see Sec. VI). For general graphs, the interaction energy of node i or community s is measured *only* by its *own* edges or missing links with each community.

B. RB Potts models

We compare the APM to the RBPM in order to demonstrate improvements in accuracy and locality despite the apparent similarity in the models. The RBPM, using an arbitrary null model, is defined as [15]

$$\mathcal{H}_{RB}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{RB}p_{ij}\right)\delta(\sigma_i,\sigma_j) \tag{5}$$

where we include the overcounting scale factor of $1/2$. The term p_{ij} is the probability that nodes i and j are connected, and it incorporates the dependence on the arbitrary null model. γ_{RB} is the weight applied to the null model. The most frequently used null models are an Erdős-Rényi null model and the configuration null model (see Sec. I). For later reference, they are explicitly given by $p_{ij} = p$ for the Erdős-Rényi null model

$$\mathcal{H}_{RB}^{ER}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{RB}p\right)\delta(\sigma_i,\sigma_j) \tag{6}$$

and by $p_{ij} = k_ik_j/(2L)$ for the configuration null model

$$\mathcal{H}_{RB}^{CM}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{RB}\frac{k_ik_j}{2L}\right)\delta(\sigma_i,\sigma_j), \tag{7}$$

where k_i is the degree of node i. Equation (7) appears to be the more preferred model since the configuration null model incorporates information about the degree distribution of the graph under consideration.

When $\gamma_{RB} = 1$, the RBCM of Eq. (7) is equivalent to modularity [15] up to a scale factor of $-1/L$. The APM can be made equivalent to the RBER model for *unweighted* graphs [24] (see also Appendix A). We address weighted generalizations of both models and their effect on model locality in Sec. VI E. Despite the similar forms of the Hamiltonians of Eqs. (2) and (5), the model weights γ and γ_{RB} perform distinctly different roles in the two models. In the APM, γ directly adjusts the weight applied to *missing edges*. In the RBCM, γ_{RB} adjusts the weight applied to the *null model*. We contrast the accuracy of the APM and the RBCM in Sec. V.

III. ALGORITHM

Our algorithm moves nodes by identifying which community they may be moved into so that the system energy

is lowered. The algorithm proceeds until no more node moves are possible. This "orthogonal steepest descent" algorithm (selecting the path of steepest descent for only one spin σ_i at a time) is extremely fast. We introduced our initial implementation of the algorithm in [20]. A summary of the efficiency of several algorithms appears in [25]. A number of algorithms were compared in [26] and [27] where algorithms similar to ours performed very well when optimizing modularity. Combined with the APM, it is exceptionally accurate. The steps of the algorithm are:

(1) *Initialize the system.* Initialize the connection matrices (A_{ij} and J_{ij}) and edge weights (a_{ij} and b_{ij}). The system begins in a "symmetric" state wherein each node forms its own individual community ($q_0 = N$). If the number of communities q is constrained (e.g., Figs. 1 and 9), we randomly initialize the system into $q_0 = q$ communities.

(2) *Optimize the node memberships.* Sequentially "pick up" each node and scan its neighbor list. Calculate the energy change as if it were moved to each connected cluster. Immediately place it in the community with the lowest energy (optionally allowing zero energy changes). Each iteration through all nodes is $O(L)$.

(3) *Iterate until convergence.* Repeat step (2) until an energy minimum is reached where no node moves will further lower the system energy.

(4) *Test for a local energy minimum.* Manually merge any connected communities if the merge(s) will further lower the energy of the system. If any merges are found, return to step (2) for any additional node-level refinements. We estimate that the computational cost is $O(L \log q)$ which is generally smaller than the node optimization cost in steps (2) and (3).

(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.

The symmetric initialization for the nodes in step (1) is not uncommon in the literature [6, 28–30]. Steps (2) and (3) are the fundamental elements of the algorithm which are similar to portions of algorithms used elsewhere [6, 29]. The number of iterations is generally $O(10)$ for large systems, but it can be higher for "hard" problems. In step (4), the community merge test is sometimes necessary because certain configurations, particularly heavily weighted graphs with $\gamma \ll 1$, more easily trap the node-level refinements [steps (2) and (3)] in local energy minima. The merge test is not generally a major concern for $\gamma \geq 1$.

The order of node moves is significant, so the additional trials in step (5) sample different regions of the energy landscape and can yield different solutions even with the symmetric initialization in step (1). We optimize solutions by increasing the number of trials where the greatest benefit occurs for problems of "intermediate" difficulty (e.g., see the data for the APM in Fig. 1). The number of trials t is generally $O(10)$ or less.

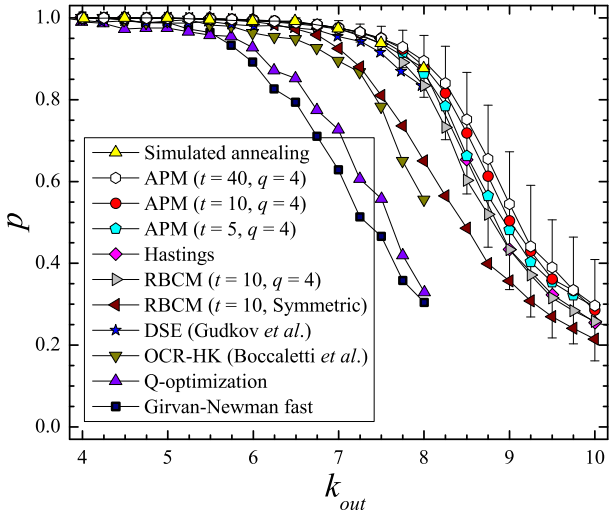

FIG. 1: (Color online) Plot of the percentage of correctly identified nodes p versus the average external degree k_{out} [32]. The average node degree is $k = 16$. The data for the APM of Eq. (2) and the RBCM of Eq. (7) both use $\gamma = \gamma_{RB} = 1$. Both models use the algorithm in Sec. III with $q = 4$ by constraint (see text regarding the RBCM/Symmetric data). The APM is at least as accurate as SA (error bars are for $t = 10$ optimization trials), and the RBCM performs excellently also. Each point is an average over 500 runs.

Empirically, the overall solution cost often scales as $O(tL^{1.3} \log k)$ where k is the average node degree. The factor of $\log k$ applies for large sparse matrix systems. The algorithm can accurately scale to at least $O(10^7)$ nodes and $O(10^9)$ edges with a calculation time of several hours [31] (see Sec. VII).

IV. ACCURACY COMPARED TO OTHER ALGORITHMS

We test the accuracy of our method compared to several other algorithms using a common benchmark [32]. The benchmark is very small by current standards with an unrealistically symmetric community structure, but its frequent use provides a means of comparing the accuracy of various algorithms that have been presented in the literature over time. The problem defines a system of $N = 128$ nodes in $q = 4$ clusters of $n = 32$ nodes each. Each node is assigned an average of $k = 16$ edges of which k_{in} are randomly assigned inside its own community. k_{out} edges are randomly assigned to nodes in other communities such that $k = k_{in} + k_{out}$. We then attempt to verify the defined community structure.

In Fig. 1, we plot the "percentage" of correctly identified nodes p as a function of k_{out}. For consistency with other data in Fig. 1, we use the same measure of percentage accuracy as Newman [32]. We use $q = 4$ communities by constraint and test several levels of optimization ($t = 5$, 10, and 40). At $t = 10$, our method maintains an

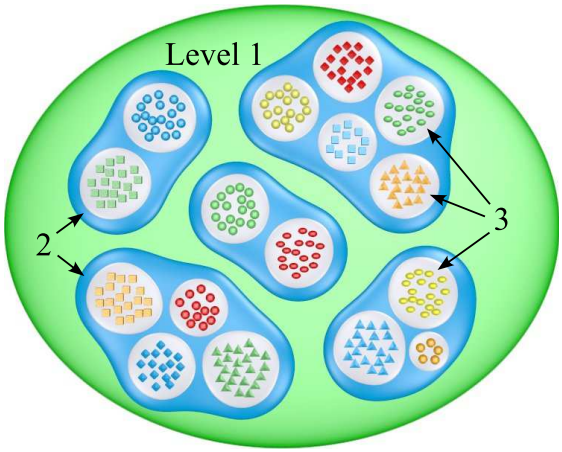

FIG. 2: (Color online) The figure depicts a simulated three-level heterogeneously-sized hierarchy with $N = 256$ nodes [22, 35]. The innermost level 3 has $q_3 = 16$ communities with a randomly assigned average density of $\overline{p}_3 = 0.90$. The intermediate level 2 has $q_2 = 5$ communities and an average density of $\overline{p}_2 = 0.47$ that is constructed by connecting the constituent level 3 sub-groups at an intercommunity edge density of $p_2 = 0.3$. Level 1 is the completely merged system with an average density of $\overline{p}_1 = 0.18$, and it is constructed by connecting nodes in different level 2 communities with an intercommunity edge density of $p_1 = 0.1$.

accuracy rate of 95% or better up to $k_{out} = 7.5$.

