5-24-2010

Physical Models in Community Detection with Applications to Identifying Structure in Complex Amorphous Systems

Peter Ronhovde
Washington University in St. Louis

WASHINGTON UNIVERSITY

Department of Physics

Dissertation Examination Committee:
Zohar Nussinov, Chair
Kenneth F. Kelton
Ralf Wessel
John W. Clark
Anders Carlsson
Roya Beheshti-Zavareh
Samuel Achilefu

PHYSICAL MODELS IN COMMUNITY DETECTION WITH APPLICATIONS TO IDENTIFYING STRUCTURE IN COMPLEX AMORPHOUS SYSTEMS

by

Peter Ronhovde

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2010

Saint Louis, Missouri

Abstract

We present an exceptionally accurate spin-glass-type Potts model for the graph theoretic problem of community detection. With a simple algorithm, we find that our approach is exceptionally accurate, robust to the effects of noise, and competitive with the best currently available algorithms in terms of speed and the size of solvable systems. Being a "local" measure of community structure, our Potts model is free from a "resolution limit" that hinders community solutions for some popular community detection models. It further remains a local measure on weighted and directed graphs. We apply our community detection method to accurately and quantitatively evaluate the multi-scale ("multiresolution") structure of a graph. Our multiresolution algorithm calculates correlations among multiple copies ("replicas") of the same graph over a range of resolutions. Significant multiresolution structures are identified by strongly correlated replicas. The average normalized mutual information and variation of information give a quantitative estimate of the "best" resolutions and indicate the relative strength of the structures in the graph. We further investigate a "phase transition" effect in community detection, and we elaborate on its relation to analogous physical phase transitions. Finally, we apply our community detection

methods to ascertain the most "natural" complex amorphous structures in two model glasses in an unbiased manner. We construct a model graph for the physical systems using the potential energy to generate weighted edge relationships for all pairs of atoms. We then solve for the communities within the model network and associate the best communities with the natural structures in the physical systems.

Acknowledgements

I would like to thank my advisor Zohar Nussinov for his exceptional patience, direction, and help throughout the research process. I am appreciative of the time and effort of Michael Ogilvie, Claude Bernard, Clifford Will, and Ramanath Cowsik. This thanks extends also to my dissertation examination committee Ralf Wessel, Kenneth F. Kelton, John W. Clark, Anders Carlson, Samuel Achilefu, and Roya Beheshti-Zavareh. I thank Mike Widom, Gilles Tarjus, and Nick Mauro for their input on elements of this work. Dandan Hu, Saurish Chakrabarty, Mousumi Sahu, and Kisor Sahu collaborated in various aspects of this work. Thanks to A. Lancichinetti and S. Fortunato for providing simulated annealing code which we modified in order to compare the accuracy of different Potts model approaches and to Wolfgang Weiser for providing a two-dimensional Ising lattice simulation. I also thank Mark Newman and UCINet for providing network data on their respective websites. This work was supported in part by the LDRD DR on the physics of algorithms at LANL and the Center for Materials Innovation at Washington University in St. Louis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

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. Figure 1.1, depicts such a network where the "natural" communities are identified by distinct node shapes and colors. "Community detection" describes the problem of finding these closely related sub-groups within a general network. Regardless of the particular approach used to solve the problem, the goal 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.

Applications of the problem are wide since an extremely broad array of applica-

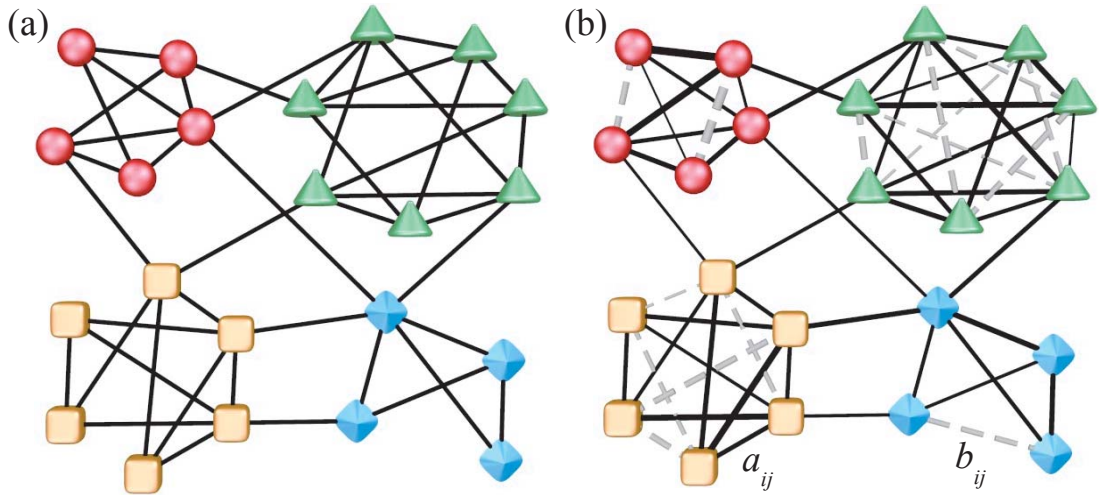


Figure 1.1: The two panels show a small network with 4 natural communities, depicted as distinct node shapes, that are strongly connected. Panel (a) depicts an *unweighted* version, and panel (b) shows the *weighted* network. In general, the edges could also be assigned in a particular direction between nodes. Regardless of the approach used, the goal in community detection is to identify any such strongly related clusters of nodes based on their defined edge relationships. In either panel, solid lines depict links corresponding to complimentary or attractive relationships where $a_{ij} > 0$ and $b_{ij} = 0$ in Eq. (2.2). In panel (b), gray dashed lines depict missing, adversarial, or repulsive relationships where $a_{ij} = 0$ and $b_{ij} > 0$. The relative link weight is indicated by the respective line thicknesses. For presentation purposes, missing intercommunity relationships are not depicted.

tion may be cast into this network representation. Examples include 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 introduction to the "physics

of networks" can be found in Ref. [10]. Reviews of the field are found in Refs. [11, 12], and the most recent thorough review is found in Ref. [13].

One method of quantitatively assessing community structure is through the use of a "quality function." A quality function essentially provides an objective measure of how "clustered" or "modular" a network is. Examples include the prominent modularity measure defined by Newman and Girvan [14], a Potts model originally proposed by Reichardt and Bornholdt (RB) [15, 16], our Potts model [17] that eliminates the random partition applied by RB, an application of a Potts model utilizing a mean-field approximation with "belief propagation" [18], and another measure "fitness" [2]. Other approaches to community detection include clique percolation [4, 19], spectral [20], information theoretic [21, 22] "label propagation" [23, 24], dynamical [25, 26], and maximum likelihood [27]. Karrer *et al.* [28] defined a measure of robustness of community structure based on random perturbations. Some efforts enhance or expand applications to more general systems such as weighted networks [29, 16, 17], heterogeneous systems [1, 17], bipartite graphs [30, 31], overlapping nodes [32, 4, 2, 31, 15], and multiresolution methods [33, 34, 35, 2, 36, 37, 38, 39, 40].

1.2 Challenges

As summarized above, there are many approaches and areas of investigation for community detection. Our community detection methods provides a strong, exceptionally accuarate, solution for the community detection problem, and they answer two par-

ticular challenges in the field. First, certain very popular models have been shown to have an implicit limitation in the smallest communities that can be resolved in a large system. Second, a more recent focus in the field is the question of how to determine the "best" scale(s) at which to solve a system.

The most popular quantitative community measure is that of "modularity" which was originally introduced by Newman and Girvan [14]. 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 [41] and [15]. Our particular model was originally inspired by a minimal cut method by Djidjev [42] which is equivalent to modularity. The resulting generalized Hamiltonian was previously presented by Reichardt and Bornholdt (RB) [16]. 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. Other Potts model approaches are by Hastings [18], which casts community detection as an inference problem, and Ispolatov *et al.* [43], which extends the Potts models in [41, 15].

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 intracommunity links and fewer intercommunity 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. 2.5.1). 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.

One challenge in the field is that Fortunato and Barthélemy [44] 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 [45] due to how it is cast by design, analogous to modularity, in terms of an arbitrary null model comparison. The number of communities 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 [45] regardless of the null model that is used.

An additional challenge is that the most natural organized community structure can depend on the scale at which the system is examined. Different scales correspond to distinct community divisions at different internal community edge densities. For many systems, including those with hierarchical organization, a "multiresolution" approach [46] is needed to capture the overall structure and the relationships between the elements at different resolutions. Examples of such systems can include biological processes [47, 48], food webs [49], air transportation networks [48], and communication networks [6]. Thus, multiresolution methods are an important extension of problems in community detection.

Hierarchical organization is the most obvious type of multiresolution structure. Some earlier work on hierarchies in graphs can be found in [50, 47]. Examples of more recent efforts in analyzing hierarchical structures in graphs are [51, 6, 2, 48, 33]. Arenas *et al.* [33] defined a multiresolution method using modularity that makes novel use of the resolution limit [52]. Reichardt and Bornholdt [15], Arenas *et al.* [33], Kumpula and co-workers [34], Heimo *et al.* [35], and Fenn *et al.* [36] also study multiresolution applications of an RB Potts model.

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 resolution limit problems encountered in some models by avoiding a null model comparison [17]. Instead, it penalizes for missing edges directly in the energy sum [53]. 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 relative null model. One consequence of our approach is that it removes the ability of the model to automatically scale the solution based on global properties of a graph (see Sec. 2.5.1), but the change results in a robust model with significant improvements to several desirable properties. Further, the multiresolution algorithm presented here *quantitatively* determines the "best" network scale(s) by evaluating the strength of correlations among independent partitions ("replicas") of the same graph over a range of resolutions.

1.3 Overview of thesis

This dissertation contains information related to the following publications or manuscripts in preparation roughly arranged into chapter divisions as indicated:

- Chapters 1 and 2: P. Ronhovde and Z. Nussinov, *Local resolution-limit-free Potts model for community detection*, Phys. Rev. E **81**, 046114 (2010).
- Chapters 1 and 3: P. Ronhovde and Z. Nussinov, *Multiresolution community detection for megascale networks by information-based replica correlations*, Phys. Rev. E **80**, 016109 (2009).
- Chapter 4: D. Hu, P. Ronhovde and Z. Nussinov, *Phase transition in the community detection problem: spin-glass type and dynamic perspectives*, e-print arXiv:1008.2699 (2010).

- Chapter 5: P. Ronhovde, S. Chakrabarty, M. Sahu, K. K. Sahu, K. F. Kelton, and Z. Nussinov, *Detecting hidden spatial and spatio-temporal structures in glasses and complex systems by multiresolution network clustering*, (in preparation, 2010).

In Chapter 2, we demonstrate a simple but effective implementation of the q-state Potts model to community detection. In Sec. 2.1, 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. 2.2, we present our algorithm, and Sec. 2.3 illustrates its accuracy compared to several other approaches. Our Potts model and the RBCM model are directly compared in Sec. 2.4. Issues regarding local and global measures and the resolution limit for general graphs are addressed in Sec. 2.5. We solve two examples in Sec. 2.6 and conclude the chapter in Sec. 2.7.

In Chapter 3 we show how information theory based measures may be used to systematically and quantitatively extract the best community partitions on *all* scales. This will enable us to methodically determine the hierarchical or multiresolution structure of arbitrary networks. In Sec. 3.1, we discuss the application of our Potts model and community detection algorithm to multiresolution analysis. We then present several examples in Sec. 3.2. The exceptional accuracy of the multiresolution algorithm is addressed in Sec. 3.3, and we conclude the chapter in Secs. 3.4 – 3.6.

Chapter 4 relates details regarding a community detection transition which elaborates on how certain physically motivated features of our Potts model manifest

themselves in terms of the community detection problem. We present the heat bath algorithm in Sec. 4.2. Static transitions are discussed for temperatures $T = 0$ in Sec. 4.3 and for $T > 0$ in Sec. 4.4. A closely related "dynamic" transition is shown in Sec. 4.5, and the chapter concludes in Sec. 4.6. In this chapter, the author's main contributions consist of specifically identifying the existence of a phase transition (particularly in the noise test benchmark in Sec. 4.3.1), developing the base community detection model and computer code along with Zohar Nussinov, working closely with Dandan Hu in the writing the heat bath algorithm that is used to further analyze phase transition in community detection, and collaboration in the resulting analyses of this aspect of the problem.

Chapter 5 illustrates a concrete application of these community detection methods to identify structures in amorphous systems, using a model metallic glass and a binary Lennard-Jones systems, in particular. In Secs. 5.1 and 5.2, we introduce a number of concepts in glasses and amorphous systems and how we will relate them to our community detection problem. We explain the simulation details in Sec. 5.3. The corresponding results are given in Sec. 5.4, and we conclude in Sec. 5.5. In this chapter the author's contributions include the multiresolution network and visualizaton analysis applied to the problem. M. Sahu and K. K. Sahu contributed to the initial stages of the multiresolution analysis. S. Chakrabarty was responsible for molecular dynamics simulations and configurations. S. Chakrabarty and M. Widom contributed to the potential models that were used for the set of potentials applied to one model glass former. K. F. Kelton oversaw experimental work on a metallic glass that is

related to our model glass former.

In Appendix A, we explain the variation of information (VI) metric, the normalized mutual information (NMI) measure, and other information measures which we use in several sections of the thesis. Appendix B argues that the *unweighted* variant of the RBER model can be strengthened to eliminate the resolution limit. Appendices C and D elaborate on some details related to comparing different Potts model approaches in the proposed noise test benchmark in Sec. 2.4.2. Appendix E explains a generalization of our replica method for other, non-graph theoretic, optimization problems. Appendix F elaborates on some details related to the benchmark accuracy test discussed in Sec. 3.3. Appendix G explains the overlapping dynamics which we use in Chapter 5. Appendices H – K present several additional test cases for the multiresolution method in Chapter 3 specifically relating to its application to complex amorphous materials in Chapter 5.

Chapter 2

Community detection

2.1 Potts model Hamiltonians

One of the most popular approaches in community detection is to define and objective quality function that will indicate the "best" community divisions when it is optimized over competing divisions of a graph. Such quality functions evaluate the best community divisions based on at least two criteria: The first obvious contribution is that edges inside a community strengthen the community. In order to consistently avoid a trivial solution (a single community) in general, a quality function must also apply a "penalty function" in some form. The most common penalty function method compares the community edge distributions to an "expected" value based on how a candidate division compares to a selected null model (a particular randomized representation of the graph). This particular penalty method has the unintended side effect of introducing an inherent limitation, a *resolution limit*, in the smallest size

communities that may be properly resolved in large networks. We elaborate on a different simple, yet powerful, approach.

2.1.1 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.

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 [16]

$$\mathcal{H}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(a_{ij}A_{ij} - b_{ij}J_{ij}\right)\left[2\delta(\sigma_i,\sigma_j) - 1\right] \tag{2.1}$$

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. (2.1) 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

$$\mathcal{H}(\{\sigma\}) = -\frac{1}{2}\sum_{i \neq j} \big(a_{ij}A_{ij} - \gamma b_{ij}J_{ij}\big)\delta(\sigma_i, \sigma_j). \tag{2.2}$$

Spins interact only with other spins in the same community ($\sigma_i = \sigma_j$). The generality of the weights ($\{a_{ij}\}$ and $\{b_{ij}\}$) [54, 37] enables the study of directed graphs, weighted graphs, and graphs with missing link weights (*i.e.*, levels of "repulsion"). Traag and Bruggeman [55] 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. (2.2) describes a system wherein spins in the same community interact ferromagnetically if they are connected and antiferromagnetically if they are not connected. We split the "attractive" (ferromagnetic) and "repulsive" (anti-ferromagnetic) contributions into two separate weighted matrices so that we can insert the model weight γ that adjusts the energy trade-off between the two types of interactions. The new parameter γ has the effect that it allows the model to adjust the scale or *resolution* of the community solution. We identify communities by minimizing Eq. (2.2), and despite a global energy sum, our model is a *local* measure

of community structure (see Sec. 2.5). We refer to Eq. (2.2) as an *"absolute" Potts model* (APM) as it is not defined relative to a null model. Although our analysis here will focus on the static APM, it is defined for both *static* systems and *dynamic* networks with time-dependent weights and adjacency matrices.