Several sets of data were assimilated by Boccaletti *et al.* [33] where the most accurate algorithm was simulated annealing (SA) although it is computationally demanding [25]. Other accurate data are by Hastings [16], Gudkov *et al.* [30], and our algorithm in Sec. III applied to the RBCM with $\gamma_{RB} = 1$ (modularity) and $t = 10$. Our algorithm is as accurate as SA when used with the APM.

The APM, one set of our data for the RBCM, and the data by Hastings impose $q = 4$ as a constraint; so using a constrained q may affect the accuracy rate in this problem. The *initial state* of the system substantially influences the accuracy of our algorithm for the RBCM when q is unconstrained and when starting from an initial state of one node per cluster (symmetric) or a random state (not depicted) [34]. A recent analysis [27] showed our multiresolution algorithm [22] applied to this benchmark using the APM with unconstrained q where it was also very accurate, among the best of tested algorithms.

V. ACCURACY COMPARISON OF POTTS MODELS

We compare the APM of Eq. (2) to the RBCM of Eq. (7) with two test systems. First, we solve for the different levels of the synthetic hierarchy depicted in Fig. 2 with the results given in Fig. 3. Second, we create a set of strongly defined systems with high community edge densities and increasing levels of noise. A sample graph is depicted in Fig. 4 with the results given in Figs. 5 and 6.

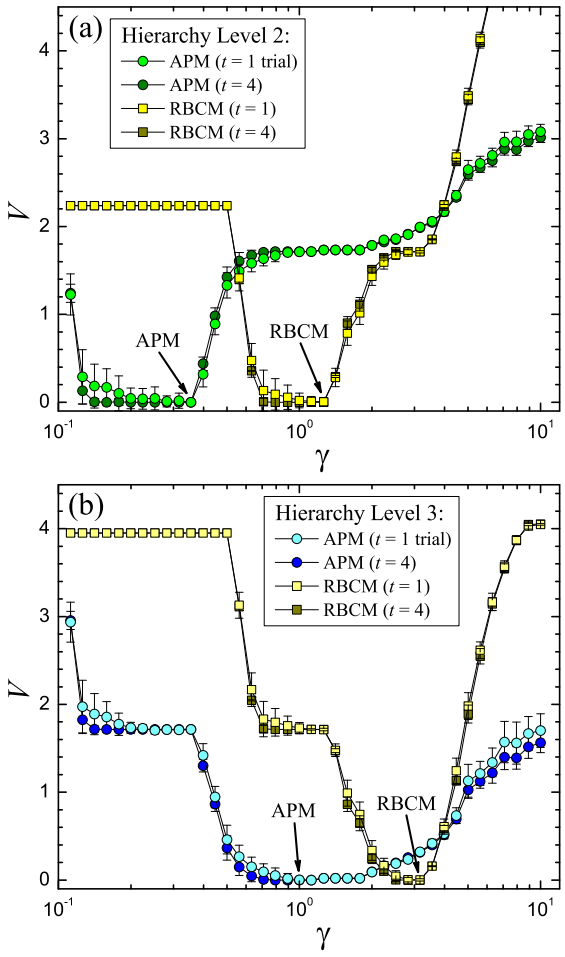

FIG. 3: (Color online) Plot of VI V vs model weights γ or γ_{RB} for the APM of Eq. (2) and the RBCM (configuration null model) of Eq. (7), respectively. The plots illustrate how the model weights operate in the respective models. We use the algorithm in Sec. III for both models to identify the hierarchy depicted in Fig. 2 using $t = 1$ and 4 trials. We calculate VI with respect to level 2 of the hierarchy in panel (a) and level 3 in panel (b). Both models exactly identify both levels of the hierarchy at $t = 4$. The APM perfectly identifies both levels at $t = 1$ which is slightly better on average than the RBCM, and it has a more stable solution for level 3. Each point is an average over 100 solutions.

The APM proves to be very robust to noise in the system. We use the VI information metric V (see Appendix B) to compare solved partitions with the constructed networks.

A. Three-level hierarchy

We identify two levels of a constructed hierarchy [22, 35] using both the APM and RBCM models. The three-level hierarchy is depicted in Fig. 2, and the results are given in Fig. 3. The system has $N = 256$ nodes divided into $q_3 = 16$ communities at level 3 with sizes as noted in

Fig. 2. Edges in each community are randomly assigned with a probability of $\overline{p}_3 = 0.90$. These communities are grouped as shown into $q_2 = 5$ communities that define level 2 of the hierarchy. The average internal density of level 2 communities is $\overline{p}_2 = 0.47$ which are defined by randomly connecting nodes in the respective sub-groups of level 3 at an intercommunity edge density of $p_2 = 0.3$. Level 1 is the completely merged system which is defined by randomly connecting nodes in sub-groups of level 2 at an intercommunity edge density of $p_1 = 0.1$.

We apply the algorithm of Sec. III to both models and solve a large range of model weights γ or γ_{RB}, respectively, in order to illustrate the differences in the two models. In Fig. 3, we plot VI V as a function γ or γ_{RB} [36] using $t = 1$ and 4 trials. VI is calculated between the respective solutions and the level 2 or 3 partitions of the hierarchy. These data are then plotted in panels (a) and (b), respectively. Both models exactly identify both levels of the hierarchy at $t = 4$. The APM is slightly better in accurately identifying them with $t = 1$, and it has a more "stable" solution in panel (b).

B. Noise test

The concept of "noise" in community detection corresponds to "extra" edges that connect a node to communities other than its best assignment(s). In general, we cannot initially distinguish between edges contributing to noise and those constituting edges within communities of the best partition(s). Community detection methods experience the effects of noise in at least two distinct ways: (1) The edges due to noise act to obscure the best partition(s) in an algorithm by creating "confusion" for early community assignments (a dynamical effect). (2) The extra edges influence the quantitative evaluation of the best community assignments (a "metric" effect). In some models, this second effect can negatively impact the contribution of edges that comprise the best communities.

1. *Benchmark*

We test the accuracy of APM and RBCM models with high levels of noise in a series of strongly defined systems with "realistic" distributions of community sizes. Specifically, we define a set of communities with a power-law distribution of community sizes specified by an exponent β, minimum size $n_{\min}$, and maximum size $n_{\max}$. We add random edges to the system, largely defining the intercommunity noise, based on a power-law distribution of node degrees given by an exponent α, average power-law degree $\langle k\rangle_\alpha$ (or minimum degree $k_{\min}$), and maximum degree $k_{\max}$ [37]. This initial framework is similar to a benchmark by Lancichinetti *et al.* [27, 38]. We then connect internal community edges at a high density p_{in}.

The strongly defined communities provide unambiguous partitions where the large level of noise does not significantly alter the optimal solutions (see Sec. V B 4). This density-based definition of community structure is consistent with concepts proposed for community identification by Palla *et al.* [4]. We solve for the systems using the algorithm in Sec. III for both models with $t = 1$ and 4 trials and using SA for the RBCM with $t = 4$.

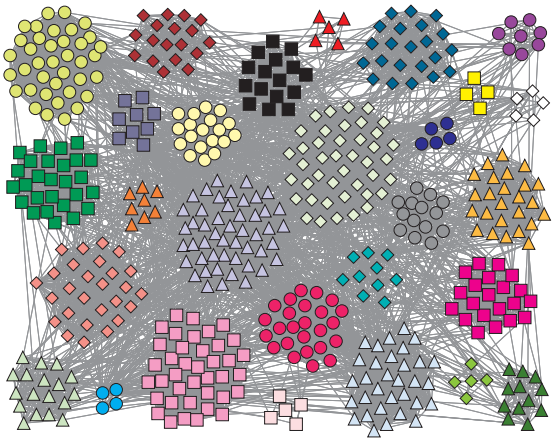

FIG. 4: (Color online) A sample graph with $N = 512$ nodes for the noise test in Sec. V B. In this sample, the node degrees are initially defined in a power-law distribution with an average $\langle k\rangle_\alpha = 5.4$, maximum $k_{\max} = 100$, and exponent $\alpha = -2$. Communities have a power-law distribution of sizes with a minimum $n_{\min} = 4$, maximum $n_{\max} = 50$, and exponent $\beta = -1$. These communities are then strongly defined by connecting all internal community edges ($p_{in} = 1$).

2. *Accuracy results*

Figures 4 and 5 show a sample system and the first test results, respectively. We use two system sizes of $N_{\rm a,b} = 512$ and $N_{\rm c,d} = 4096$ nodes, respectively. The initial power-law degree distribution uses $\alpha = -2$; and the maximum degree constraints are $k^{\max}_{\rm a,b} = 100$ and $k^{\max}_{\rm c,d} = 1000$, respectively. We increment the average power-law degree $\langle k\rangle_\alpha$ to vary the system noise (the average external degree $k_{out} \simeq \langle k\rangle_\alpha$ for large systems). Community sizes range from $n_{\min} = 4$ to $n_{\max} = 50$ nodes and are distributed according to $\beta_{\rm a,c} = -1$ or $\beta_{\rm b,d} = -2$, respectively. The internal community edges are maximally connected at a density of $p_{in} = 1$.

We plot VI V versus k_{out} for both Potts models where VI is calculated between the solved partition and the generated graph ($V^{\max}_{\rm a,b} = 9$ and $V^{\max}_{\rm c,d} = 12$). For the APM, we use $\gamma = 1$ for every solution, and we allow zero energy moves after the system reaches an initially converged state. For the RBCM, we subjectively select the *best* solution corresponding to the *highest accuracy* γ_{RB} independently determined for each k_{out} given the *known* answer (see Appendix C). We further solve the system via SA at this best value of γ_{RB} for comparison. We average over 100 graphs for each point in panels (a)

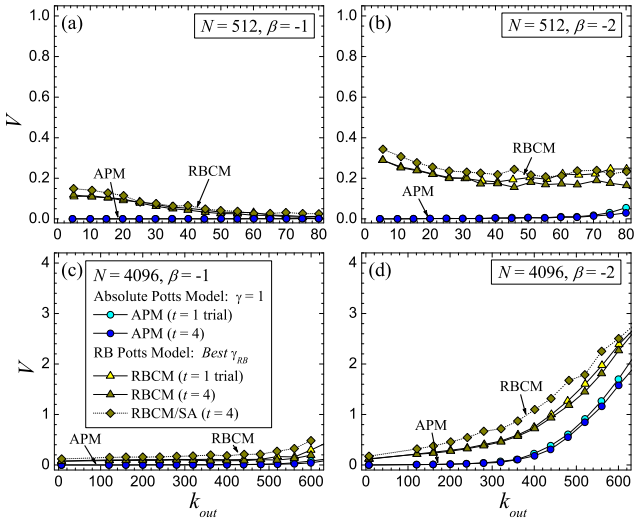

FIG. 5: (Color online) Plot of VI V between solved and known test systems in Sec. V B as a function of the average external node degree k_{out}. The system sizes are $N_{\rm a,b} = 512$ in panels (a) and (b) (with a sample system depicted in Fig. 4) and $N_{\rm c,d} = 4096$ nodes in panels (c) and (d). The graphs are solved with the APM and the RBCM using the algorithm in Sec. III with $t = 1$ and 4 trials. We use $\gamma = 1$ for the APM on all solutions, and we subjectively select the *best* γ_{RB} for the RBCM independently for each k_{out} (see Appendix C). For comparison, we also solve the system at this *best* γ_{RB} using SA. System noise is randomly assigned in an approximate power-law degree distribution [37] with an exponent $\alpha = -2$, an average degree $\langle k\rangle_\alpha$, and maximum degrees of $k^{\rm max}_{\rm a,b} = 100$ or $k^{\rm max}_{\rm c,d} = 1000$, respectively. Constructed communities are randomly assigned in a power-law size distribution [38] specified by an exponent $\beta_{\rm a,c} = -1$ or $\beta_{\rm b,d} = -2$, minimum size $n_{\rm min} = 4$, and maximum size $n_{\rm max} = 50$. Communities are then maximally connected with $p_{in} = 1$. Even with $t = 1$, the APM is almost perfectly accurate for most tested parameters in this problem. See Appendix E regarding the accuracy transitions in panel (d). Data are averaged over 100 graphs in panels (a) and (b) and 25 graphs in panels (c) and (d).

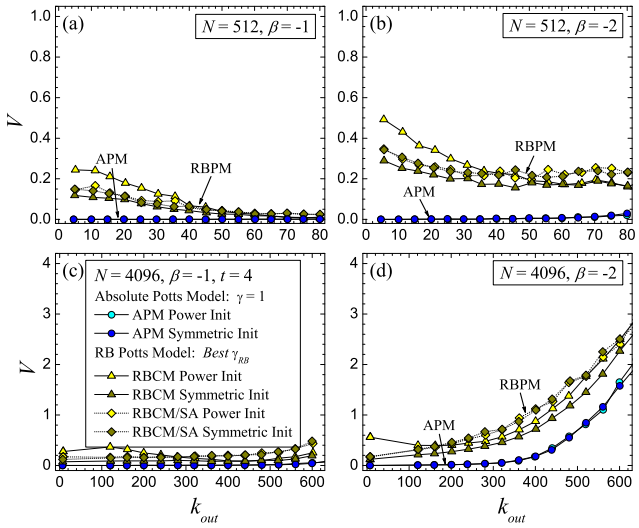

FIG. 6: (Color online) Plot of VI V between solved and known test systems in Sec. V B vs the average external node degree k_{out}. In panels (a) through (d), system sizes are $N_{\rm a,b} = 512$ and $N_{\rm c,d} = 4096$ nodes, respectively. The graphs are solved with the APM and the RBCM using the algorithm in Sec. III with $t = 4$ trials. The constructed configurations are identical to those used in Fig. 5. In this plot, we test two different initial states for the solutions: a symmetric initial state and a random power-law distribution (see text). We use $\gamma = 1$ for the APM on all solutions, and we subjectively choose the *best* γ_{RB} for the RBCM for each k_{out} (see Appendix C). For comparison, the results for SA with $t = 4$ are also depicted and are solved using this best value of γ_{RB}. The APM and SA with the RBCM show no difference in accuracy between the symmetric and random initial states. A symmetric state appears to be the favored starting configuration for the RBCM when using a greedy algorithm in this benchmark. In fact, this symmetric initial state allows the RBCM to slightly *outperform* SA in accuracy (see text). We average over 100 graphs in panels (a) and (b) and 25 graphs in panels (c) and (d).

and (b) and 25 graphs in panels (c) and (d).

In panels (a) and (c), the advantage in accuracy for the APM is modest. The accuracy of the RBCM increases in panels (a) and (b) at higher levels of noise due in part to the fact that the degree distribution is becoming more uniform as we increase $\langle k\rangle_\alpha$ but keep $k_{\rm max}$ constant. While the RBCM performs excellently in many cases, the APM outperforms it to varying degrees for most tested parameters and levels of noise. Moreover, the APM is often able to almost perfectly solve the system.

The rapid increases in VI for both models in Fig. 5(d) are due to transition effects described in Appendix E. We subjectively select the best γ_{RB} in this paper, but note also that our algorithm can slightly outperform SA in accuracy in many cases *for either Potts model* (see Sec. V B 3 and Appendix D). See Sec. VI D regarding how the high levels of noise in this test actually mitigate the effect of the resolution limit for the RBCM.

3. *Dependence on initial condition and SA accuracy*

A community detection algorithm should ideally be robust with respect to the initial state that is used to solve the system. We show that APM displays this feature, and we contrast the result with the RBCM when using the greedy algorithm in Sec. III. We further elaborate on the accuracy of SA compared to this greedy algorithm.

In Fig. 6, we plot VI V vs k_{out} where increasing k_{out} corresponds to higher levels of system noise. We measure V between the solved and constructed systems where the defined systems are identical to the previous subsection. We test both models beginning from two different initial states: a symmetric initial state of one node per cluster with $q_0 = N$ and a random power-law configuration with $q_0 \simeq q$ which is different than the defined answer. For simplicity, this random initial state uses the same distribution parameters ($\beta_{\rm a,c} = -1$ or $\beta_{\rm b,d} = -2$, $n_{\rm min} = 4$, and $n_{\rm max} = 50$) that are used to generate the answers.

The best solutions for the APM are robust to the initial state of the system in this benchmark, including during the major accuracy transition in panel (d), despite using a greedy algorithm. The symmetric initial state performs very well for the RBCM and is the favored starting configuration compared to the random power-law state. The situation is reversed for the RBCM on the benchmark in Fig. 1 in Sec. IV where the symmetric initialization performs worse than a random initial state with $q_0 = 4$ communities (with or without constraining q in the dynamics [34]) although that benchmark has an unrealistically symmetric community structure. While optimizing the RBCM often provides excellent partitions, this difference in accuracy between initial states indicates that it is more easily trapped in unfavorable regions of the energy landscape than the APM when using a greedy algorithm.

As expected, SA with the RBCM shows no difference in accuracy for either initial state, but the *greedy* algorithm *outperforms* SA in terms of accuracy when using a symmetric initial state (see also Appendix D). This reduced accuracy for SA compared to a greedy algorithm is not isolated to this benchmark. Lancichinetti and Fortunato [27] compared the accuracy of several algorithms using their benchmark [38]. One result in [27] showed that a similar greedy algorithm optimizing modularity [equivalent to $\gamma_{RB} = 1$ in Eq. (7)] by Blondel *et al.* [6] also outperformed SA in accuracy on that benchmark. See also Good *et al.* [39] regarding difficulties associated with modularity optimization in practical problems.