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. (2.2) is related to the *minimum* edge density of each community,

$$p_{\min} \geq \frac{\gamma}{\gamma+1}, \tag{2.3}$$

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. 2.5).

Community and node stability

From Eq. (2.2), the interaction energy E_{rs} between communities r and s is

$$E_{rs} = -w_{rs} + \gamma u_{rs} \tag{2.4}$$

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. (2.4). 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 (2.4) indicates the strong local behavior of the APM (see Sec. 2.5). 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.

2.1.2 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 [16]

$$\mathcal{H}_{RB}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{_{RB}} p_{ij}\right)\delta(\sigma_i,\sigma_j) \tag{2.5}$$

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. $\gamma_{_{RB}}$ 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. 1.1). For later reference, they are explicitly given by $p_{ij} = p$ for the Erdős-Rényi null model

$$\mathcal{H}^{ER}_{RB}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{_{RB}} p\right)\delta(\sigma_i,\sigma_j) \tag{2.6}$$

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

$$\mathcal{H}^{CM}_{RB}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left(A_{ij} - \gamma_{_{RB}}\frac{k_i k_j}{2L}\right)\delta(\sigma_i,\sigma_j), \tag{2.7}$$

where k_i is the degree of node i. Equation (2.7) 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. (2.7) is equivalent to modularity [16] up to a scale factor of $-1/L$. The APM can be made equivalent to the RBER model for *unweighted* graphs [56] (see also Appendix B). We address weighted generalizations of both models and their effect on model locality in Sec. 2.5.5. Despite the similar forms of the Hamiltonians of Eqs. (2.2) and (2.5), 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. 2.4.

2.2 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 [54]. A summary of the efficiency of several algorithms appears in [57]. A number of algorithms were compared in [58] and [59] 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. 2.1 and 2.9), 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, 60, 23, 25]. Steps (2) and (3) are the fundamental elements of the algorithm which are similar to portions of algorithms used elsewhere [6, 23]. 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. 2.1). The number of trials t is generally $O(10)$ or less.

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 [61] (see Sec. 2.6).

2.3 Accuracy compared to other algorithms

We test the accuracy of our method compared to several other algorithms using a common benchmark [62]. 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

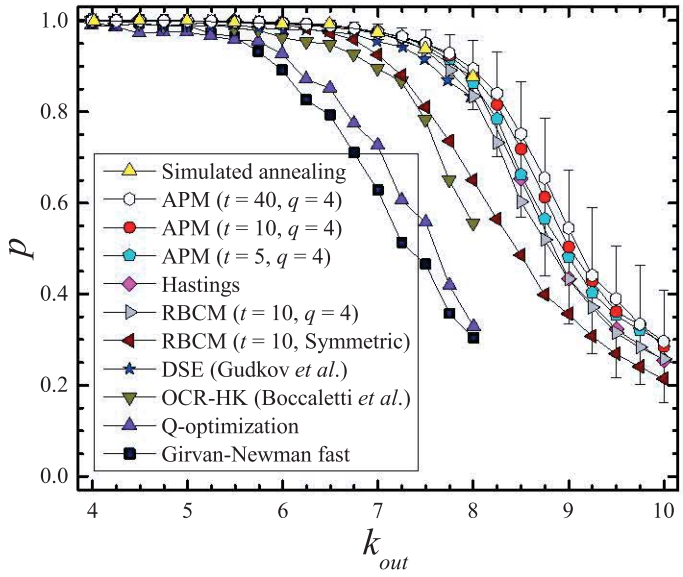


Figure 2.1: Plot of the percentage of correctly identified nodes p versus the average external degree k_{out} [62]. The average node degree is $k = 16$. The data for the APM of Eq. (2.2) and the RBCM of Eq. (2.7) both use $\gamma = \gamma_{RB} = 1$. Both models use the algorithm in Sec. 2.2 with $q = 4$ 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.

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. 2.1, we plot the "percentage" of correctly identified nodes p as a function of k_{out}. For consistency with other data in Fig. 2.1, we use the same measure of percentage accuracy as Newman [62]. 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 accuracy rate of 95% or better up to $k_{out} = 7.5$.

Several sets of data were assimilated by Boccaletti *et al.* [26] where the most accurate algorithm was simulated annealing (SA) although it is computationally demanding [57]. Other accurate data are by Hastings [18], Gudkov *et al.* [25], and our algorithm in Sec. 2.2 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) [63]. A recent analysis [59] showed our multiresolution algorithm [37] applied to this benchmark using the APM with unconstrained q where it was also very accurate, among the best of tested algorithms.

2.4 Accuracy comparison of Potts models

We compare the APM of Eq. (2.2) to the RBCM of Eq. (2.7) with two test systems. First, we solve for the different levels of the synthetic hierarchy depicted in Fig. 2.2 with the results given in Fig. 2.3. 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. 2.4 with the results given in Figs. 2.5 and 2.6. The APM proves

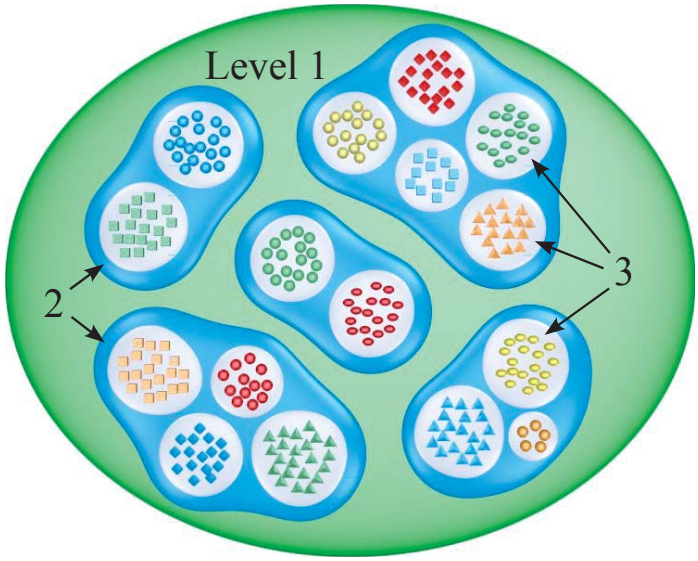


Figure 2.2: The figure depicts a simulated three-level heterogeneously-sized hierarchy with $N = 256$ nodes [37, 64]. The innermost level 3 has $q_3 = 16$ communities with a randomly assigned average density of $\overline{p}_3 = 0.90$. The intermediate level 2 has $q_2 = 5$ communities and an average 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 $p_1 = 0.1$.

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

2.4.1 Three-level hierarchy

We identify two levels of a constructed hierarchy [64, 37] using both the APM and RBCM models. The three-level hierarchy is depicted in Fig. 2.2, and the results are

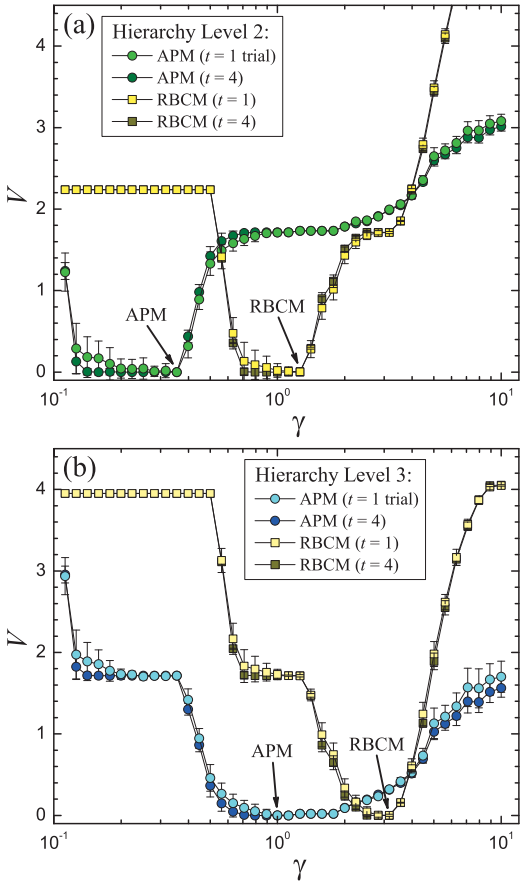


Figure 2.3: Plot of VI V vs model weights γ or γ_{RB} for the APM of Eq. (2.2) and the RBCM (configuration null model) of Eq. (2.7), respectively. The plots illustrate how the model weights operate in the respective models. We use the algorithm in Sec. 2.2 for both models to identify the hierarchy depicted in Fig. 2.2 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.

given in Fig. 2.3. The system has $N = 256$ nodes divided into $q_3 = 16$ communities at level 3 with sizes as noted in Fig. 2.2. 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. 2.2 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. 2.3, we plot VI V as a function γ or γ_{RB} [65] 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).

2.4.2 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

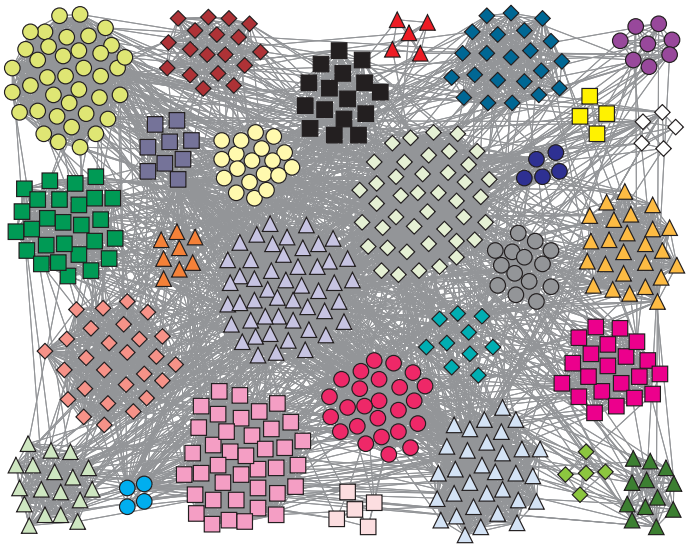


Figure 2.4: A sample graph with $N = 512$ nodes for the noise test in Sec. 2.4.2. In this sample, the node degrees are initially defined in a power-law distribution with an average $\langle k \rangle_\alpha = 5.4$, maximum $k_{\max} = 100$, and exponent $\alpha = -2$. Communities have a power-law distribution 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 ($p_{in} = 1$).

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.

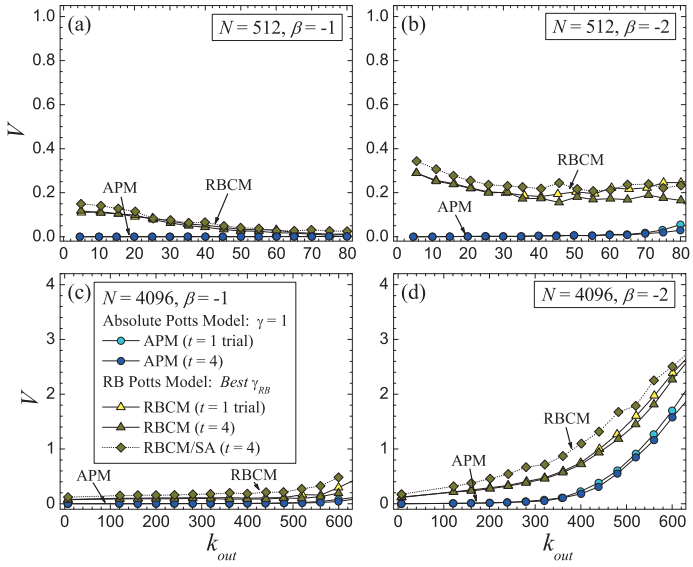


Figure 2.5: Plot of VI V between solved and known test systems in Sec. 2.4.2 as a function of the average external node degree k_{out}. The system sizes are $N_{\mathrm{a,b}} = 512$ in panels (a) and (b) (with a sample system depicted in Fig. 2.4) and $N_{\mathrm{c,d}} = 4096$ nodes in panels (c) and (d). The graphs are solved with the APM and the RBCM using the algorithm in Sec. 2.2 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 [66] with an exponent $\alpha = -2$, an average degree $\langle k \rangle_\alpha$, and maximum degrees of $k^{\max}_{\mathrm{a,b}} = 100$ or $k^{\max}_{\mathrm{c,d}} = 1000$, respectively. Constructed communities are randomly assigned in a power-law size distribution [67] specified by an exponent $\beta_{\mathrm{a,c}} = -1$ or $\beta_{\mathrm{b,d}} = -2$, minimum size $n_{\min} = 4$, and maximum size $n_{\max} = 50$. 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 Sec. 4.3.1 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).

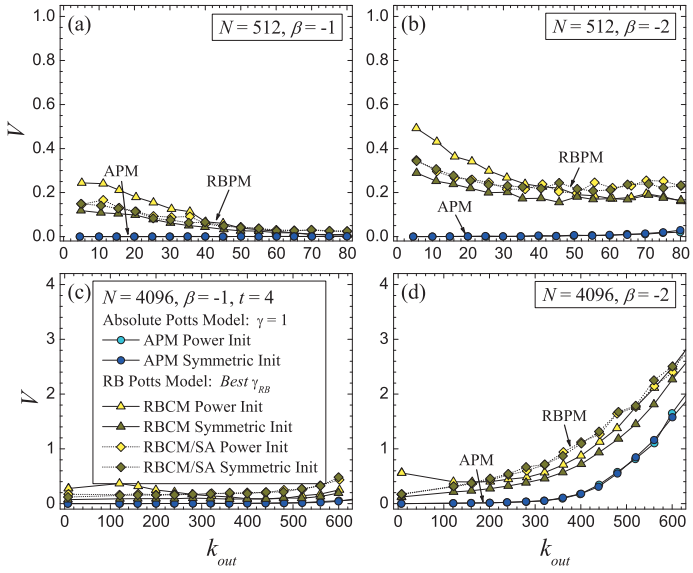


Figure 2.6: Plot of VI V between solved and known test systems in Sec. 2.4.2 vs the average external node degree k_{out}. In panels (a) through (d), system sizes are $N_{\mathrm{a,b}} = 512$ and $N_{\mathrm{c,d}} = 4096$ nodes, respectively. The graphs are solved with the APM and the RBCM using the algorithm in Sec. 2.2 with $t = 4$ trials. The constructed configurations are identical to those used in Fig. 2.5. 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).

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 $k_{\max}$ [66]. This initial framework is similar to a benchmark by Lancichinetti *et al.* [67, 59]. 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. 2.4.2). 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. 2.2 for both models with $t = 1$ and 4 trials and using SA for the RBCM with $t = 4$.

Accuracy results

Figures 2.4 and 2.5 show a sample system and the first test results, respectively. We use two system sizes of $N_{\mathrm{a,b}} = 512$ and $N_{\mathrm{c,d}} = 4096$ nodes, respectively. The initial power-law degree distribution uses $\alpha = -2$; and the maximum degree constraints are $k_{\mathrm{a,b}}^{\max} = 100$ and $k_{\mathrm{c,d}}^{\max} = 1000$, 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 $\beta_{\mathrm{a,c}} = -1$ or $\beta_{\mathrm{b,d}} = -2$, 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 ($V_{\mathrm{a,b}}^{\max} = 9$ and $V_{\mathrm{c,d}}^{\max} = 12$). 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* $\gamma_{_{RB}}$ 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 $\gamma_{_{RB}}$ for comparison. We average over 100 graphs for each point in panels (a) 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 $k_{\max}$ 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. 2.5(d) are due to transition effects described in Chapter 4. We subjectively select the best $\gamma_{_{RB}}$ here, but note also that our algorithm can slightly outperform SA in accuracy in many cases *for either Potts model* (see Sec. 2.4.2 and Appendix D). See Sec. 2.5.4 regarding how the high

levels of noise in this test actually mitigate the effect of the resolution limit for the RBCM.

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. 2.2. We further elaborate on the accuracy of SA compared to this greedy algorithm.

In Fig. 2.6, 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 ($\beta_{\mathrm{a,c}} = -1$ or $\beta_{\mathrm{b,d}} = -2$, $n_{\min} = 4$, and $n_{\max} = 50$) 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. 2.1 in Sec. 2.3 where the symmetric initialization performs worse than a random initial state with $q_0 = 4$

communities (with or without constraining q in the dynamics [63]) 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 [59] compared the accuracy of several algorithms using their benchmark [67]. One result in [59] showed that a similar greedy algorithm optimizing modularity [equivalent to $\gamma_{RB} = 1$ in Eq. (2.7)] by Blondel *et al.* [6] also outperformed SA in accuracy on that benchmark. See also Good *et al.* [68] regarding difficulties associated with modularity optimization in practical problems.

Noise tolerance discussion

Even at low levels of noise, these benchmark graphs exceed the proposed definition of so-called "weak" communities [69], 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 $(q_{\mathrm{b}} - 1) \simeq 370$ 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. (2.7)] 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. 2.5 and 2.6 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. (2.2)]. 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 [37, 59].

2.5 Resolution limit

The quantitative approaches of modularity [14] and the RBPM [15, 16] in Sec. 2.1.2 were implemented by incorporating global graph parameters into the models. Both models marked important progress in the field of community detection, but Refs. [44] and [45] 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 $\sqrt{L}$ [44] and $\sqrt{\gamma_{RB} L}$ [45], 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.

2.5.1 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 [44, 45] 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, 19], analyzing random walks [21, 39], "label propagation" [23] (and one variant in [24]), and local variants of modularity [5, 70]. See Sec. 2.5.5 and Appendix B 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 [71] 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 [59, 39] are based on essentially local methods or models of community detection [21, 37, 39].

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. (2.2)] 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. (2.3) 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. 2.1.1) are the best partitions for the network. Arenas *et al.* [33] varied a weight parameter with modularity and tracked stable partitions. Kumpula and co-workers [34, 35] as well as Fenn *et al.* [36] also explored stability approaches for the RBCM. Our multiresolution method [37] utilized information comparisons among independent solutions to quantitatively evaluate the best resolutions. Zhang *et al.* [38] used a topological weighting strategy. Cheng and Shen [39] used the stability of random walker diffusion dynamics to identify the most relevant resolutions.

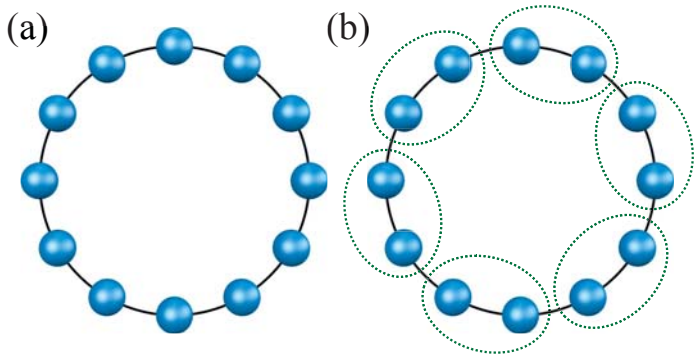


Figure 2.7: Each graph is a circle of cliques with N total nodes and L total edges. (a) A set of q cliques with m nodes each are connected in a circle by q links. (b) r consecutive cliques are each grouped together. Intuitively, one would expect that any measure should resist merging these communities on any system scale (e.g., N, L, or q) if $m \geq 3$.

2.5.2 Circle of cliques

Fortunato and Barthélemy [44] and Kumpula *et al.* [45] identified a resolution limit in the respective models in part by considering the unweighted system shown in Fig. 2.7, a set of q cliques (maximally connected communities) connected in a circle by single edges. In Fig. 2.7(a), 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. (2.2), the APM energy is

$$E_{\mathrm{a}} = -\frac{1}{2}qm(m-1). \tag{2.8}$$

This energy E_{a} has *no finite extremum* with respect to *any* global parameters of the graph. The analogue to Eq. (2.8) 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 2.7(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. (2.7), using the configuration null model, includes modularity as a special case when $\gamma_{RB} = 1$. Two neighboring cliques ($r = 2$) [45] will merge if

$$\gamma_{RB} < \frac{q}{m(m-1)+2}. \tag{2.9}$$

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. 2.5.4). 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. (2.6), neighboring cliques merge if

$$\gamma_{RB} < \frac{q - 1/m}{m(m-1)+2}. \tag{2.10}$$

We can always choose q large enough to induce a merger of neighboring cliques for any γ_{RB} (see also Appendix B). These results generalize so that a resolution limit can be found to apply for an arbitrary choice of null model [45] when using the RBPM of Eq. (2.5). The APM energy E_{b} of the configuration in Fig. 2.7(b) with r merged cliques is

$$E_{\mathrm{b}} = -\frac{\gamma+1}{2} qm(m-1)\left[1 - \frac{\gamma}{\gamma+1}\frac{rm-1}{m-1} + \frac{2(r-1)}{rm(m-1)}\right]. \tag{2.11}$$

We compare the energies in Eqs. (2.8) and (2.11) and find that $r = 2$ cliques merge if

$$\gamma < \frac{1}{m^2 - 1}. \tag{2.12}$$

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 $\gamma < 1/8$.

2.5.3 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 [62]. The APM deals with heterogeneous communities naturally.

Figure 2.8(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. (2.5), using a generic null model, the number of edges l that causes communities A and B to merge is of the order [45]

$$l \gtrsim \frac{\gamma_{RB} nm}{N}. \tag{2.13}$$

The RBER model yields a merge condition of

$$l > \frac{2L}{N(N-1)} \gamma_{RB} nm. \tag{2.14}$$

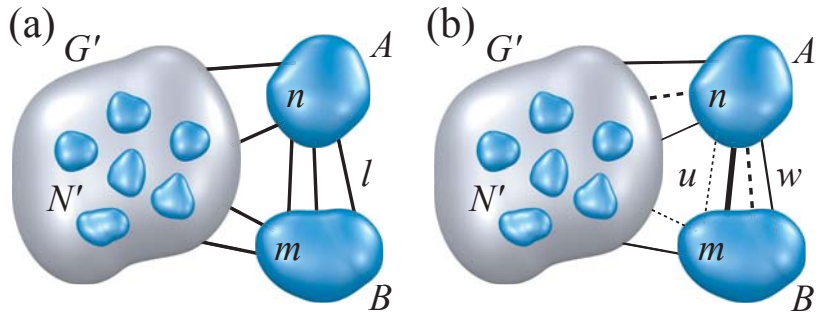


Figure 2.8: A large graph G has N nodes and three sub-divisions depicted, one potentially large sub-graph G' with N' nodes and two distinct communities A and B with n and m nodes, respectively. In panel (a), G is unweighted, and communities A and B are joined by l edges. In panel (b), G is a weighted 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.

In Eqs. (2.13) and (2.14), 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. 2.5.4 and Appendix B).

Our APM model merges communities A and B if

$$l > \frac{\gamma}{\gamma + 1} nm. \tag{2.15}$$

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 inter-connections. The dependence on γ is consistent with the purpose of its introduction

in Eq. (2.2) — to allow the model to vary the system resolution.

2.5.4 Mitigated resolution limit

We return to the unweighted system of q cliques in Fig. 2.7 and Sec. 2.5.2 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. (2.9). The new merge condition is

$$\gamma_{RB} < \frac{q\ell_{out}}{(2\ell_{in} + s\ell_{out})}. \tag{2.16}$$

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. 2.4.2, Eq. (2.16) 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.

2.5.5 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.

Absolute Potts model

We generalize results from Sec. 2.5.3 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. (2.2) becomes

$$\mathcal{H}_s(\{\sigma\}) = \sum_s \big(- w_s + \gamma\, u_s \big) \tag{2.17}$$

where w_s and u_s are the energy sums of *connected* and *missing* edges of community s, respectively. For reference in Sec. 2.5.5, the unweighted version of Eq. (2.17) is

$$\mathcal{H}_s(\{\sigma\}) = \sum_s \big[- (\gamma + 1)\, l_s + \gamma l_s^{\max} \big] \tag{2.18}$$

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

In Fig. 2.8(b), we use Eq. (2.17) 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

$$w > \gamma u. \tag{2.19}$$

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. 2.5.5).

Weighted configuration RB Potts model

A weighted generalization of the RBCM model is

$$\mathcal{H}^{w}_{CM}(\{\sigma\}) = \sum_{s} \left(-w_s + \gamma_{RB} \frac{W_s^2}{4W} \right) \tag{2.20}$$

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.

Weighted Erdős-Rényi RB Potts model

The weighted generalization of the RBER model of Eq. (2.6) increases the global dependence of the model as it is proposed in [15]. We rewrite the *unweighted* RBER model as a sum over communities

$$\mathcal{H}_{ER}(\{\sigma\}) = \sum_{s} (-l_s + \gamma_{RB} p\, l_s^{\max}). \tag{2.21}$$

Equations (2.18) and (2.21) 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 B).

We write a conceptual generalization of Eq. (2.21) for weighted graphs which we use again in Sec. 2.5.5,

$$\mathcal{H}^{w}_{ER}(\{\sigma\}) = \sum_{s} (-w_s + \gamma_{RB} p\, w_s^{\max}). \tag{2.22}$$

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 [15]

$$\mathcal{H}^{w}_{ER}(\{\sigma\}) = \sum_{s} \left(-w_s + \gamma_{RB} p \overline{W}\, l_s^{\max}\right) \tag{2.23}$$

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

In Fig. 2.8(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. (2.23), the merge condition is

$$w > \gamma_{RB} p \overline{W} nm. \tag{2.24}$$

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 sub-graph 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.

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

One can modify the weighted RBER model of Eq. (2.22) 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. (2.22) 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 [15]. See also Appendix B.) In removing the density dependence p, the model is technically no longer an "Erdős-Rényi" Potts model; however, in this interpretation, Eq. (2.22) simplifies to

$$\mathcal{H}_{ER}^{local}(\{\sigma\}) = \sum_s \overline{w}_s\,(-l_s + \gamma_{_{ER}}\, l_s^{\max}). \tag{2.25}$$

This variant uses almost the same energy sum as the *unweighted* RBER model in Eq. (2.6) except that total energy of community s is weighted by $\overline{w}_s$. Equation (2.25) 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.

2.6 Examples

We demonstrate the Potts model with (A) one real-world example and (B) a very large constructed system of 40×10^6 nodes and over 1×10^9 edges.

2.6.1 Zachary karate club

A common test is the Zachary karate club [72]. 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. (2.2) with $b_{ij} = 1$ and divide the graph into two parts by constraint as shown in Fig. 2.9. 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 [15].) In the actual division, node 10 associated with the group depicted by circles. A more complete multiresolution

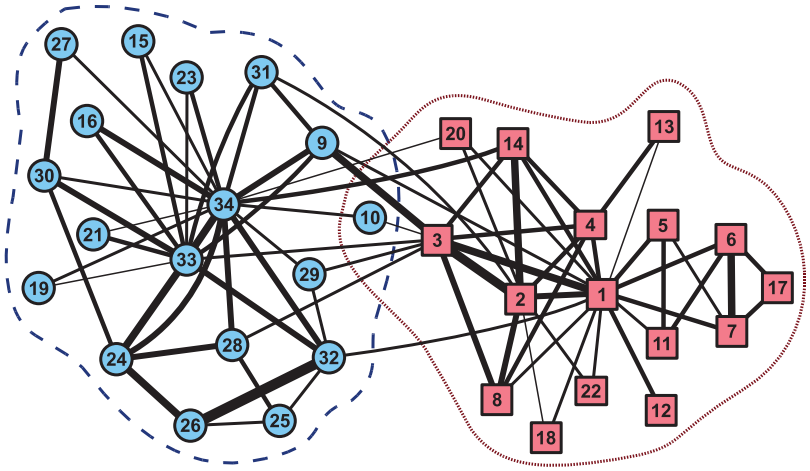


Figure 2.9: Graph depicts the Zachary karate club [72] solved with the APM using weighted edges (relative line thickness), $\gamma = 1$, and $q = 2$ communities by constraint. All nodes except 10 are correctly assigned. By our analysis, this node appears frequently in both groups (see text).

analysis in [37] would correctly place node 10.

2.6.2 Very large system

We also construct a very large system similar to those defined in Sec. 2.4.2. 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. (2.2) with the algorithm in Sec. 2.2 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 [61].

2.7 Conclusion

We present an exceptionally accurate 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 system noise. (4) It is a local measure in a strong sense for unweighted, weighted (including weighted "adversarial" relationships), and directed graphs. As such, it corrects a resolution-limit problem that affects other popular measures [44, 45]. (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 network. We have accurately solved synthetic systems as large as 40×10^6 nodes and over 10^9 edges [61]. In Ref. [37], we illustrated in detail how this core community detection method may be extended to systematically, accurately, and rapidly identify general multiresolution structures.

Chapter 3

Multiresolution community detection

3.1 Multiresolution approach

One challenge in developing a multiresolution community detection algorithm is that of selecting the best resolution(s) for the system. A straight-forward method that avoids the choice of the best resolution is to iteratively solve the system (with a necessary change in γ for our model) and collapse the communities into "supernodes" until the system is organized into a forced hierarchical structure. This approach is viable; but even when the system is hierarchical in nature, there is the question of whether the best resolutions were resolved at each stage. Our algorithm enables a quantitative analysis that determines the best resolutions and applies to general types of multiresolution structure.

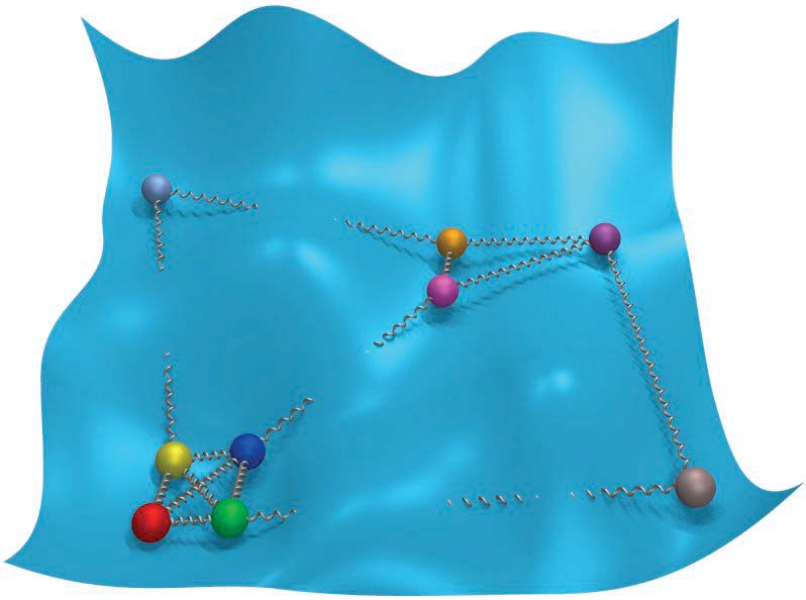


Figure 3.1: A depiction of several replicas, represented as "marbles," in an energy landscape. "Interactions" between replicas are depicted by springs which represent information correlations among the replicas (with only a few shown for presentation purposes).