4. Noise tolerance discussion

Even at low levels of noise, these benchmark graphs exceed the proposed definition of so-called "weak" communities [40], but the communities are not ill-defined from an intuitive standpoint within the tested range of noise. In panel (d) for example, at $k_{out} \simeq 370$ the average number of edges connecting a given node to another community is $\ell \simeq 1$ because the k_{out} edges are randomly spread over $(q_{\rm b} - 1) \simeq 370$ communities. This value is small compared to the average internal degree $k_{in} \simeq 10$ *and* the average number of *missing* links with an external community $(\langle m \rangle - \ell) \simeq 10$ where $\langle m \rangle$ is the average community size. Thus, the communities remain well-defined particularly given their high edge density $p_{in} = 1$.

The two Potts models respond to the noise in the system in distinctly different ways in terms of how the community measure is calculated. Noise complicates community assignment decisions for the RBCM because the configuration null model [the second term in Eq. (7)] incorporates the contribution of *all* edges, including noise, for every node assignment evaluation even after a reliable solution "kernel" is located during early stages of the solution dynamics (the metric effect of noise).

The APM evaluates all edges for community assignment decisions through relative energy calculations, but Figs. 5 and 6 demonstrate that the best solution is often completely unaffected by the system noise if the algorithm can navigate sufficiently close to the solution. Once an initial solution kernel evolves during the early stages of the algorithm dynamics, confusion caused by random system noise is often easily mitigated by the missing edge energy penalty [the second term in Eq. (2)]. That is, the metric effect of noise on the APM is very favorable so that the main challenge caused by noise in the network is often due to the dynamical effect of noise (early incorrect assignments) that affects both models.

We could further improve the accuracy of the APM using a more robust, but much slower, algorithm such as SA. Nevertheless, this benchmark illustrates that the energy landscape of the APM is more easily navigated, particularly for $\gamma = 1$, than the RBCM. The energy landscape of the APM is more difficult to navigate for $\gamma \ll 1$ as compared to $\gamma \geq 1$ or when communities are not as strongly defined as they are in this test, but the model maintains an exceptional accuracy [22, 27].

VI. RESOLUTION LIMIT

The quantitative approaches of modularity [11] and the RBPM [13, 15] in Sec. II B were implemented by incorporating global graph parameters into the models. Both models marked important progress in the field of community detection, but Refs. [18] and [19] noted an unintended consequence of using global community measures — an imposed resolution limit. The resolution limit restricts the solutions of affected models so that they cannot correctly resolve all communities of a system in certain non-pathological cases. For modularity and the RBCM, the number of communities in a system tends toward $\sqrt{L}$ [18] and $\sqrt{\gamma_{RB} L}$ [19], respectively. The models have difficulty properly resolving small communities in large systems and may incorrectly divide large communities.

We first discuss local versus global measures, and we then illustrate the resolution limit for the RB Potts models, including modularity as a special case. We also show that the APM is free of resolution-limit effects because it is a local measure of community structure.

A. Local vs global measures

Including global dependences in quantitative community detection models was apparently rooted in the need to objectively determine the community structure of arbitrary graphs. The assumption is that the global properties of the graph should imply its local community structure. Global dependences make the partition solution objective since it allows the same quantitative model to automatically rescale to any graph, but they also became the central element that caused a resolution limit.

One suggested solution [18, 19] to the resolution limit is to define a *local* measure of community structure. That

is, community evaluations are made based only on local features of the graph in the neighborhood of the involved nodes and communities. Some approaches that provide local community detection methods include clique percolation [4, 41], analyzing random walks [42, 43], "label propagation" [29] (and one variant in [44]), and local variants of modularity [5, 45]. See Sec. VI E 4 and Appendix A regarding the RBER model. The APM is also a strongly local measure of community structure.

Local models possess beneficial properties for solving some networks such as: large networks that are "defined" as the network is explored (e.g., the World Wide Web), incompletely known networks (e.g., social interactions), coarse partitioning and refinement algorithms, and dynamic networks. Communities sufficiently isolated from graph changes do not have to be repetitively updated as the network is modified. There exists work for modularity [46] that preserves the measure as the system is scaled, but a local model does not require this extra buttressing. Further, several of the most accurate community detection methods [27, 43] are based on essentially local methods or models of community detection [22, 42, 43].

However, using a local measure of community structure returns to the subjectivity problem. How does the model *objectively* determine the community structure of arbitrary graphs? Stated specifically for the APM, how does one choose the "correct" resolution(s) [*i.e.*, value(s) of γ in Eq. (2)] that will best solve the system? Several answers to this problem are as follows although the concepts are not restricted to local models.

One approach is to define a community independent of the graph being solved. For example, we might seek to identify all communities of "close friends" in a social network regardless of the size of the graph. For the APM, Eq. (3) relates the model weight γ to the minimum community edge density for all communities in a partition.

Some other methods, which are beyond the scope of this paper, define an algorithm or measure that can determine which resolutions (see Sec. II A 2) are the best partitions for the network. Arenas *et al.* [47] varied a weight parameter with modularity and tracked stable partitions. Kumpula and co-workers [48, 49] as well as Fenn *et al.* [50] also explored stability approaches for the RBCM. Our multiresolution method [22] utilized information comparisons among independent solutions to quantitatively evaluate the best resolutions. Zhang *et al.* [51] used a topological weighting strategy. Cheng and Shen [43] used the stability of random walker diffusion dynamics to identify the most relevant resolutions.

B. Circle of cliques

Fortunato and Barthélemy [18] and Kumpula *et al.* [19] identified a resolution limit in the respective models in part by considering the unweighted system shown in Fig. 7, a set of q cliques (maximally connected communities) connected in a circle by single edges. In Fig. 7(a),

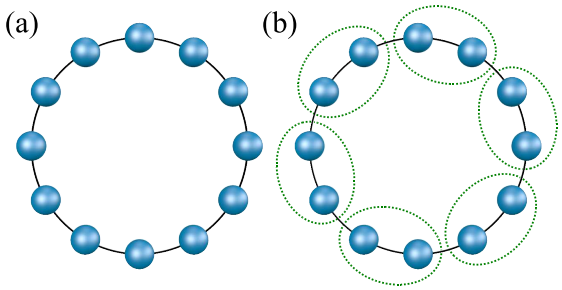

FIG. 7: (Color online) Each graph is a circle of cliques with N total nodes and L total edges. (a) A set of q cliques with m nodes each are connected in a circle by q links. (b) r consecutive cliques are each grouped together. Intuitively, one would expect that any measure should resist merging these communities on any system scale (e.g., N, L, or q) if $m \geq 3$.

each clique is a separate community. The total number of links is L and the number of nodes in the system is N. The total number of links between the cliques is q. The number of nodes in each clique m can be varied independently of q. From Eq. (2), the APM energy is

$$E_{\rm a} = -\frac{1}{2} q m (m-1). \tag{8}$$

This energy $E_{\rm a}$ has *no finite extremum* with respect to *any* global parameters of the graph. The analogue to Eq. (8) for modularity and the RBCM is where the resolution limit was demonstrated. That is, those models have minima, $q^*_{mod} = \sqrt{L}$ and $q^*_{RB} = \sqrt{\gamma_{RB} L}$, respectively, in the expected number of communities. Neither of these values correspond to the intuitive partition (q clique communities) for all system sizes.

Figure 7(b) depicts sets of r cliques grouped together. The specific conditions, based on γ_{RB}, for r neighboring cliques to merge are given by the following relations. The RBCM of Eq. (7), using the configuration null model, includes modularity as a special case when $\gamma_{RB} = 1$. Two neighboring cliques ($r = 2$) [19] will merge if

$$\gamma_{RB} < \frac{q}{m(m-1)+2}. \tag{9}$$

The dependence on the number of cliques q is a problem since this condition for γ_{RB} can always be satisfied if q is large enough (see also Sec. VI D). For example, if $m = 3$ and $\gamma_{RB} = 1$, the cliques merge if $q > 8$.