3.1.1 Motivation

Ideally, we desire an algorithm that allows the system to communicate what the best resolutions are; but without *a priori* information, the correct weights for these resolutions are not obvious in general. In order to identify the proper resolutions, we examine information-based correlations among independent replicas (independent solutions) via NMI or VI over a range of resolutions as depicted in Fig. 3.1. Rather than using the replicas to simply identify a unique optimized solution for each resolution, we examine correlations among the entire set. We then select the strongest correlations as the best resolutions.

From a global perspective, the average NMI (between all pairs of replicas) indicates how strongly a given structure dominates the energy landscape by measuring how well the replicas agree with each other. High values of the NMI (often manifested as peaks) correspond to more dominant, and thus more significant, structures. From a local perspective, at resolutions where the system has well-defined structure, a set of independent replicas should be highly correlated because the individual nodes have strongly preferred community memberships. Conversely, for resolutions "in-between" two strongly defined configurations, one might expect that independent replicas will be less correlated due to "mixing" between competing divisions of the graph. Random effects will usually reduce the correlations between independent solutions.

A similar argument applies to VI where, as an information distance, low values of VI correspond to better agreement among replicas. With these information-based correlations, we obtain a set of multiresolution partitions of the graph, but we also obtain an estimate of the relative strength of the structures at each resolution. Note that this argument does not distinguish between unrelated multiresolution structures or those that are strictly hierarchical in nature although nothing prevents the imposition of additional hierarchical constraints if desired.

Implicit in this argument is the idea that local minima in the energy landscape represent meaningful, even if perhaps incomplete, information about the graph. The same assertion was made in [48, 15] for modularity and the RB Potts model. Moderate levels of "confusion" caused by random or competing effects within a graph do not destroy information contained in the global energy landscape, and the replica

correlations of our algorithm are a measure of the "complexity" of that landscape. As the noise in the system is increased we expect that the transition to incoherence (where replicas are weakly correlated) to occur rapidly (see end of Sec. 3.4 and two brief examples of accuracy transitions in Sec. 4.3. If an algorithm can verifiably solve for the global minima of a system in most cases, the problem of community detection is solved in principle. Since this is difficult to do in practice, the replica correlations in our algorithm take advantage of the fact that we cannot always locate the optimal ground state(s).

In principle, one can also include in Eq. (2.2) interactions between each of the r replicas to produce a "free energy" type functional of the form

$$F = \sum_i \mathcal{H}_i(\{\sigma\}) - T \sum_{i \neq j} S(i,j). \tag{3.1}$$

where $S(i,j)$ is an information-based measure (e.g., I_N, V, etc.) between all replica pairs and T is a scale for this information measure. $S(i,j)$ is maximized when the community partitions are identical in all replicas. This information theory measure formally plays a role analogous to entropy in a free energy functional. T then plays the role of a "temperature." Sans the first term, the minima of F in Eq. (3.1) produce highly correlated random configurations (a "random high temperature configuration" of the system which appears without change in all replicas). Our algorithm in this work will amount to initially minimizing the first term in F, *i.e.*, $\sum_i \mathcal{H}_i(\{\sigma\})$, for a set of fixed $\{\gamma_i\}$. Out of this set of replica configurations, we then ask for which γ_i do we find a maximum of the correlations, $\sum_{i \neq j} S(i,j)$, when this information theory

measure is plotted as a function of γ. A more sophisticated version of our algorithm minimizes F directly with both terms included in each step as depicted in Fig. 3.1. The information theory measures that we employ may also be written for other (non-graph theoretic) optimization problems with general Hamiltonians, or cost functions, $\mathcal{H}$ (see Appendix E).

3.1.2 Algorithm

We start the algorithm with a weighted or unweighted graph. In Eq. (2.3), p_{in} is the *minimum* internal edge density for each community, and it is equivalent to the *resolution* of the system when we minimize Eq. (2.2). The algorithm uses Eq. (2.2) to solve a range of resolutions $\{p_i\} = [p_0, p_f]$ (decrementing p_i) corresponding to a particular set of model weights $\{\gamma_i\} = [\gamma_0, \gamma_f]$ as determined by Eq. (2.3). It is almost always sufficient to have $\gamma_0 \lesssim 19$ since it corresponds to a *minimum* community edge density of $p_0 \geq 0.95$. The final weight γ_f is found when the system is completely reduced. A completely reduced system is one that is fully collapsed into one community or one where disjoint sub-graphs will not allow the system to collapse any further.

Each iteration, we decrement the density p_i by a small value $\Delta p = 0.05$ (or 0.025 for smaller graphs) and calculate the corresponding γ_i. After a threshold value (say $p_t = 0.1$), we scale p_i by a factor of 1/2 (or 3/4 for smaller graphs) in order to take sizable steps towards a fully reduced system (necessary for large systems). One could readily implement an adaptable step or "fill-in" process since the order of trials is irrelevant for the result.

The algorithm takes three input parameters: the number of independent replicas r that will be solved at each tested resolution, the number of trials per replica t, and the starting density which we set to be $p_0 \simeq 0.95$ corresponding to $\gamma_0 = 19$. The number of replicas is typically $8 \leq r \leq 12$ and is selected based upon how much averaging (over all replica pairs) is needed or desired. The number of trials t per replica is generally $2 \leq t \leq 20$. For each replica, we select the lowest energy solution among the t trials as was discussed in Sec. 2.2. The value of t is chosen based on how much optimization is necessary to identify a strong low-energy configuration [73].

The r replicas (and t optimization trials) are generated by reordering the "symmetric" initialized state of one node per community. That is, even though the initialized state is symmetric, the order that we traverse the list also affects the answer that we obtain. This occurs because the node-level dynamics of the underlying community detection algorithm in Sec. 2.2 moves a node immediately upon identifying the best community membership given the current state of the system. Utilizing the r replicas, we then use the information-based measures of Sec. A to determine the multiresolution structure. Our algorithm is given by the following steps:

(1) *Initialize the system.* Initialize adjacency matrices (A_{ij} and J_{ij}) and weights (a_{ij} and b_{ij}) based on the system definition. Use Eq. (2.3) and p_0 to calculate the initial model weight γ_0.

(2) *Solve all replicas at this resolution p_i.* Initialize the current replica to a symmetric state of one node per community. Use Eq. (2.2) to solve each replica with model weight γ_i at a cost of $O(N^{1+\beta} Z^{1+\beta} t \log Z)$ per replica [73, 74]. Repeat the

process independently for all r replicas. Each trial and replica randomly permutes the order in which nodes are initially traversed in the respective solutions.

(3) *Calculate the replica I_N, V, I, and H information measures.* Use Eq. (A-1) to calculate H for all replicas and Eqs. (A-2) – (A-4) to calculate I, I_N, and V between all pairs of replicas for this resolution p_i [75]. Calculate the average (see Eqs. (A-5) and (A-6)) and the standard deviation for each measure.

(4) *Decrement to the next resolution p_{i+1}.* If $p_i > 0.1$, decrement $p_{i+1} = p_i - 0.05$ or 0.025 for smaller graphs. If $p_i \leq 0.1$, $p_{i+1} = p_i/2$ or $3p_i/4$ for smaller graphs. Calculate the model weight γ_{i+1} by Eq. (2.3). Return to step (2) until the system is not further reducible (fully collapsed or disjoint sub-graphs will not collapse).

(5) *Evaluate results.* For the range of model weights $\{\gamma_i\}$, plot each average $I_{N,i}$, V_i, I_i, and H_i versus γ_i. Determine the strongest correlations (I_N high or V low) in these plots (see Figs. 3.3 – 3.5, 3.7, 3.9, and 3.11). These strongly correlated regions correspond to the best multiresolution structure(s) in the graph. If the correlation is less than "perfect" ($I_N < 1$ and $V > 0$), we choose the lowest energy replica to be the partition solution. One could also choose to construct a "consensus" partition between all of the replicas [23, 76] at each notable resolution.

We estimate that the number of resolutions $\{p_i\}$ required to adequately specify an arbitrary system scales as $O(\log N)$. The dominant scaling of the algorithm is almost always step (2), so we estimate that the overall scaling is $O(N^{1+\beta} Z^{1+\beta} rt \log N \log Z)$ for some small β [74, 77].

Structures identified by this algorithm are not necessarily hierarchical; however,

one can augment the algorithm by imposing an additional hierarchical constraint on some fraction of the replicas. Comparisons would then be made strictly between all pairs with and without this additional constraint. We applied this variation in both divisive and agglomerative approaches, but in our testing it only resulted in a modest improvement to the algorithm's ability to identify the best resolutions. Therefore, we use the above algorithm in order to take advantage of its generality and relative simplicity.

3.2 Examples

3.2.1 Three-level hierarchy

The system in Fig. 3.2(a) depicts a set of 256 nodes for a constructed three-level heterogeneously-sized hierarchy. The results are seen in Fig. 3.3. The unweighted edge connection probabilities are p_k for $k = 1, 2, 3$. Level 3 has a density $p_3 = 0.9$ between nodes in the *same* community with community sizes from 5 to 22 (average 16) nodes. Level 2 has a density $p_2 = 0.3$ between nodes in *different* constituent sub-communities and is divided into five groups with merged sizes from 33 to 76 nodes. Level 1 is the completely merged system that has a density $p_1 = 0.1$ between nodes in *different* sub-communities. These edges provide some system noise. The average densities of communities at levels 1 and 2 are $p = \overline{p}_1 = 0.182$ and $\overline{p}_2 = 0.470$. We use eight replicas and four trials per replica at a total run time of 6.1 s [78].

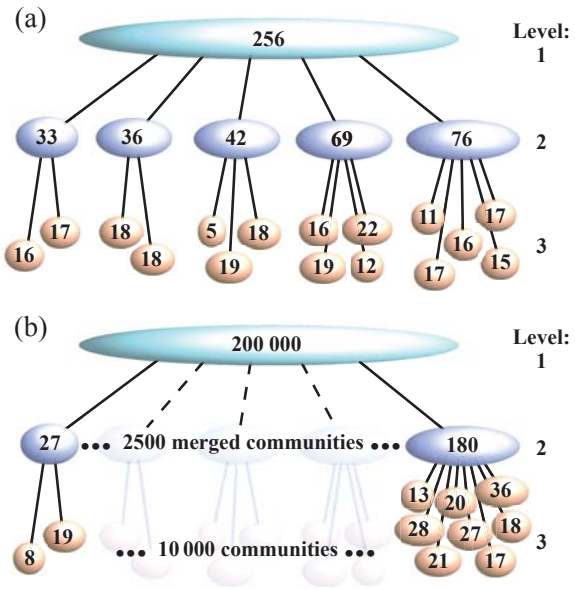


Figure 3.2: Heterogeneous hierarchical systems corresponding to the plots in Fig. 3.3 for panel (a) and the plots in Fig. 3.5 for panel (b). In panel (a), the 256 node system is divided into a three-level hierarchy where the unweighted edge connection probabilities at each level are the following: level 3 has $p_3 = 0.9$ between nodes in the *same* community with community sizes from 5 to 22 nodes (average 16). Level 2 has $p_2 = 0.3$ between nodes in *different* constituent sub-communities with merged community sizes from 33 to 76 nodes. Level 1 is the completely merged system of 256 nodes with $p_1 = 0.1$ between nodes in *different* sub-communities. The average edge density is $p = \overline{p}_1 = 0.182$. In panel (b), we increase the system size to 200 000 nodes. Level 3 has 10 000 communities with sizes from 6 to 37 nodes (average 20). Level 2 has 2500 communities with sizes from 27 to 180 nodes which are formed by merging two to eight communities from level 3. The density p_1 is changed from panel (a) to $p_1 = 0.000\,31$, and the average edge density is $p = \overline{p}_1 \simeq 0.0005$. This larger system has over ten million edges with approximately 62% of the edges being random noise between level 2 communities.

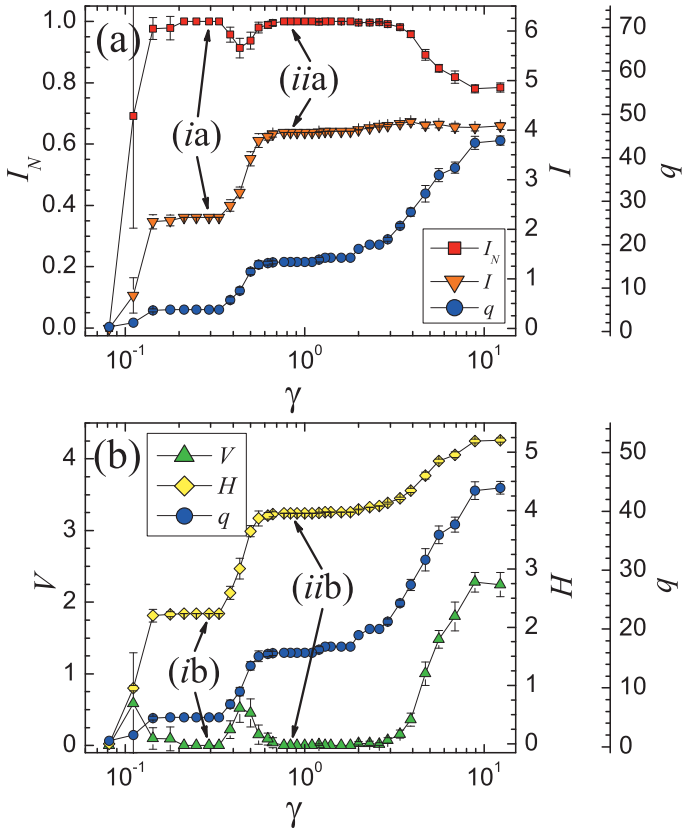


Figure 3.3: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs. the Potts model weight γ in Eq. (2.2) for the three-level heterogeneous hierarchy depicted in Fig. 3.2(a). In panel (a), the squares represent the average replica NMI I_N (left axis), and the inverted triangles represent the average mutual information I (right axis). In panel (b), the triangles represent the average VI V (left axis), and the diamonds represent the average Shannon entropy H (right axis). Circles in both panels represent the average number of clusters q (right-offset axes). In each panel, the peak I_N values at (ia) and (iia) and the corresponding minimum V values at (ib) and (iib) accurately correspond to levels 2 and 3, respectively, of the hierarchy depicted in Fig. 3.2(a). In panels (a) and (b), both the mutual information I and Shannon entropy H display a "plateau" behavior corresponding to the correct solutions. Plateaus in q [79] also indicate important structures as in [33].

We show the results of the multiresolution algorithm of Sec. 3.1 applied to several test cases [64]. In Secs. 3.2.1 and 3.2.3, we illustrate a small 256 node and a larger 200 000 node hierarchy respectively with both systems depicted in Fig. 3.2. In Sec. 3.2.2, we examine the structure of an Erdős-Rényi random graph for comparison to graphs with known internal structure. We then analyze two real social networks in Secs. 3.2.4 and 3.2.5 where the respective systems are depicted in Figs. 3.6 and 3.8. In Sec. 3.3, we also demonstrate the algorithm's exceptional accuracy for large systems.

In Fig. 3.3(a), the squares represent NMI averages over all replica pairs (left axis). The inverted triangles represent the mutual information I averages for the same replica pairs (right axis). In Fig. 3.3(b), the triangles represent VI averages over all replica pairs (left axis), and the diamonds represent the Shannon entropy H averages for the replicas (right axis). In both panels, the circles represent the average number of clusters across the replicas (right offset axes). All parameters are plotted versus the model weight γ where we use a logarithmic scale to facilitate comparing the behavior of a large range of system sizes from $N = 16$ nodes in Figs. 3.8 and 3.9 to as large as $N = 200\,000$ nodes in Figs. 3.2(b) and 3.5 [65].

The extrema (ia,b) and (iia,b) are the correctly determined levels 2 and 3 respectively of the test hierarchy depicted in Fig. 3.2(a). Peaks (ia) and (iia) have $I_N = 1$ and minima (ib) and (iib) have $V = 0$ which indicate perfect correlations among the replicas for both levels of the hierarchy. The "plateaus" in H and I are a second indication of the significant system structure whose importance will become more apparent in later examples. The plateau in the average q [79] is also an impor-

tant indicator of system structure as used in [33]. However, Figs. 3.4, 3.7, and 3.9 discussed later demonstrate that some caution should be exercised when using the plateau criterion (in H, I, or q) for determining multiresolution structure.

At level 3 in Fig. 3.2(a), the average number of externally connected edges for each node is $Z_{out} \simeq 32.0$ with a random noise component of $Z_{out}^{noise} \simeq 19.8$. Both of these numbers are larger than the average number of internal edges, $Z_{in} \simeq 14.3$. Despite this imbalance, the algorithm easily identifies level 3 of the hierarchy because the external edges (particularly those due to the random noise) are not concentrated strongly enough into any one external cluster. This behavior is important for smaller communities on level 3 where Z_{out} is substantially larger than Z_{in}, and it illustrates that the model is robust to noise in the system.

The VI peaks at $\gamma_1 = 0.111$ and $\gamma_2 = 0.435$ in Fig. 3.3(b) correspond to the average inter-community edge densities, $p_1 = 0.1$ for sub-communities at level 2 and $p_2 = 0.3$ for sub-communities at level 3. Equation (2.3) relates the *minimum* internal edge density $p_{in} \geq \gamma/(\gamma+1)$ for each community in a solved partition. We can arrive at this inequality, using inductive reasoning, by considering the minimum inter-community edge density required for two arbitrary communities A and B to merge. We apply the relation as an equality (*i.e.*, energy difference between the merged and unmerged states is approximately zero) for the peak VI values at γ_1 and γ_2. The respective densities are $p_1^{AB} = 0.100$ and $p_2^{AB} = 0.303$. These values correspond closely to the constructed inter-community densities p_1 and p_2 above. The local VI maxima show that "complexity" of the energy landscape increases at resolutions where $\gamma/(\gamma+1)$ is

equal to the mean inter-community edge density. The more intuitive interpretation is that the "complexity" of the energy landscape increases substantially when the energy difference between different states is approximately zero.

3.2.2 Erdős-Rényi random graph

In Fig. 3.4, for comparison purposes we show the results for a purely (Erdős-Rényi) random graph at the same average edge density $p = 0.182$ as the hierarchy in Figs. 3.2(a) and 3.3. We use eight replicas and four trials per replica at a total run time of about 6.9 sec [78]. The only peak (ia) in the random graph corresponds to a trivial division into groups with sizes of approximately $\{1, 2, 253\}$ among the various replica solutions. This peak indicates transitional behavior to lower density, essentially trivial, structures. Peaks such as (i) can be distinguished from more meaningful ones by the cluster size distribution or the corresponding information measures. The value of I at (ia) or V and H at (ib) all have very low information values. Otherwise, the random graph displays no significant multiresolution structure.

All of the information measures display a plateau behavior at (iia,b). The plateaus in NMI or VI do not indicate a clear multiresolution structure because the correlations are relatively poor ($I_N \simeq 0.70$ and $V \simeq 3.6$) for both measures. If we examine the detailed solutions across the plateaus (separate from our multiresolution algorithm), the average NMI and VI are $I_N = 0.644$ and $V = 4.04$ both of which indicate poor agreement. There is no consistent structure identified by the community detection algorithm in this region. Instead, the weak plateaus in NMI and VI indicate that

the system is constrained within a set of similarly sized partitions that have similarly high community edge densities. This example also illustrates that if we use *only* the plateaus (in H, I, or q), there is a potential to incorrectly identify significant structure(s) in the system. This possibility can be remedied by information checks on nearby solutions in the plateau, but the poor NMI and VI correlations already appear to indicate the lack of consistent structure in the region.

3.2.3 Large hierarchy

A much larger hierarchy is depicted in Fig. 3.2(b). The system has 200 000 nodes and 10 011 428 edges. Approximately 62% of these edges are due to random noise between level 2 communities. For this system, $p_1 = 0.000\,31$, but $p_2 = 0.3$ and $p_3 = 0.9$ are unchanged from Fig. 3.2(a). There are 10 000 sub-communities at level 3 with sizes ranging from 6 to 37. Level 3 communities are combined in groups of two to eight to form the 2500 communities of level 2 with sizes ranging from 27 to 180. We use eight replicas and two trials per replica with a run time of about 4.6 hours [78]. In Fig. 3.5, extrema (*i*a,b) exactly identify level 2 of the hierarchy with perfect NMI and VI correlations, and extrema (*ii*a,b) accurately identify ($I_N = 0.999\,995$ and $V = 1.42 \times 10^{-4}$) all but 5 merged clusters out of 10 000 and 15 nodes out of 200 000 nodes for level 3. Due to random fluctuations, all of these nodes have a random connectedness of 50% or less for their intended communities. This result is therefore consistent with the model and algorithm.

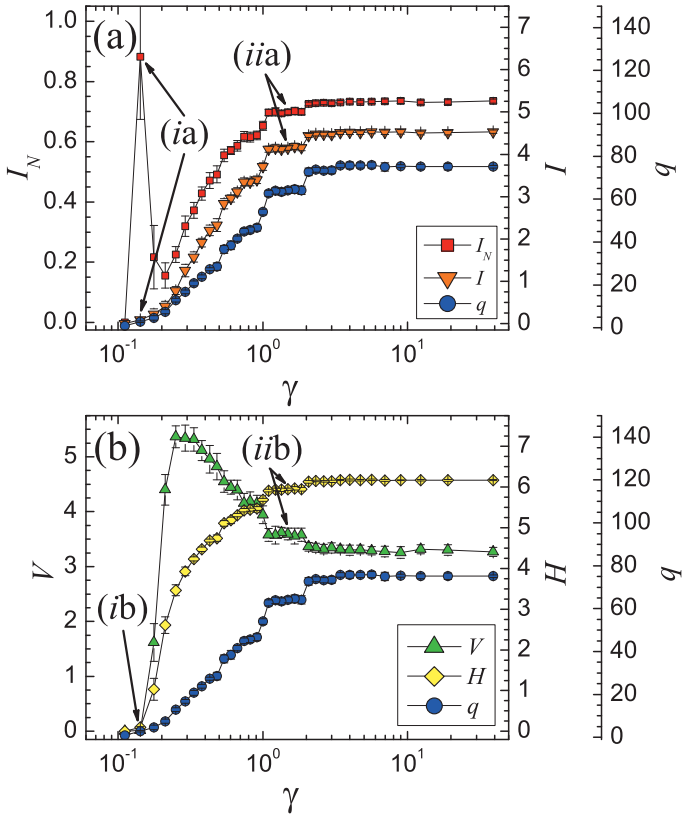


Figure 3.4: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs the Potts model weight γ for a (Erdős-Rényi) random graph that has the same average density $p = 0.182$ as the hierarchy in Fig. 3.2(a) and the corresponding results in Fig. 3.3. The right-offset axes plot the number of clusters q. See Fig. 3.3 for a full description of the legends and axes. In panel (a), the peak (*i*a) corresponds to a trivial partition of the system into groups with sizes of approximately $\{1, 2, 253\}$ among the different replicas. The trivial structure change in the NMI spike is indicated by its the low value of mutual information I at (*i*a) and by its low VI V and Shannon entropy H at (*i*b). The plateaus at (*ii*a,b) do not correspond to a consistent multiresolution structure as evidenced by the poor NMI and VI correlations. Rather, they indicate multiple similarly sized configurations that have similar community edge densities.

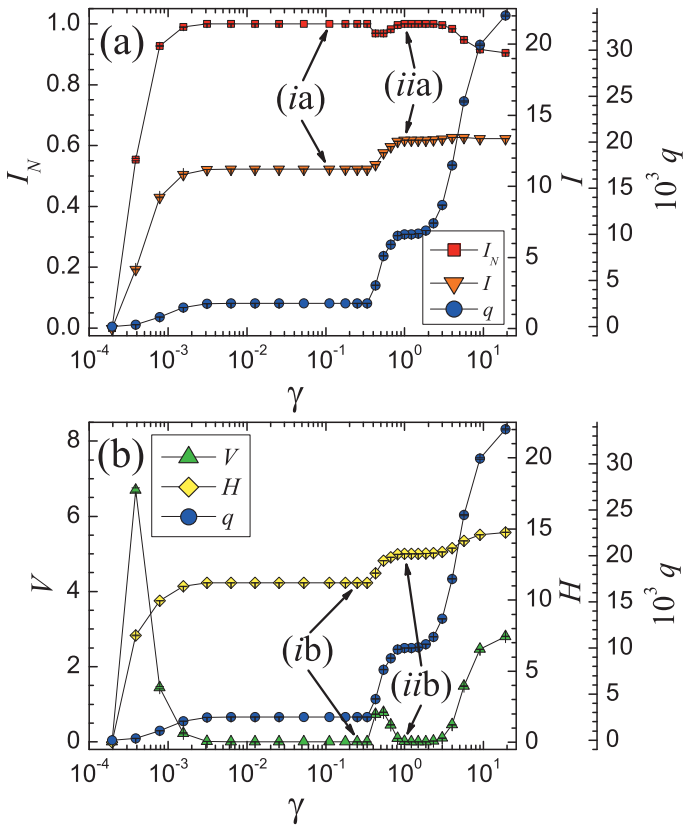


Figure 3.5: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs. the Potts model weight γ for the large three-level heterogeneous hierarchy depicted in Fig. 3.2(b). The right-offset axes plot the number of clusters q. See Fig. 3.3 for a complete description of the legends and axes. With the exception of 15 weakly connected nodes (out of 200 000) and 5 merged clusters (out of 10 000) at (*ii*a,b), the extremal values of I_N and V at (*i*a,b) and (*ii*a,b) both accurately correspond to levels 2 and 3 respectively of the hierarchy depicted in Fig. 3.2(b).

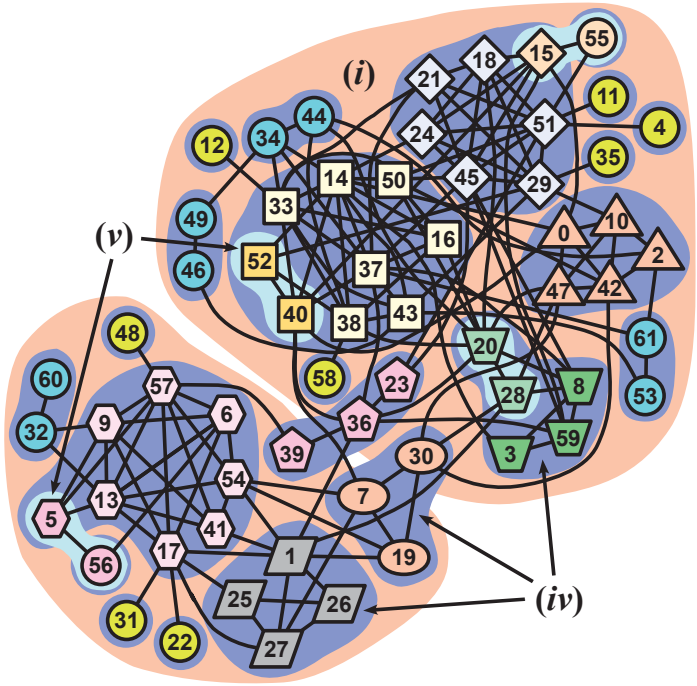


Figure 3.6: Pictorial representation of a social network of 62 bottlenose dolphins in Doubtful Sound, New Zealand[80, 81, 82]. These groupings correspond to structures (i), (iv), and (v) in Fig. 3.7 in order of smaller group sizes. The two-cluster partition (i) corresponds to a known split of the dolphin community [80]. In partition (iv), sub-groups are assigned distinct node shapes except for circles which indicate various one and two member groups. Structure (v) is identified from configuration (iv) when the four highlighted dyads of dolphins ($\{5, 56\}$, $\{15, 55\}$, $\{20, 28\}$, and $\{40, 52\}$) form distinct sub-groups. Note that sub-groups $\{7, 19, 30\}$ and $\{23, 36, 39\}$ in (iv) have nodes that are separated in their respective super-groups. These groups are examples of how our algorithm does not restrict node assignments between different resolutions, and they illustrate how the algorithm can apply to general types of multiresolution structure.

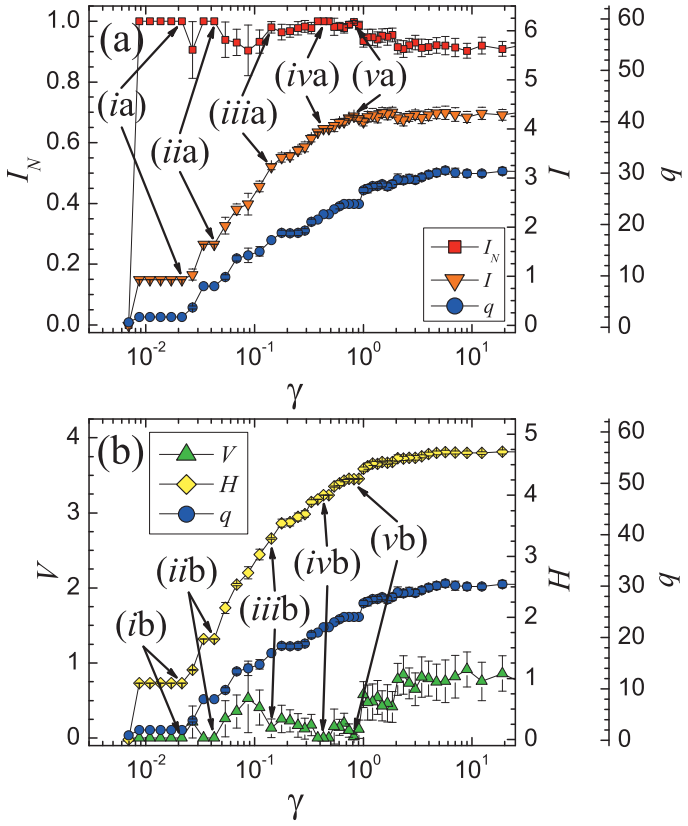


Figure 3.7: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs. the Potts model weight γ for a social network of 62 bottlenose dolphins in Doubtful Sound, New Zealand [81, 80, 82]. A summary of results is depicted in Fig. 3.6 for configurations (i), (iv), and (v). The right-offset axes plot the number of clusters q. See Fig. 3.3 for a complete description of the legends and axes. One notable grouping is configuration (i) which corresponds to a known split of the dolphin community [80]. The structures represented by (ii) – (v) are other potential well-defined partitions explained in the text.