When using the RBER model with an Erdős-Rényi null model in Eq. (6), neighboring cliques merge if

$$\gamma_{RB} < \frac{q - 1/m}{m(m-1)+2}. \tag{10}$$

We can always choose q large enough to induce a merger of neighboring cliques for any γ_{RB} (see also Appendix A). These results generalize so that a resolution limit can be found to apply for an arbitrary choice of null model [19] when using the RBPM of Eq. (5). The APM energy $E_{\rm b}$

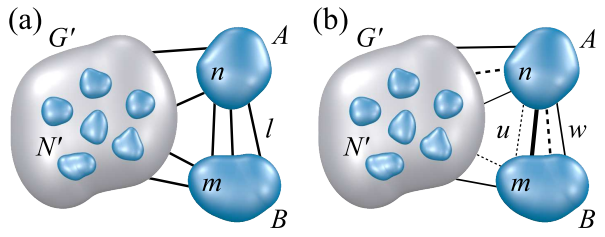

FIG. 8: (Color online) A large graph G has N nodes and three sub-divisions depicted, one potentially large sub-graph G' with N' nodes and two distinct communities A and B with n and m nodes, respectively. In panel (a), G is unweighted, and communities A and B are joined by l edges. In panel (b), G is a weighted graph. For visualization purposes, solid lines depict weighted edges, dashed lines depict weighted missing links, and the link thickness depicts a relative link weight. Communities A and B are joined by weighted edges with a summed weight of w. Weighted *missing* links have a summed weight of u. All other graph features are left unspecified but are consistent with the community designations.

of the configuration in Fig. 7(b) with r merged cliques is

$$E_{\rm b} = -\frac{\gamma+1}{2}qm(m-1) \times \left[1-\frac{\gamma}{\gamma+1}\frac{rm-1}{m-1}+\frac{2(r-1)}{rm(m-1)}\right]. \quad (11)$$

We compare the energies in Eqs. (8) and (11) and find that $r=2$ cliques merge if

$$\gamma < \frac{1}{m^2-1}. \quad (12)$$

This merge condition depends *only* on the *local* variable m. Solving the system with $\gamma = 1$ will ensure that neighboring cliques will not merge on *any* global scale (e.g., N, L, or q) of the system. Since γ adjusts the weight applied to missing links, we can *force* a merger of neighboring cliques if we reduce γ to a sufficiently low value. At $m=3$ for example, we can force a merger if $\gamma < 1/8$.

C. Heterogeneous communities

Resolution-limit effects can be exacerbated when communities of substantially different sizes are present. Danon *et al.* addressed improvements for modularity to better resolve heterogeneous structures [1] with Newman's algorithm in [32]. The APM deals with heterogeneous communities naturally.

Figure 8(a) depicts a large graph G with three divisions. Communities A and B have n and m nodes, respectively, and are connected by l edges. Sub-graph G' has N' nodes with an unspecified structure. For the RBPM of Eq. (5), using a generic null model, the number of edges l that causes communities A and B to merge is of the order [19]

$$l \gtrsim \frac{\gamma_{RB} nm}{N}. \quad (13)$$

The RBER model yields a merge condition of

$$l > \frac{2L}{N(N-1)}\gamma_{RB} nm. \quad (14)$$

In Eqs. (13) and (14), even for $l=1$ the merge conditions can be readily satisfied in large graphs for any reasonable value of γ_{RB} due to the dependences on global graph parameters L or N (see also Sec. VI D and Appendix A).

Our APM model merges communities A and B if

$$l > \frac{\gamma}{\gamma+1}nm. \quad (15)$$

The merge condition is based only on γ and the local community sizes n and m. For $\gamma = 1$, even small communities merge with large ones only if there are many interconnections. The dependence on γ is consistent with the purpose of its introduction in Eq. (2) — to allow the model to vary the system resolution.

D. Mitigated resolution limit

We return to the unweighted system of q cliques in Fig. 7 and Sec. VI B to show that certain conditions will mitigate resolution-limit effects. By design, this circle of cliques was constructed to have an unambiguous intuitive answer. Communities are not so clearly defined in practice, so we convert the cliques to communities with ℓ_{in} edges each, not necessarily maximally connected. We also increase the number of intercommunity edges so that each community has an average of ℓ_{out} edges connected to s other communities ($qs\ell_{out}/2$ total external edges). The original condition for the RBCM for neighboring cliques to merge [with $l_{in} = m(m-1)/2$, $l_{out}=1$, and $s=2$] is given by Eq. (9). The new merge condition is

$$\gamma_{RB} < \frac{q\ell_{out}}{(2\ell_{in}+s\ell_{out})}. \quad (16)$$

High levels of noise [$s \simeq O(q)$ and $\ell_{out} \gtrsim O(1)$] tend to *reduce* the effect of the resolution limit because the ratio is asymptotic to $\gamma_{RB} \simeq 1$.

For the benchmark in Sec. V B, Eq. (16) explains how the RBCM can perform very well, despite a resolution limit, even when a large number of communities q are present (we also subjectively evaluate many values of γ_{RB}). On the other hand, more weakly defined communities [$\ell_{in} < m(m-1)/2$] tend to increase resolution-limit effects, but system noise can substantially and positively influence the effects of the resolution limit.

E. Locality of weighted Potts models

When considering weighted graphs, the introduction of (additional) global dependences should be a "warning flag" because global dependences are the source of the resolution limit. We show that the APM is remains a local model for weighted and directed graphs.

1. Absolute Potts model

We generalize results from Sec. VI C for the APM with an emphasis on weighted graphs including those with weighted missing links. Missing link weights correspond to levels of adversarial relations between nodes. "Neutral" relations use a weight $b_{ij} = 1$ since a weight of 0 is an inconsistent community detection model in general. The following result also applies to directed graphs. Represented as a sum over communities, Eq. (2) becomes

$$\mathcal{H}_s(\{\sigma\}) = \sum_s \big(- w_s + \gamma\, u_s \big) \tag{17}$$

where w_s and u_s are the energy sums of *connected* and *missing* edges of community s, respectively. For reference in Sec. VI E 3, the unweighted version of Eq. (17) is

$$\mathcal{H}_s(\{\sigma\}) = \sum_s \big[- (\gamma + 1)\, l_s + \gamma l_s^{\max} \big] \tag{18}$$

where l_s is the number of edges and $l_s^{\max}$ is the maximum number of possible edges in community s.

In Fig. 8(b), we use Eq. (17) to calculate the condition for two arbitrary communities A and B to merge in a general graph. A and B are connected by edges with a total weight of w and a total missing link weight of u. The merge condition is almost trivially given by

$$w > \gamma u. \tag{19}$$

Note that this merge condition is based only on γ and the connected or missing edges *between* A and B. The APM remains a local measure for general graphs in the strongest sense because node assignments are *independent of the internal structure* of the communities (see also Sec. VI E 4).

2. Weighted configuration RB Potts model

A weighted generalization of the RBCM model is

$$\mathcal{H}^w_{CM}(\{\sigma\}) = \sum_s \left(-w_s + \gamma_{RB} \frac{W_s^2}{4W} \right) \tag{20}$$

where we express it as a sum over all communities. W is the total weight of all edges in the system and W_s is the total weight of *all* edges in community s (including edges connected to *other* communities). As with the unweighted variant, this weighted model is necessarily already a global measure due to W in the sum over W_s^2.

3. Weighted Erdős-Rényi RB Potts model

The weighted generalization of the RBER model of Eq. (6) increases the global dependence of the model as it is proposed in [13]. We rewrite the *unweighted* RBER model as a sum over communities

$$\mathcal{H}_{ER}(\{\sigma\}) = \sum_s (-l_s + \gamma_{RB} p\, l_s^{\max}). \tag{21}$$

Equations (18) and (21) show that in the special but important case of *unweighted* graphs, the APM and RBER models are coincidentally equivalent if we *rescale* the null model weight by $\gamma_{ER} \equiv \gamma_{RB} p$ to explicitly remove the global density dependence (see Appendix A).

We write a conceptual generalization of Eq. (21) for weighted graphs which we use again in Sec. VI E 4,

$$\mathcal{H}^w_{ER}(\{\sigma\}) = \sum_s (-w_s + \gamma_{RB} p\, w_s^{\max}). \tag{22}$$

Analogous to $l_s^{\max}$, $w_s^{\max}$ is the "maximum weight sum" of community s which must be defined. RB used one natural definition of (i) $w_s^{\max} \equiv \overline{W} l_s^{\max}$ to obtain [13]

$$\mathcal{H}^w_{ER}(\{\sigma\}) = \sum_s \big(-w_s + \gamma_{RB} p \overline{W}\, l_s^{\max} \big) \tag{23}$$

where $\overline{W}$ is the average weight over *all* edges.

In Fig. 8(b), an arbitrary graph G has three parts: two communities A and B, and an arbitrary sub-graph G'. Communities A and B are connected by a summed edge weight w. We ignore the missing link weight sum $(u = 0)$ since the model does not account for them. Using Eq. (23), the condition for the communities to merge is

$$w > \gamma_{RB} p \overline{W} nm. \tag{24}$$

The dependence on $\overline{W}$ allows arbitrary changes to independent parts of a graph to unintuitively affect each other. For example, if we alter the edge weights in subgraph G', we change the average edge weight $\overline{W}$. As a result, we indirectly change the condition for communities A and B to merge even though there are no local changes that affect A, B, or the links between them. This type of indirect effect caused by global parameters of the graph is at the heart of the resolution limit.