3.2.4 Dolphin social network

We tested a social network of 62 bottlenose dolphins in Doubtful Sound, New Zealand [80, 81, 82]. Three of the strongest partitions ((i), (iv), and (v)) are depicted in Fig.

3.6 using the results in Fig. 3.7. We use ten replicas with ten trials per replica at a total run time of about 0.78 sec [78]. We use a density scaling of 0.8 rather than 0.75 for $p_i < 0.1$ for step (4) of the algorithm in order to more easily observe the transition between structures (i) and (ii) in Fig. 3.7. Configuration (i) identifies a grouping of 21 and 41 dolphins with perfect NMI and VI correlations ($I_N = 1$ and $V = 0$). This configuration agrees with an observed split of the dolphin network when a dolphin left the school [80], but our algorithm also suggests that this configuration is not the only strongly defined partition for the system.

Our algorithm further identifies partitions (ii) – (v) as important candidate partitions based on the strong NMI and VI information correlations. Partition (ii) separates weakly connected dolphins ($\{4\}$, $\{11\}$, $\{12\}$, $\{35\}$, $\{58\}$, and $\{46, 59\}$) in the larger super-group of Fig. 3.6 into distinct sub-groups. Configuration (iii) is slightly less well-defined with information correlations of $I_N \simeq 0.980$ and $V \simeq 0.132$. It separates weakly connected dolphins ($\{22\}$, $\{31\}$, $\{39\}$, $\{48\}$, and $\{32, 60\}$) of the smaller super-group of partition (i) and also begins a coarse division of the larger super-group. Configuration (iv) is perfectly correlated and is the first major reconfiguration of both super-groups of structure (i). The data in the three largest groups of (iv) are largely divided along gender lines according to details presented in [81]. Configuration (v) is a slight variation of (iv) with $I_N \simeq 0.998$ and $V \simeq 0.0178$ which separates four dyads of dolphins ($\{15, 55\}$, $\{46, 49\}$, $\{32, 60\}$, and $\{20, 28\}$) into distinct groups. Among different tests, there is some variation in the predicted groupings where a few nodes can be reassigned between groups or separated into distinct communities. Sub-groups

$\{7, 19, 30\}$ and $\{23, 36, 39\}$ of configuration (iv) have nodes that are split between the two super-groups of (i). These groups show that our algorithm does not restrict node assignments between different resolutions. This behavior allows our algorithm to solve general types of multiresolution structures.

All measures show a strong plateau for configuration (ia,b). The mutual information I shows weak plateaus at (iia) and (iva) but no plateau at (iiia) and (va). Similarly, the Shannon entropy H shows weak plateaus at (iib) and (vb) but no plateau for (iiib) and (ivb). The average number of clusters q as used in [33] also indicates the presence of structures (ii) and (v), but it misses partition (iv). Additionally, a weak plateau in q near configuration (iii) predicts a slightly different resolution than the extremal NMI and VI correlations. The weak plateau behavior of H, I, or q at different configurations of (iia,b) – (va,b) do not contradict the existence of valid structures. Rather, missing plateaus in the supplemental measures H, I, or q can indicate a noisy graph in general or a strongly defined but transient resolution.

3.2.5 Highland Polopa tribe relations

Figures 3.8 and 3.9 show the results for 16 Polopa tribes of Highland New Guinea [83, 84]. These data feature allied, neutral, and antagonistic relations between the subtribes of the region. Hage and Harary [84] used symmetric edge weights of $+1$ for allied relations, 0 for neutral relations, and -1 for antagonistic relations in their analysis; but these "intuitive" weight assignments are inconsistent if extended to systems that include few or no antagonistic relations (such systems would tend to "collapse" into

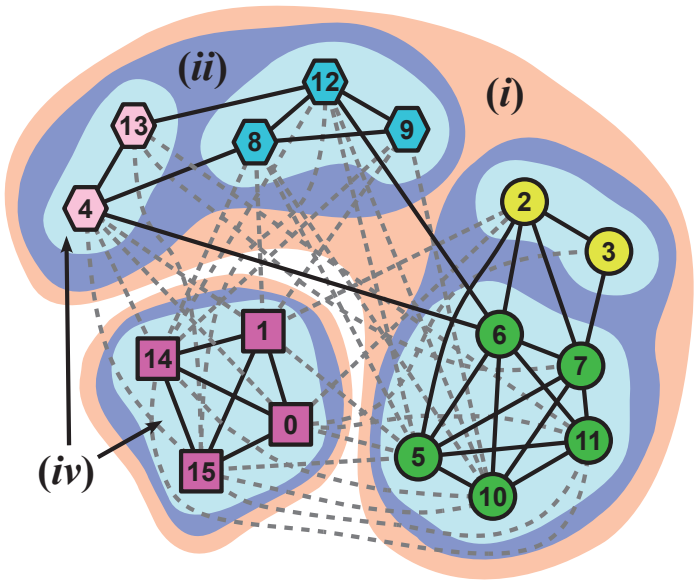


Figure 3.8: Pictorial representation of 16 Polopa tribes of Highland New Guinea [83, 84]. Solid lines represent allied relationships, and gray dashed lines represent antagonistic relationships. The three main levels of the structure are indicated by shaded areas. These groupings of tribes correspond to structures (i), (ii), and (iv) in Fig. 3.9 in order of smaller group sizes. Distinct node shapes (intermediate grouping) also correspond to structure (ii). The three-cluster structure (ii) corresponds exactly to the analysis in [83, 84]. Structure (iii) in Fig. 3.9 is formed when node 2 joins the group at the bottom-right of the figure.

large groups). Therefore, our model uses the more consistent assignments of -1 for "neutral" relations and -2 for antagonistic relations. Interestingly, Hage and Harary [84] related the fact that the sub-tribes did not consider the possibility of strictly neutral relations among tribes. We use 12 replicas with 10 trials per replica to limit fluctuations in this very small data set at a total run time of about 0.46 sec [78]. We use an array data structure due to the missing edge weights.

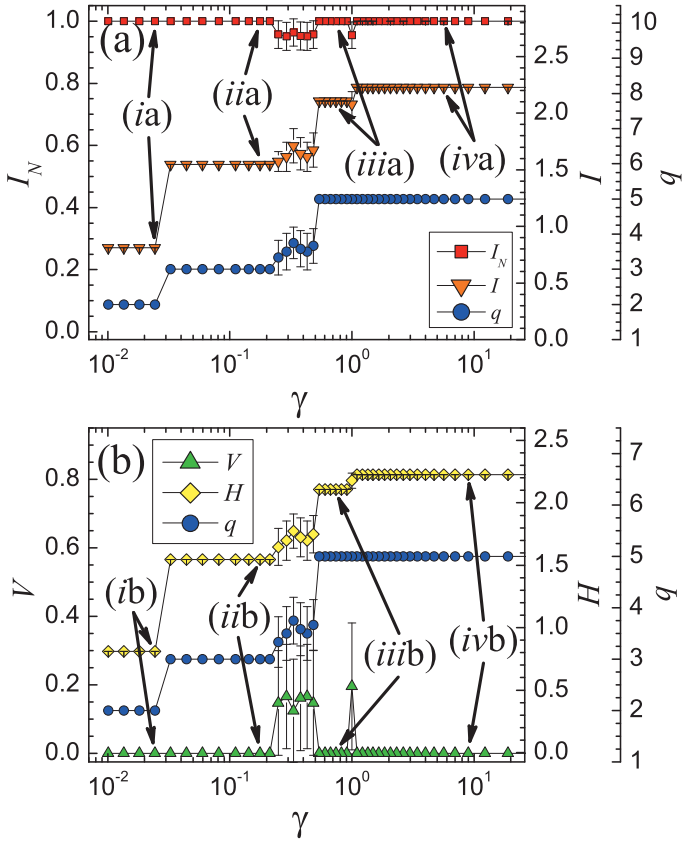


Figure 3.9: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs. the Potts model weight γ for 16 Polopa tribes of Highland New Guinea. The results are summarized in Fig. 3.8. The right-offset axes plot the number of clusters q. See Fig. 3.3 for a complete description of the legends and axes. The most important structure represented in the figure is at (iia,b) where the strong correlations agree exactly with analysis presented in [83, 84]. See the text for a full discussion of the other structures indicated in the figure.

Figure 3.8 depicts configurations (i), (ii), and (iv) from Fig. 3.9 in order of smaller group sizes. For presentation purposes, we allow three additional resolutions to be solved after the algorithm detects disjoint subgraphs at (ia,b). Our three-cluster partition (ii) agrees exactly with those discussed in [84]. All configurations indicated

in Fig. 3.9 are strongly defined with $I_N = 1$ and $V = 0$. The first configuration (i) is a two-cluster solution which merges two sets of clusters of configuration (ii). The small size of the system causes the transition between configurations (i) and (ii) to be sharply defined. To resolve the ambiguity, we must reference the plateaus in the information measures H or I (or the number of clusters q [33]).

Strong NMI and VI values at $(iii$a,b) and $(iv$a,b) correspond to two five-cluster solutions. These solutions sub-divide the three-cluster system into two slightly different dense configurations of allied tribes. In configuration (iii), node 2 is associated with the group on the bottom-right of Fig. 3.8. In configuration (iv), all groups are cliques (maximally connected sub-graphs). Both NMI and VI detect the transition between (iii) and (iv) with a short-lived spike. The information measures H and I also show the transition with plateaus at different values. Here, the number of clusters q does not detect the transition since q does not actually change. Again, this is due to the limited variability in this system, but the same ambiguity occurs in Fig. 3.4 for all three supplemental measures H, I, and q.

3.3 Accuracy

In Figs. 3.10 – 3.12, we test the accuracy of the multiresolution algorithm of Sec. 3.1 with a recently proposed benchmark in [67]. An example graph with $N = 1000$ nodes is depicted in Fig. 3.10. This new benchmark can pose a significant challenge since it incorporates a more realistic heterogeneous distribution of community sizes and

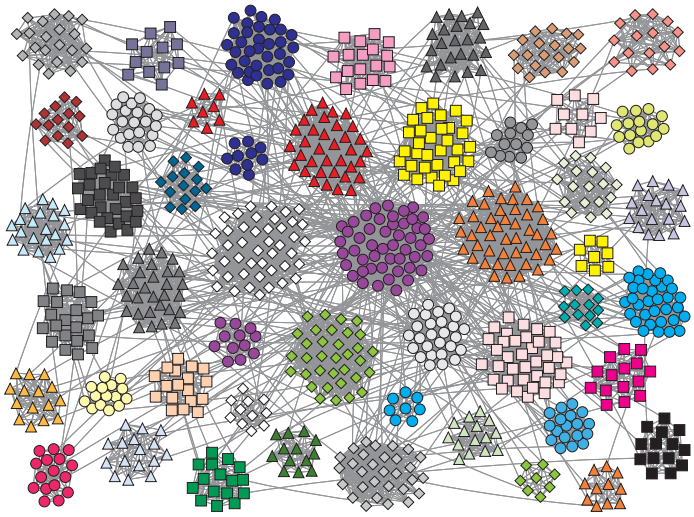


Figure 3.10: A sample graph with $N = 1000$ nodes from the new benchmark proposed in [67]. For presentation purposes, this depiction uses $\mu = 0.05$. Other parameters are $\alpha = 2$, $\beta = 1$, $\langle k \rangle = 15$, and $k_{max} = 50$ (see text).

node degrees, and it allows for testing across a large range of system sizes. It divides a set of N nodes into q communities with sizes assigned according to a power-law distribution with an exponent β. The community sizes are optionally constrained by minimum and maximum sizes of n_{min} and n_{max}. The degrees of the nodes are also assigned in a power-law distribution with an exponent α with constraints specified by the maximum degree k_{max} and the mean degree $\langle k \rangle$. The minimum degree k_{min} is set so that the distribution gives the correct mean $\langle k \rangle$. A fraction $(1 - \mu)$ of the edges of each node are connected to nodes within their own communities. The remaining fraction μ are assigned to nodes in other communities.

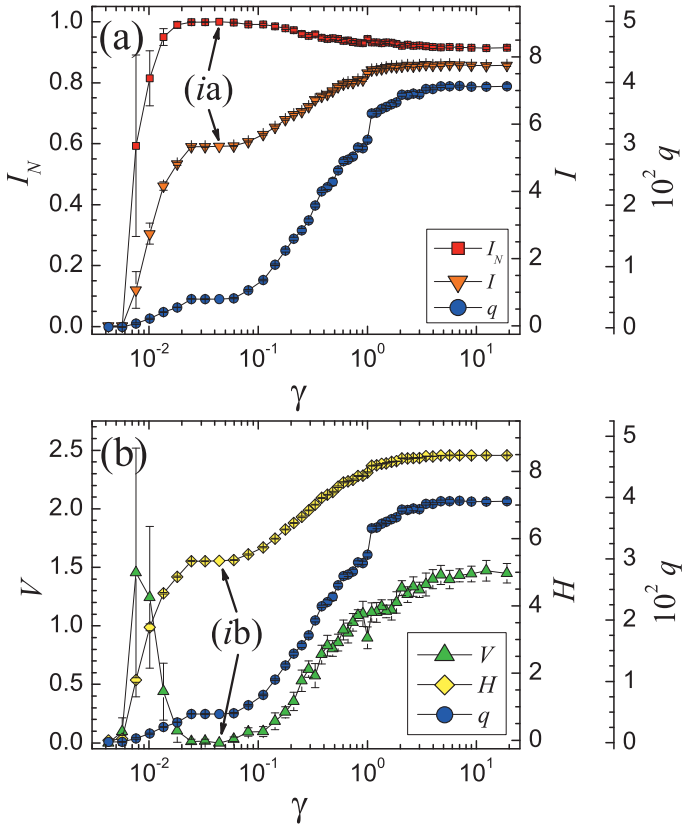


Figure 3.11: Plot of information measures I_N, V, H, and I in panels (a) and (b) vs. the Potts model weight γ for a single realization of the benchmark suggested in [67]. The right-offset axes plot the number of clusters q. See Fig. 3.3 for a complete description of the legends and axes. Figure 3.10 depicts a sample system from the benchmark. This example plot is for $N = 1000$, $\mu = 0.5$, $\alpha = 2$, and $\beta = 1$ (see text). Using the algorithm in Sec. 3.1, we identify the strongest NMI and VI replica correlations among the different resolutions as the "best" answer for the graph. For this graph, there is only one extremal value of I_N and V which indicates that there is only one "best" resolution for the defined system (see also Appendix F). Note that these information values are the averages *among the replicas*. The full accuracy plot in Fig. 3.12 plots the average I_N between the "*best*" partitions and the *known* benchmark graphs for a range of the mixing parameter μ.

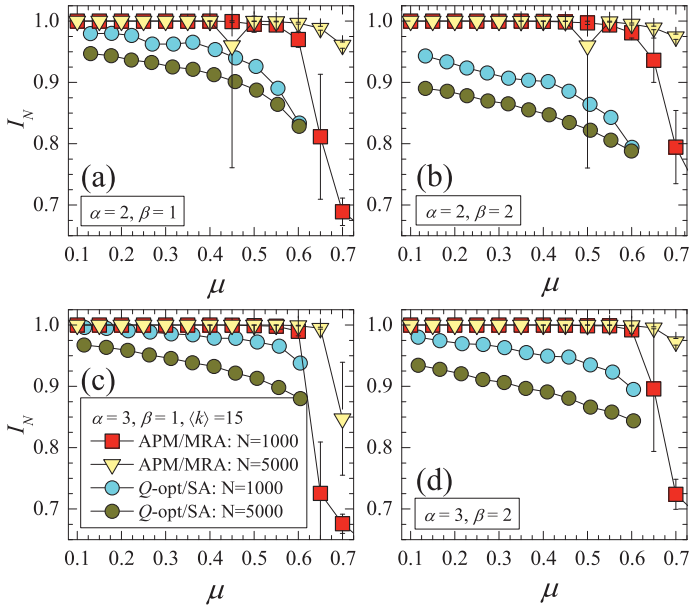


Figure 3.12: A plot of I_N vs. μ for a new benchmark problem proposed in [67]. I_N is calculated between the solved answer, by means the multiresolution algorithm in Sec. 3.1 using the APM of Eq. (2.2), and the constructed benchmark graphs. An example multiresolution analysis for a sample graph is in Fig. 3.11. μ is the fraction of edges of each node (on average) that are assigned outside its own community. We tested the power-law distribution exponents $\alpha = 2$ and 3 and $\beta = 1$ and 2 for the node degrees and the community sizes, respectively. For comparison, we also plot the results from [67] determined by modularity optimization (Q-opt) using SA. With the APM, our multiresolution algorithm demonstrates extremely high accuracy for large systems (see text). Appendix F discusses the accuracy perturbations in panels (a) and (b) for $N = 5000$ nodes. Data for $N = 1000$ and $N = 5000$ nodes are averaged over 100 and 25 graphs, respectively.

We test systems with $N = 1000$ and 5000 nodes and power-law exponents of $\alpha = 2$ and 3 for the degree distribution and $\beta = 1$ and 2 for the community size distribution.

We do not specify the optional community size constraints n_{min} or n_{max} allowing the benchmark program to specify them by the degree distribution. The node degree distribution is specified by $\langle k \rangle = 15$ and $k_{max} = 50$ where the mean degree $\langle k \rangle = 15$ was the most difficult of the tested values in [67]. We vary the mixing parameter μ in the range $0.1 \leq \mu \leq 0.7$. The accuracy results are summarized in Fig. 3.12.

We apply the multiresolution algorithm of Sec. 3.1 to identify the "best" system partition. Figure 3.11 shows an application of the algorithm for a single benchmark graph with $N = 1000$, $\mu = 0.5$, $\alpha = 2$, and $\beta = 1$. In this plot, we identify the "best" system resolution by the strongest average NMI correlation between all pairs of replicas. We use $r = 8$ replicas with $t = 4$ energy optimization trials per replica. As seen in Fig. 3.11, both I_N and V (almost always) show only one extremal value which is the strongly defined system at (ia,b). Plateaus in H, I, and q qualitatively confirm the structure indicated by the extrema in I_N and V. From these data, we determine that there is only one "best" resolution for the defined system. See Appendix F for additional considerations in identifying the "best" benchmark resolution.

In Fig. 3.12, we identify the "best" partition for a set of benchmark graphs over a range of the mixing parameter $0.1 \leq \mu \leq 0.7$. We then compare each solution via NMI with the "known" partition. We average over 100 graphs for $N = 1000$ and over 25 graphs for $N = 5000$ for each tested μ. For comparison, we also include the results given in [67] for modularity optimization using a simulated annealing algorithm. Combined with the APM of Eq. (2.2), our multiresolution algorithm performs excellently, achieving almost perfect accuracy for each tested distribution

exponent α and β and for a large range of the mixing parameter μ. The accuracy perturbations in panels (a) and (b) for $N = 5000$ nodes are due to benchmark graphs with more than one local extremum in I_N and V. These perturbations are a result of the automated selection of the single "best" resolution based on I_N and V extrema. We can largely eliminate them by a simple extension of the basic multiresolution algorithm (see Appendix F). They are also nearly eliminated for these values of N if we specify the default community size constraints of $n_{min} = 20$ and $n_{max} = 50$.

The absolute Potts model has little difficulty accurately solving the harder problem with $N = 5000$ nodes because the edges connected to external communities are spread over more communities on average. This construction causes a greater contrast of interior and external edge densities (considering edges connecting *pairs* of communities). This larger contrast allows the benchmark graph to be easily identified by the multiresolution algorithm. The converse occurs for small systems in the benchmark.

Our multiresolution algorithm has some difficulty in identifying all communities in this benchmark for exceptionally small systems ($N \lesssim 300$) where we achieve $I_N \simeq 1.0$ for a range of μ that increases with N (for $N = 300$, $I_N \simeq 1.0$ for $\mu \leq 0.45$). Communities are partitioned locally, independent of any *global* parameters of the system; so this limitation is not a resolution limit effect. Rather, this behavior is due to simultaneously resolving communities with substantially different relative densities [85]. Palla *et al.* [4] stated that the community density should be used in identifying communities, which our Potts model does in effect. In Sec. 2.1.1, we suggested that it

is the typical community edge density that characterizes the *resolution* of a partition. The difficulty in this benchmark is due to defining communities by the fraction of each node's edges $(1-\mu)$ that lie within its own community. Each community contains $\ell_s = n_s\langle k\rangle(1-\mu)/2$ edges on average where n_s is the size of community s. The average edge density p_s of community s is

$$p_s = \frac{\langle k\rangle(1-\mu)}{(n_s-1)}. \tag{3.2}$$

The numerator is constant on average across all communities. Our Potts model solves heterogeneously-sized systems well (see Secs. 3.2.1 and 3.2.3), but one notable implication of Eq. (3.2) is that the realistic distribution of community sizes leads to a substantial distribution of community edge densities with substantially different character for this benchmark.

Note also that our highly accurate results for $\mu = 0.6$ and 0.65 for most values of N, α, and β in Fig. 3.12 show that the concept of a weak community structure [69], where some nodes have more total edges connected to other communities than within their own, is not too restrictive because the external edges can be dispersed among many other communities. Indeed for $\mu > 0.5$, all clusters in this benchmark on average exceed the definition of a weak community since most, if not all, nodes have more exterior than internal edges. So-called weak communities can occur frequently in social networks for example. Individuals often know far more people than the size of the local "community" group(s) (friends, associates, etc.) of which they are members. We showed a similar, but more striking, result when identifying level 3 of

the constructed hierarchy in Figs. 3.2(a) and 3.3 where the smallest communities had many more external than internal edges. Nevertheless, the model could easily resolve the communities at the correct resolution.

3.4 Discussion

In Figs. 3.3 – 3.5, 3.7, 3.9, and 3.11, strong correlations in NMI and VI appear to be consistent indicators of important multiresolution structures. In most cases the assessments of the "best" partitions are confirmed by "plateaus" in the mutual information I and the Shannon entropy H. These information plateaus are similar to those seen in the number of clusters q in [33] and that are also observed in our data [79]. In Ref. [33], the Arenas *et al.* indicated that plateaus in q correspond to the most relevant system structures. Our results largely affirm but also extend that observation.

In many pertinent applications of our algorithm, the final results (including, by fiat, our synthetic networks in Secs. 3.2.1 and 3.2.3) are indeed hierarchical in the conventional sense. That is, solving the Hamiltonian of Eq. (2.2) anew with a different model weight γ may break the communities apart, but it does not swap vertices between different communities at the correct resolutions. As each resolution is solved independently in our algorithm, we may (and indeed do) find more complicated multiresolution partitions where node reassignments lead to overlaps between communities that are perhaps disjoint on another level. This latter case is more sub-

tle and appears in systems such as the dolphin social network of Sec. 3.2.4 and other individually oriented networks.

Variations in run time scaling among the different tests is influenced, sometimes strongly, by different levels of effective noise in each system (aside from differing numbers of replicas and trials; see Sec. 4.3.2). For example, the hierarchy for Fig. 3.3 had a run time of 6.1 s. The corresponding random graph in Fig. 3.4, with nearly the exact same density and number of nodes, finished in 6.9 sec.

NMI and VI possess different strengths for quantitatively assessing multiresolution structure. (**1**) Of course, NMI is normalized and VI is not (although one normalization for VI is $1/\log_2 N$ [86]). Both of these features are useful. (**2**) Figures 3.3–3.5 show that VI more clearly identifies poor configurations. In the high density regime ($\gamma \gtrsim 5$) of Figs. 3.3 and 3.5, NMI shows a lower correlation compared to the peak values at (*i*) and (*ii*); but VI clearly indicates poor agreement. In Fig. 3.4, VI in panel (b) visually indicates a much poorer correlation in the $\gamma \simeq 0.3$ region as compared to NMI in panel (a). (**3**) In Fig. 3.4(a), we identified peak (*i*a) as a "trivial" division with a huge component weakly connected to some small branch elements. If one was actually interested identifying these very low-density solutions, NMI does identify them. In panel (b), V and I simply indicate a very low-information configuration.

In many cases, extrema in either NMI or VI are sufficient to identify the multiresolution structure of a system. Occasionally, we need to additionally reference the mutual information I or the Shannon entropy H (or the number of clusters q [33]). For example, in Fig. 3.3 NMI and VI almost do not distinguish between the $\gamma = 0.83$

partition (the exactly correct one) and the $\gamma = 1.6$ partition (one weakly connected node separates to form a new community) because the separation between the two configurations is almost imperceptible. Both of these partitions correspond to level 3 of the hierarchy depicted in Fig. 3.2(a), and both partitions have perfect correlations ($I_N = 1$ and $V = 0$). In this case, the small changes in information measures H and I indicate a redundant $\gamma = 1.6$ partition. Also in Figs. 3.11 and 3.12, we used the plateau to distinguish, when needed, between strongly correlated transient partitions (due to random elements of the benchmark generation process) and the more stable partition corresponding to the intended solution.

A similar challenge can occur for very small systems, such as in the transition from (i) to (ii) in Fig. 3.9, or for systems with few intercommunity connections. As the resolution is adjusted in these systems, variability can be more limited; and system transitions can be sharply defined. For these systems, it is possible that the NMI and VI correlations can remain strong and constant while crossing a structural transition. In Fig. 3.9, we avoid this ambiguity by noting that H and I clearly show a transition between structures (i) and (ii). Such systems can also accentuate the perceived plateaus in the multiresolution data because the variation in different configurations is small and transitions between major configurations can be sharp.

Given the distinctions, the two evaluations of multiresolution structure ("plateau" behavior in H, I, and q or strongly defined I_N and V correlations) are complimentary. While the plateau behavior is important, it is a more qualitative assessment of the "best" resolutions for the system. At least for our Potts model, under some

conditions the plateaus in H, I, or q can be weak enough to prevent them being used as the universal indicator of multiresolution structure. In Fig. 3.4, the plateaus even corresponded to a set of similarly sized partitions with similar densities rather than consistent structure. The NMI and VI approach can more easily identify short-lived, but nevertheless strongly defined, structures (such as configuration (*iv*) in Fig. 3.7) that the plateau criterion can miss. In all Figs. 3.3 – 3.5, 3.7, 3.9, and 3.11, the major benefit of using the NMI and VI evaluations is that it appears to give a *quantitative* estimate of the "best" resolutions. Together, the information measures appear to provide a consistent, accurate, and quantitative method of identifying general multiresolution structure.

3.5 Further work

In further work, we will also consider a different method of adjusting the resolution of the system using the Hamiltonian

$$\mathcal{H}_{vt}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left[\,(a_{ij}+\alpha_{ij})\,A_{ij} - (b_{ij}+\beta_{ij})\,J_{ij}\right]\delta(\sigma_i,\sigma_j) \tag{3.3}$$

where α_{ij} and β_{ij} are the new model weights as compared to γ in Eq. (2.2). This *variable topology* Potts Hamiltonian is a generalized and continuous version of threshold cut-offs in weighted graphs. It presents an alternative method of continuously scaling the system by using an additive rather than a multiplicative scaling. It differs from Eq. (2.2) in that it progressively adjusts the topology of the system where multiplicative scaling does not change the system's connectedness. Additive scaling may

provide a different perspective on the evolution of the system structure over different scales, and it may better simulate how some real world models are "stressed."

Additionally, it may be possible to probe the system at a local level by using either localized partitions or by analyzing details within the confusion matrix at each resolution. With this approach, we may be able to identify stable, but localized, structures beyond the information conveyed in the global information-based correlations.

In a future work, we will detail the minimization of the "free energy" type functional of Eq. (3.1). This functional contains both the Potts model energy and the composite information function. This latter information theory measure is maximized when the correlation between replicas is maximal.

3.6 Conclusion

We use a Potts model measure for community detection and apply it to detecting multiresolution structures: (**1**) Our approach identifies and *quantitatively* evaluates the 'best' multiresolution structure(s), or lack thereof, in a graph. (**2**) All resolutions are solved independently, so the algorithm allows for the identification of completely general types of multiresolution structure. (**3**) It is based on information comparisons, so in principle is should apply to any community detection model that can examine different resolutions. (**4**) The underlying Potts model and algorithm are as accurate as the best methods currently available. [54, 59]). The model is a local measure of community structure, so it is free from the 'resolution limit' as discussed

in the literature [54, 44, 45, 33, 34, 52]. (**5**) Building on this foundation, the multiresolution algorithm demonstrates extremely high accuracy for large systems using a recent benchmark proposed in [67] (see Sec. 3.3). (**6**) We estimate that the computational cost scales as $O(N^{1+\beta}Z^{1+\beta}rt \log N \log Z)$ for some small β [74, 77] where r is the number of replicas, t is the number of optimization trials per replica, Z is the average node degree, and N is the number of nodes. We have tested our community detection algorithm on systems as large as $O(10^7)$ nodes and $O(10^9)$ edges (see Sec. 2.6.2) [78]. The multiresolution algorithm requires a substantial number of individual community solutions; but due to the speed of the underlying algorithm, it can nevertheless examine systems over $O(10^5)$ nodes and $O(10^7)$ edges on a single-user workstation. The algorithm should extend very efficiently to parallel or distributed computing methods allowing larger systems to be studied.

Chapter 4

Phase transitions in community detection

4.1 Introduction

Phase transition effects appear in many computational problems, and they constitute one of the most important applications of statistical methods to these problems. In these applications, a phase transition is defined as a situation where small changes in local behavior will significantly change the overall algorithm performance (accuracy and/or computational cost). Examples include the k-Satisfiability (k-SAT) problem [87, 88, 89], search problem in artificial intelligence [90], Steiner trees [91], random vertex-covers [92], Hamiltonian circuits [93], graph coloring [93], image restoration and error-correction [94], and others. We will elaborate on a phase transition effect in community detection that manifests in both static and dynamic aspects of the

problem. Our results constitute a new perspective for singular transition from a typical-easy to a rare-hard region in the community detection problem.

The static phase transition is studied via the system energy, "noise" (density of intercommunity edges), time and temperature curves, and also the time-correlation function. For a particular system, the energy as a function of noise p_{out} reaches a peak at some critical value of p_{out} which is a sign of crossing from an easy to a hard solution region much like that which occurs in the k-SAT problem [87]. We determine the critical value of the noise for this transition, and we then study the properties of system in the transition region. As it turns out, the system shows non-equilibrium phenomenon, i.e., a breakdown of ergodicity; and further studies show that the system in this region has spin-glass-like properties. For example, it has a memory effect.

The dynamic transition is analyzed by testing the fluctuations in the node trajectories (community memberships) as a function of time. The node trajectories change from convergent (well-defined memberships) to chaotic status (rapidly changing memberships) as more noise is added. The convergence behavior in low noise corresponds to system ergodicity, and the chaotic behavior in high noise corresponds to the breakdown of ergodicity. For a fixed system, the transition found by dynamic and static facets always correspond to the same transition point which indicates that the phase transition is an inherent property of the community detection problem.