4. Local "Erdős-Rényi" Potts model and "weak" locality

One can modify the weighted RBER model of Eq. (22) to create a local variant. We briefly introduce our own variant since the comparison illustrates how "strongly" the APM defines a local measure of community structure.

Another natural interpretation of $w_s^{\max}$ in Eq. (22) is (ii) $w_s^{\max} \equiv \overline{w}_s\, l_s^{\max}$ where $\overline{w}_s$ is the average edge weight in the *local* community s. We also define $\gamma_{ER} \equiv \gamma_{RB} p$ to explicitly remove any dependence on the global density of the system. (This was the initial form of the RBER model [13]. See also Appendix A.) In removing the density dependence p, the model is technically no longer an

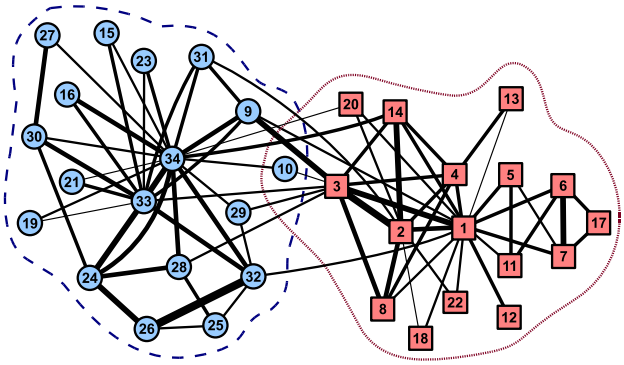

FIG. 9: (Color online) Graph depicts the Zachary karate club [52] solved with the APM using weighted edges (relative line thickness), $\gamma = 1$, and $q = 2$ communities by constraint. All nodes except 10 are correctly assigned. By our analysis, this node appears frequently in both groups (see text).

"Erdős-Rényi" Potts model; however, in this interpretation, Eq. (22) simplifies to

$$\mathcal{H}_{ER}^{local}(\{\sigma\}) = \sum_s \overline{w}_s \left(-l_s + \gamma_{ER}\, l_s^{\max}\right). \qquad (25)$$

This variant uses almost the same energy sum as the *unweighted* RBER model in Eq. (6) except that total energy of community s is weighted by $\overline{w}_s$. Equation (25) is a *local* model in the sense that only parameters in the "neighborhood" of the local communities contribute to the energy, but it is a local model in a "weaker" sense than the APM because node assignments depend on the *internal structure* (edge weights in this case) of the communities.

One can devise applications for such weakly local quality functions when influences within a graph need to be abstracted for efficiency or due to limited knowledge of the full details of the network (e.g., social networks with influential personalities). However, despite being a local model, using these indirect dependences for community assignments (without associated edges between nodes) should elicit some caution because similar indirect effects on a global level are the source of the resolution limit for modularity and the RBPM.

VII. EXAMPLES

A common test is the Zachary karate club [52]. It provides a small and real example of a social division that occurred while the group was under study. The graph consists of 34 people with 78 recognized relationships between them that are weighted according to the strength of the friendships (depicted by the relative line thickness). We use the *weighted* relations in Eq. (2) with $b_{ij} = 1$ and divide the graph into two parts by constraint as shown in Fig. 9. Our algorithm correctly identifies the communities except for node 10 which appears frequently in *both* groups because there is no energy difference between the two assignments at $\gamma = 1$. (This is a rudimentary identification of an overlapping node using a method such as in [13].) In the actual division, node 10 associated with the group depicted by circles. A more complete multiresolution analysis in [22] would correctly place node 10.

We also construct a very large system similar to those defined in Sec. V B. The system has 40×10^6 nodes and $L = 1\,157\,634\,899$ edges assigned in a power-law distribution of node degrees with an exponent $\alpha = -2$. We specify the minimum and maximum degrees as $k_{\min} = 20$ and $k_{\max} = 500$, respectively. The system is randomly partitioned into $q = 2\,443\,782$ communities in a power-law distribution of community sizes with an exponent $\beta = -1$ with sizes ranging from $n_{\min} = 10$ to $n_{\max} = 25$ nodes. The average internal community density is set to $p_{in} = 0.95$. The average graph density is $p = 1.45 \times 10^{-6}$.

We solve the system with $\gamma = 1/2$ in Eq. (2) with the algorithm in Sec. III using $t = 1$ trial. (Random community edge assignments allow some nodes to be weakly connected to their intended communities. Using $\gamma = 1/2$ ensures that all but the extreme outliers are properly assigned.) The system was solved very accurately with $V = 1.17 \times 10^{-7}$ in 3.9 hours on a single processor [31].

VIII. CONCLUSION

We present an exceptionally accurate and local spin-glass-type Potts model for community detection: (1) our approach employs an absolute energy evaluation as opposed to a null model comparison. (2) Its accuracy, even when using a greedy algorithm, is among the best of currently available algorithms. (3) The model is robust to noise in the system. (4) It is a local measure in the strongest sense of the term for unweighted, weighted (including weighted "adversarial" relationships), and directed graphs. As such, it corrects a resolution-limit problem that affects other popular measures [11, 15, 18, 19]. (5) Heterogeneous community sizes are naturally resolved. (6) The computational demand often scales as $O(tL^{1.3})$ where t is the number of optimization trials [generally $O(10)$ or less] and L is the number of edges in the system. We have been able to accurately solve synthetic systems as large as 40×10^6 nodes and over 10^9 edges [31]. In Ref. [22], we illustrated in detail how this core community detection method may be extended to systematically, accurately, and rapidly identify general multiresolution structures.

ACKNOWLEDGMENTS

We thank A. Lancichinetti and S. Fortunato for providing the SA code and UCINet for network data made available on their website. This work was supported by the LDRD DR on the physics of algorithms at LANL.

Appendix A: RESOLUTION LIMIT AND THE ERDŐS-RÉNYI POTTS MODEL

For unweighted graphs, the RBER model of Eq. (21), based on the Erdős-Rényi null model, is not inherently a global measure of community structure as is the RBCM, based on the configuration null model. The original model [13] was defined without the density p where $\gamma_{ER} \equiv \gamma_{RB} p$. The *ad hoc* inclusion of the density carried an implicit assumption that γ_{RB} is constrained to some range, perhaps by $\gamma_{RB} \simeq O(1)$ (otherwise, introducing a second constant is not meaningful). It then became a Potts model based on an Erdős-Rényi null model.

The justification for including the graph density in the model was initially based on heuristic arguments about density inequalities that bounded the behavior of γ_{ER}. Data were also presented using a common, but very small, benchmark (discussed in Sec. IV) that supported the approximation of $\gamma_{ER} \propto p$. However, the approximation is not generally applicable. For example, between the systems in Secs. IV, V, and VII, we would need to vary γ_{RB} by at least 3 orders of magnitude (and arbitrarily larger if we increase the system size in Sec. V B) if we wish to consistently identify the most accurate solution for each system. If we remove the constant (but graph dependent) density p, we trivially remove from Eq. (21) any dependence on global graph parameters.

This change is more than a pedantic exercise. Connecting the RBER model to the system density allowed it to automatically scale to solve arbitrary graphs in a semiobjective manner (see Sec. VI A), but it also appeared to impose a resolution limit [19]. Trivially removing the global dependence on p effectively "eliminates" the resolution limit for the model if one reinterprets the meaning of the original model weight γ_{ER}. With this change, we assert that there is *no genuine resolution limit* in the *unweighted* RBER model as it was originally presented in [13] without the density dependence p.

The second term in Eq. (21) indicates that γ_{ER} specifies the fraction of $l_s^{\max}$ that each community must have before it has an energy less than zero. Thus, we reinterpret γ_{ER} as the minimum edge density of each community in a solved partition (or the maximum external edge density [13]), but this minimum density is enforced through only *local* constraints. The cost for this freedom is that we must *choose* the "correct" weights γ_{ER} for each graph, but the best choices are not arbitrary.

After removing p, we re-analyze the resolution-limit results obtained for the RBER model in Secs. VI B and VI C. Using Fig. 8(a), the original condition for two arbitrary unweighted communities A and B to merge is given by Eq. (14). Without p, the new merge condition is

$$l > \gamma_{ER} nm \tag{A1}$$

which is based only on *local* variables of communities A and B and the independently set γ_{ER}.

For the circle of cliques depicted in Fig. 7, the original merge condition is given by Eq. (10). The new condition for two neighboring cliques to merge is

$$\gamma_{ER} < \frac{1}{m^2}. \tag{A2}$$

Using the reinterpretation of γ_{ER}, the value of $\gamma_{ER} = 1/2$ demands at least a 50% edge density for each community to be valid. At $m = 3$, Eq. (A2) demands $\gamma_{ER} < 1/9$ for a merger to occur. Therefore, at $\gamma_{ER} = 1/2$ the model will not experience a resolution limit effect for *any* global scale of N, L, or q for cliques of size $m \geq 3$.

After removing the density and reinterpreting γ_{ER}, the model is not genuinely subject to a resolution limit because the constraints that define the community structure are enforced *locally*. We can then apply concepts mentioned in Sec. VI A to solve graphs with a local community measure. One caveat is that the locality of the RBER model does not extend as naturally to weighted systems (see Sec. VI E).

Appendix B: VARIATION OF INFORMATION

We use the variation of information metric [53] for our accuracy tests in Sec. V. The probability that a randomly selected node from partition A will be a member of community k is $P(k) = n_k/N$, where n_k is the number of nodes in community k and N is the total number of nodes in the system. The Shannon entropy is

$$H(A) = -\sum_{k=1}^{q_A} \frac{n_k}{N} \log \frac{n_k}{N} \tag{B1}$$

where q_A is the number of communities in partition A.

The mutual information $I(A, B)$ evaluates the level of interdependence in two sets of data. We define a "confusion matrix" for partitions A and B by identifying how many nodes n_{ij} of community i of partition A are in community j of partition B. The mutual information is

$$I(A, B) = \sum_{i=1}^{q_A} \sum_{j=1}^{q_B} \frac{n_{ij}}{N} \log \left(\frac{n_{ij} N}{n_i n_j} \right) \tag{B2}$$

where n_i is the number of nodes in community i of partition A and n_j is the number of nodes in community j of partition B. The variation of information $V(A, B)$ is

$$V(A, B) = H(A) + H(B) - 2I(A, B), \tag{B3}$$

which measures the "distance" in information between the two partitions A and B. The range of values for VI is $0 \leq V(A, B) \leq \log N$. We use base 2 logarithms.

Appendix C: EXAMPLE NOISE TEST SOLUTION WITH THE RBCM

In Sec. V B, we add noise to a strongly defined system to test the accuracy of the RBCM of Eq. (7) compared

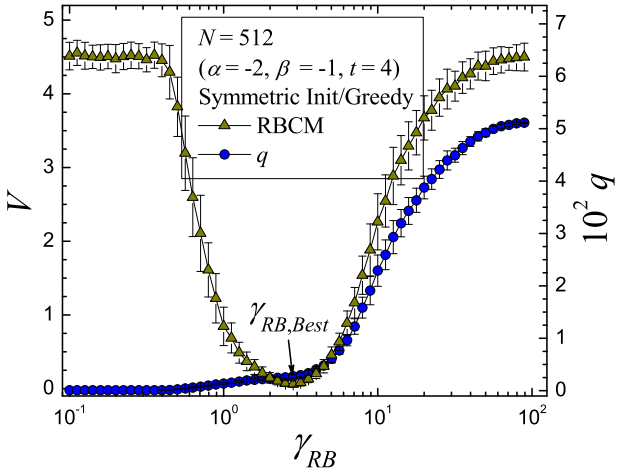

FIG. 10: (Color online) Plot of VI V vs γ_{RB} for the RBCM. In Sec. V B, we generate a set of strongly defined communities with varying levels of intercommunity noise k_{out}. Using the greedy algorithm in Sec. III, we compare the accuracy of solutions found with the RBCM to our APM. Since the models operate differently, we compare the results in Figs. 5 and 6 for our APM using $\gamma = 1$ to the *best result* for the RBCM *independently determined* for each k_{out}. To this end, we increment γ_{RB} by 20 steps per decade and calculate VI for each solution using the *known answer*. We then select the best γ_{RB} corresponding to the lowest VI average. This example is for a system with $N = 512$ nodes (see Fig. 4) with an average external degree $k_{out} \simeq 10$. See the text regarding other parameters defining the distribution of initial node degrees and community sizes. Each point is an average over 100 graphs.

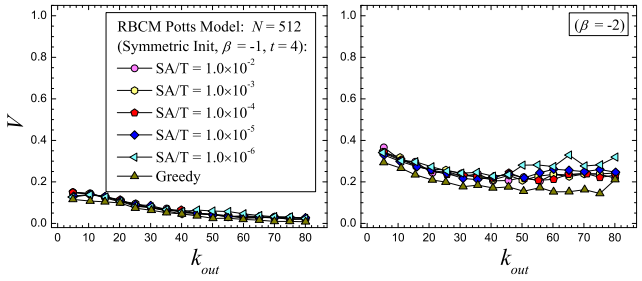

FIG. 11: (Color online) Plot of VI V vs the average external degree k_{out} for the RBCM. In Sec. V B, we generate a set of strongly defined communities with high levels of intercommunity noise. In Figs. 5 and 6, we vary the initial average power-law degree $\langle k \rangle_\alpha \simeq k_{out}$ and solve the networks using the greedy algorithm in Sec. III and SA where we use a starting temperature of $T_0 = 1.0 \times 10^{-4}$ for the $N = 512$ node systems. The greedy algorithm initialized into a symmetric initial state ($q_0 = N$) *outperforms* SA in accuracy when using the best γ_{RB} (see Appendix C). In these plots, we further examine SA for a range of starting temperatures. Even with significantly higher starting temperatures, SA cannot exceed the accuracy of the greedy algorithm in this problem.

to the APM of Eq. (2). A sample system is depicted in Fig. 4, and the accuracy results are summarized in Figs. 5 and 6. For the APM, we solve the system with the model weight $\gamma = 1$ for all graphs. Figure 10 shows an example of how we select the best result for the RBCM as compared to the known answer.

We start with $\gamma_{RB} = 0.1$ and geometrically increase the step size by $10^{1/20}$ (*i.e.*, 20 steps per decade of γ_{RB}). This example is for $N = 512$ nodes. The power-law distribution exponents are $\alpha = -2$ and $\beta = -1$ for the power-law degree and the community size distributions, respectively. Other parameters are: minimum and maximum community sizes $n_{\min} = 4$ and $n_{\max} = 50$, community edge densities $p_{in} = 1$, average external degree (noise) $\langle k \rangle_\alpha \simeq k_{out} \simeq 10$, maximum external degree $k_{\max} = 100$, and $t = 4$ trials per solution. Figure 10 shows only one *best* answer for the RBCM. We average over 100 graphs for each γ_{RB}, and the best VI average is plotted in Figs. 5 and 6 with the result for the APM.

Appendix D: NOISE TEST ANALYSIS OF SA AT DIFFERENT STARTING TEMPERATURES

In Sec. V B, we construct a set of maximally connected communities with varied levels of intercommunity noise. The systems are defined with a power-law distribution of community sizes and an approximate power-law distribution of external noise (see Sec. V B). We solve each system using the RBCM with the greedy algorithm in Sec. III and SA both with $t = 4$ trials. In Figs. 5 and 6, we use starting temperatures of $T_0 = 1 \times 10^{-4}$ for $N = 512$ and $T_0 = 1 \times 10^{-5}$ for $N = 4096$. Note that we scale the model energy by the number of edges in the system ($-1/L$) so that it is explicitly equivalent to the normalized modularity when $\gamma_{RB} = 1$. SA performs slightly *worse* in accuracy than the greedy algorithm of Sec. III using a symmetric initial state ($q_0 = N$).

Given this counter-intuitive result, in Fig. 11 we examine how the accuracy of the SA algorithm is affected by the algorithm's starting temperature. We plot the average VI V for the *best* RBCM result (see Appendix C) versus the average external degree k_{out}. We test starting temperatures spanning 5 orders of magnitude for $N = 512$ nodes. The cooling rate is fixed at $T_{i+1} = 0.999 T_i$ where each step i consists of N randomly proposed state changes. For the highest temperatures, the computational time dramatically increases due to a significantly longer cooling time with no significant improvement in accuracy. Thus, a higher starting temperature for SA cannot improve its performance sufficiently to match the accuracy of the greedy algorithm in this problem.

Appendix E: EVIDENCE OF A PHASE TRANSITION IN COMMUNITY DETECTION

In Fig. 12, we construct a set of well-defined but noisy systems with $N = 4096$ nodes from the benchmark in Sec. V B. An initial random degree distribution is as-

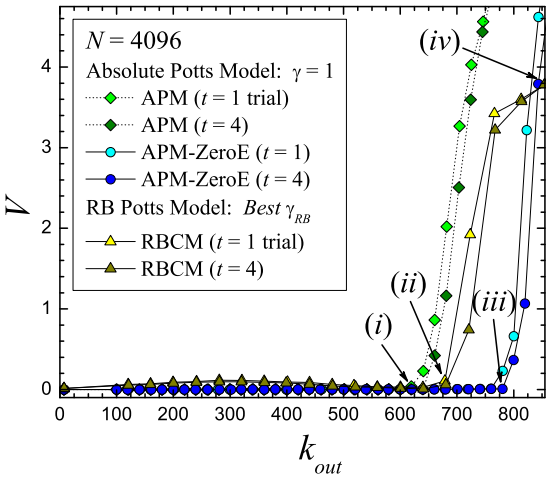

FIG. 12: (Color online) Plot of VI V vs the average external degree k_{out} for the APM and RBCM models. Similar to Sec. V B, we generate strongly defined communities with $N = 4096$ nodes and high levels of intercommunity noise. We use the greedy algorithm in Sec. III to solve the systems for both models. The system is initially assigned a random power-law degree distribution with an exponent $\alpha = -2$, maximum degree $k_{\max} = 1200$, and average degree $\langle k \rangle_\alpha$ ($k_{out} \simeq \langle k \rangle_\alpha$). Communities are assigned in a power-law distribution with an exponent $\beta = -1$, minimum size $n_{\min} = 8$, maximum size $n_{\max} = 24$, and density $p_{in} = 1$. The APM shows sharp accuracy transitions at (i) near $k_{out} \simeq 620$ (no zero energy moves) and at (iii) near $k_{out} \simeq 770$ (using zero energy moves). These roughly correspond to a similar transition for the RBCM at (ii). See the text regarding (iv). Each point is an average over 25 graphs.

signed according to a power law with an exponent of $\alpha = -2$, maximum degree $k_{\max} = 1200$, and average degree $\langle k \rangle_\alpha$. Community sizes are assigned in a power-law distribution with an exponent of $\beta = -1$. Minimum and maximum community sizes are $n_{\min} = 8$ and $n_{\max} = 24$, respectively. We then maximally connect internal community edges (density $p_{in} = 1$). We vary the average power-law degree $\langle k \rangle_\alpha$ (the average external degree $k_{out} \simeq \langle k \rangle_\alpha$) and solve the system with the APM and RBCM models using the algorithm in Sec. III.

Features (i) and (iii) correspond to two related "phase transitions" of the system into a "glassy" state for our APM. The "complexity" of the energy landscape dramatically increases near the "critical points" of $k_{out}^{(i)} \simeq 630$ (where we *disallow* zero energy moves) and $k_{out}^{(iii)} \simeq 770$ (where we *allow* zero energy moves), respectively. After the transitions, our algorithm in Sec. III is more easily trapped in a metastable state when navigating the energy landscape. In the intermediate region, the APM can still *almost perfectly* solve the system (in general problems, allowing zero energy moves does not usually result in such a drastic difference in accuracy). A second aspect of these transitions (not depicted) is a generally rapid rise in the computational effort required to solve the system which peaks near the respective critical points.

Feature (iii) at $k_{out}^{(iii)} \simeq 680$ shows that the RBCM displays a similar transition. We speculate that the more complicated energy landscape of the RBCM actually allows further optimization as compared to the APM when not utilizing zero energy moves in this problem. At feature (iv), the best RBCM solution (see Appendix C) approaches a trivial partition with $q > 3200$ communities.

This *community detection transition* is similar to transitions in the k-SAT (k-SATisfiability) problem found by Mézard *et al.* [54]. The authors showed that the most difficult solutions for k-SAT problems are found along well-defined loci in the phase diagram of random satisfiability problems. Figure 12 illustrates a similar transition in community detection. We will address the phase-transition aspects of community detection in more detail in an upcoming publication.

[1] L. Danon, A. Díaz-Guilera, and A. Arenas, J. Stat. Mech.: Theory Exp. 2006, P11010.
[2] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. **11**, 033015 (2009).
[3] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[5] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.: Theory Exp. 2008, P10008.
[7] J. J. Xu and H. Chen, ACM Trans. Inf. Sys. Secur. **23**, 201 (2005).
[8] N. Masuda, New J. Phys. **11**, 123018 (2009).
[9] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[10] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[12] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[13] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[14] H. N. Djijev, in *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006, Revised Papers* (Springer-Verlag, Berlin, Heidelberg, 2007), Vol. 4936, pp. 117–128.
[15] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[16] M. B. Hastings, Phys. Rev. E **74**, 035102(R) (2006).
[17] I. Ispolatov, I. Mazo, and A. Yuryev, J. Stat. Mech.: Theory Exp. 2006, P09014.
[18] S. Fortunato and M. Barthélemy, Proc. Natl. Aca. Sci. U.S.A. **104**, 36 (2007).
[19] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Euro. Phys. J. B **56**, 41 (2007).
[20] P. Ronhovde and Z. Nussinov, e-print arXiv:0803.2548v1.
[21] The Potts model in [16] also implicitly uses a version of a

missing edge penalty, but the (accurate) implementation of the model is defined with a mean-field approximation.
[22] P. Ronhovde and Z. Nussinov, Phys. Rev. E **80**, 016109 (2009).
[23] V. A. Traag and J. Bruggeman, Phys. Rev. E **80**, 036115 (2009).
[24] For *unweighted* graphs, the RBER model should be equivalent to the APM if we ignore the density dependence p representing the Erdős-Rényi null model. For example, on the benchmark in Sec. V B, the model would use $\gamma_{ER} \equiv \gamma_{RB} p = 1/2$ regardless of the system size or level of noise. See Appendix E for more discussion and Secs. VI E 1 and VI E 3 for differences between the models on weighted graphs.
[25] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. 2005, P09008.
[26] A. Noack and R. Rotta, in *Experimental Algorithms*, edited by J. Vahrenhold (Springer-Verlag Berlin, Heidelberg, 2009), Vol. 5526, pp. 257–268.
[27] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
[28] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[29] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).
[30] V. Gudkov, V. Montealegre, S. Nussinov, and Z. Nussinov, Phys. Rev. E **78**, 016113 (2008).
[31] Systems were solved on AMD Opteron computers at 2.2 – 2.8 GHz with up to 48 GB of random access memory.
[32] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[33] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, Phys. Rev. E **75**, 045102(R) (2007).
[34] In Fig. 1, q is *not constrained* during the solution dynamics for the RBCM/Symmetric data which results in a significantly lower accuracy than the data where $q = 4$ is fixed. However, the model can achieve essentially the same accuracy for the unconstrained solution if we begin with a random state near $q_0 = 4$ communities. Work by Chauhan *et al.* [55] may indicate the expected q in general problems, but it is uncertain whether this knowledge can be leveraged to improve the accuracy of a solution in general cases. In Sec. V B 3, for example, we use a random initial state with $q_0 \simeq q$, but the accuracy *decreases* compared to a solution beginning with $q_0 = N$.
[35] Data for the constructed system in Fig. 2 can be found at http://physics.wustl.edu/zohar/communitydetection/.
[36] We use a log scale for the model weight γ because of its relation to the minimum community edge density $p_{\min}$ in Eq. (3). γ identifies the targeted *resolution* of the partition; and a log scale better captures, as compared to a linear scale, how the typical community density varies in the range of practical importance $0 < \gamma \lesssim 19$ for unweighted graphs.
[37] We generate a power-law degree distribution and randomly fill it in decreasing order of node degree left to fill. Given the high level of noise in our benchmark in Sec. V B, if one randomly selects successive pairs of nodes when adding new edges (such as is done in the PLOD algorithm [56]), nodes with smaller degrees will "fill up" first leaving nodes with larger degrees to be increasingly connected to each other as a group.
[38] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[39] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
[40] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[41] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki, Phys. Rev. E **78**, 026109 (2008).
[42] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).
[43] X.-Q. Cheng and H.-W. Shen, e-print arXiv:0911.2308.
[44] M. J. Barber and J. W. Clark, Phys. Rev. E **80**, 026129 (2009).
[45] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[46] A. Arenas, J. Duch, A. Fernández, and S. Goméz, New J. Phys. **9**, 176 (2007).
[47] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. **10**, 053039 (2008).
[48] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Fluct. Noise Lett. **7**, L209 (2007).
[49] T. Heimo, J. M. Kumpula, K. Kaski, and J. Saramäki, J. Stat. Mech.: Theory Exp. 2008, P08007.
[50] D. J. Fenn, M. A. Porter, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones, Chaos **19**, 033119 (2009).
[51] J. Zhang, K. Zhang, X. ke Xu, C. K. Tse, and M. Small, New J. Phys. **11**, 113003 (2009).
[52] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[53] M. Meilă, J. Multivariate Anal. **98**, 873 (2007).
[54] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[55] S. Chauhan, M. Girvan, and E. Ott, Phys. Rev. E **80**, 056114 (2009).
[56] C. R. Palmer and J. G. Steffan, in *Global Telecommunications Conference, 2000. GLOBECOM '00* (IEEE, San Francisco, CA, 2000), Vol. 1, pp. 434–438.