4.2 Heat bath algorithm

Using the cavity method [95], we can analytically solve for the communities in certain graphs, e.g., when all nodes have a fixed degree of $k = 3$. However, most general graphs (with arbitrary degree and cluster size distributions) require computer simulation. The system energy is the most intrinsic parameter for our problem, and we search for the ground state of the above Hamiltonian by iteratively moving nodes between candidate communities based on a heat bath algorithm (HBA). Our community detection heat bath algorithm is as follows:

(1) *Initialize the system.* Except for special cases, initialize the network into a "symmetric" state where each node forms its own community ($q_0 = N$).

(2) *Test for a node move.* Select a candidate node for a possible move to a new state $\{\sigma_i\}$ (see below). The probability of moving a given node is based on a Boltzman weight $P(\{\sigma_i\}) = e^{-\Delta E/T}$ (where we take the Boltzman constant to be $k_B = 1$) for the next candidate configuration $\{\sigma_i\}$ at a specified temperature T and an energy change ΔE (see Ref. [16], for example). We sum the probabilities for all candidate moves where the node may (*i*) stay in its current cluster, (*ii*) move to each *connected* cluster, (*iii*) or move to a new empty cluster (except when the current cluster size is already $n = 1$). We normalize the probability distribution, generate a random number, and move the node accordingly. The node is then "frozen" (not allowed to move again) for the remainder of the current iteration (see step 3).

(3) *Iterate through all nodes.* Rather than randomly selecting nodes for candidate

moves, we sequentially allow each node an opportunity to move, per step 2, on each algorithm iteration.

(4) *Test for community merges.* After each iteration through all nodes in step 3, we also allow the possibility for pairs of communities to merge based on the Boltzman weight given in step 2.

(5) *Stop after fixed number of iterations.* Because we are looking at the solution behavior at a fixed temperature in this chapter, we run for a (large) fixed number of iterations. Generally, the number of iterations *required* for a "stable" solution is $O(10^2)$, but for "hard" problems, such as those encountered in this paper during the phase transitions, the number of iterations is $O(10^3)$. In general practice, we would implement a cooling schedule to model a simulated annealing process with an associated convergence criterion.

4.3 Static transition for $T = 0$

In this section, we use the *greedy* algorithm in Sec. 2.2 to solve the community partitions. We illustrate a community detection phase transition via the accuracy (as measured by VI) on the noise test benchmark in Sec. 2.4.2 and a "susceptibility" on a much smaller benchmark discussed in Sec. 2.3.

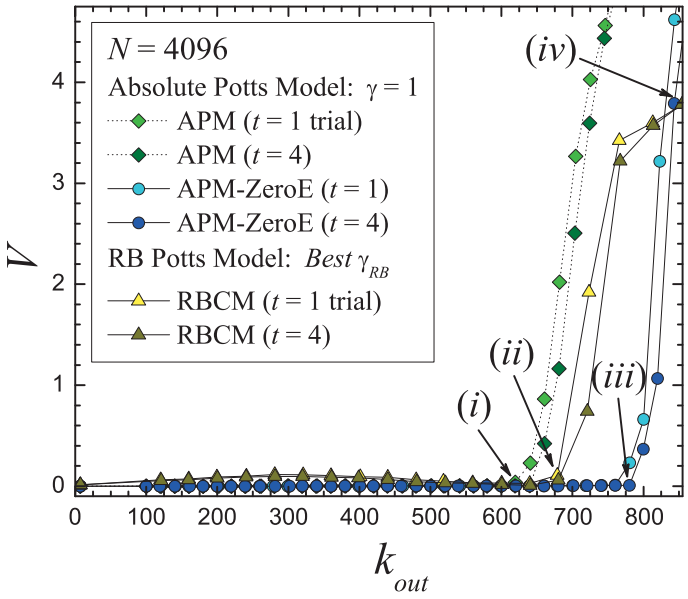


Figure 4.1: (Color online) Plot of VI V vs the average external degree k_{out} for the APM and RBCM models. Similar to Sec. 2.4.2, we generate strongly defined communities with $N = 4096$ nodes and high levels of intercommunity noise. We use the greedy algorithm in Sec. 2.2 to solve the systems for both 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.

4.3.1 Accuracy transition

In Fig. 4.1, we construct a set of well-defined but noisy systems with $N = 4096$ nodes from the benchmark in Sec. 2.4.2. An initial random degree distribution is assigned 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. 2.2.

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. 2.2 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.* [87]. 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 4.1 illustrates a similar transition in community detection.

4.3.2 Transition via a "Susceptibility"

Another approach to view the phase transition is depicted by the data in Fig. 4.2. The benchmark problem that serves as the basis for this data is discussed in detail in Sec. 2.3. In Fig. 4.2, we plot for several numbers of trials n, the "susceptibility" $\chi_n \equiv p(t = n) - p(t = 4)$ versus Z_{out}, the average number of edges that each node has connected exterior to its own community. The average number of total edges per node is $Z = 16$. p is the percentage of correctly identified nodes from Fig. 2.1 (see Ref. [19] in [62]), and t is the number of trials at each test. The ordinate χ in Fig. 4.2 is the percentage improvement in accuracy based on the number of optimization trials that are used.

As Z_{out} increases, the noise in the system increases. Figure 4.2 illustrates how the noise in the system affects the effort required to solve the system as accurately as

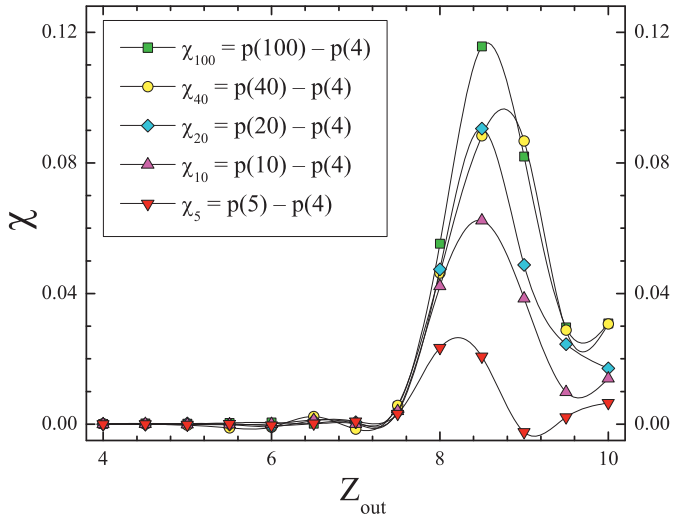


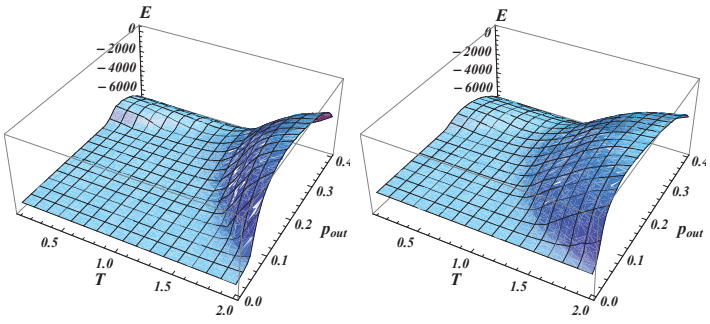
Figure 4.2: (Color online) A plot of the susceptibility $\chi_n \equiv p(t=n) - p(t=4)$ versus Z_{out}, the average number of edges that each node has connected exterior to its own community. χ_n is the percentage increase in the accuracy of each test as the number of trials $t = n$ is increased from $n = 5$ to $n = 100$. The average number of total edges per node is $Z = 16$. p is the percentage of correctly identified nodes from Fig. 2.1. The curves are spline fits and are intended for visualization purposes only. Additional trials are unnecessary in the easy region $Z_{out} \lesssim 7$. The benefit of extra trials is largest in the short transition region $8 \leq Z_{out} \leq 9$, and the benefit diminishes into the hard region $Z_{out} \gtrsim 9.5$ where the accuracy improvement is small even with a large number of attempted optimization trials.

possible. The benefit of extra optimization trials is negligible for the easy region up until about $Z = 7$. Additional trials are more important for a short transition region ($8 \leq Z_{out} \leq 9$). After this region, the benefit quickly reaches a point of diminishing returns in the hard region $Z_{out} \gtrsim 9.5$ where it fails to produce large improvements in accuracy despite significantly more computational effort.

As the number of trials n increases, the "susceptibility" χ_n progressively exhibits a more pronounced peak. Such a trend is also evidenced in the susceptibility of finite size physical systems. We have also identified a similar and related dynamic feature of the transition that is quantified by the increased computational time required for a single solution [74] (beyond any added computational cost due to extra energy optimization trials).

4.4 Static transition for $T > 0$

We again use the noise test benchmark in Sec. 2.4.2 to study phase transitions in our problem for simulation temperatures $T > 0$. That is, for each constructed benchmark graph, we start with N nodes divided into q communities with a power law size distribution ($\beta = -1$). In order to create a strongly defined community structure, we connect all intracommunity edges at a high average edge density $p_{in} = 0.95$. One distinction in this section is that we implement a fixed density of random external edges $p_{out} < 0.5$ (noise) as opposed to using a power law distribution of noise as in Sec. 2.4.2. We use the above HBA in Sec. 4.2 at a temperature $T > 0$ to solve each test network. Obviously, the higher the noise level, the more difficult the system will be to solve. Similarly, a higher solution temperature T makes the constructed configuration harder to detect.



(a) $q = 80$ (b) $q = 100$

Figure 4.3: Plots of energy $E(T, p_{out})$ where T is the heat bath temperature and p_{out} is the level of network noise for a system with $N = 1024$ nodes. The number of communities is $q = 80$ and 100 in panels (a) and (b), respectively. The energy here refers to an ensemble average energy over 100 replicas at time $t = 1000$. Both panels show a phase transition from a "flat region" (solvable) to a "hill region" (difficult-to-solve) as T and p_{out} increase. From panels (a) to (b), the solvable region decreases as q increases, which matches the complexity trend.

4.4.1 Energy transition

In panel (a) of Fig. 4.3, we show a 3-dimensional (3D) plot of the system energy $E(T, p_{out})$ for $q = 80$. As expected, the energy exhibits a sudden jump as the noise p_{out} exceeds some critical value p_c. A similar critical behavior is observed for solution temperature T (the algorithm misplaces some nodes). We can determine the transition region in each system by observing the values of p_{out} and T that show rapid increase in E. We can solve the system perfectly in the "flat" region which shows relatively low noise for many temperatures. In panel (b), we also show that a larger

q (smaller communities) makes the system more difficult to solve (transitions occur earlier in p_{out} and T) because the noise has a larger relative energy contribution when used to evaluate the best community memberships. At low T and high p_{out}, we see an additional small bump in E where we observe that the near "greedy" application of the solver cannot adequately optimize the solution.

The relation between energy and noise can be extracted from this 3D plot by fixing the temperature. The resulting curve shows a peak behavior as the noise passes a critical point p_c, which is a site of the easy-hard transition. In other words, the system moves from solvable, to difficult-to-solve, to unsolvable as noise increases. This kind of transition is confirmed by the computational effort required to solve the system as a function of noise [96], which displays a peak behavior as well.

A simple explanation of the energy transition is as follows: As p_{out} increases from 0, the system is able to stay ergodic at low levels of noise. That is, the algorithm can still traverse any encountered local energy minima to find the optimal solution, so the energy stays constant (there is only one global energy minimum); however, it does take progressively more time, in a non-linear relation, to locate the global minimum state. Secondly, as p_{out} passes the critical value p_c, the system is still ergodic, but it takes a very long time to find the lowest energy state. In a finite time scale, the system stays near a local minimum state thus yielding a higher energy. Lastly, as p_{out} becomes significantly larger than p_c, the system starts to lose ergodicity, and so it takes shorter time to converge.

Following this explanation, increasing the running time would help increase the

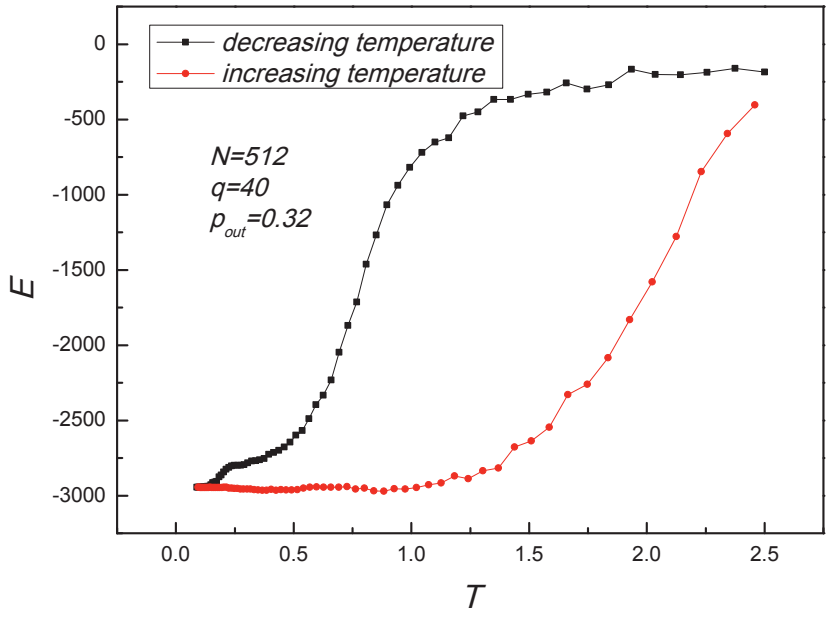


Figure 4.4: Plot of energy E vs temperature T for a system with $N = 512$ nodes, $q = 40$ communities, and $p_{out} = 0.32$. The HBA temperature begins at $T = 2.5$ and decreases by $T_{k+1} = 0.95T_k$ per step (one loop through all nodes) for each solution step k on the network. After a steady-state solution is obtained, the process is reversed by $T_{k+1} = 1.05T_k$. Note that the network solution shows a clear hysteresis-like effect.

accuracy of the solution in the hard region (the peak area). After this region, the system requires essentially an infinite amount of time to solve accurately. The accuracy aspect of the transition was observed in [17] at $T = 0$. The non-zero temperature case can also be verified by plotting the energy versus time in the hard region [96].

The breakdown of ergodicity indicates a non-equilibrium system in that region. This non-equilibrium behavior would also cause a memory effect which has been

previously studied for other spin glass systems, both in experiment and theory [97, 98]. When a spin glass is cooled down, a memory of the cooling process is imprinted in the spin structure, and this process will be reproduced if one heats the system up.

In order to determine whether our system has a similar memory effect, we conduct a similar computational "experiment." Basically, we put our system in a heat bath, and we lower the heat bath temperature T by a small amount each step k (a single iteration through all nodes), i.e., $T_{k+1} = 0.95T_k$. After we reach a steady-state solution, we then reverse the process and increase T after each step, i.e., $T_{k+1} = 1.05T_k$.

In Fig. 4.4, we plot the system energy E as a function of T during this process. The energy curve as T decreases follows a different path than when T increases which strongly implies a hysteresis-like effect. This effect reinforces the similarity between the community detection and a spin glass system.

Examples that show a memory effect are not limited to this one [99]. For instance, if we add noise to the same system and then sequentially remove them, the accuracy of the solution also forms a hysteresis loop at low temperature [96]. Similar to a real spin glass system, the magnitude of this effect also decreases as the temperature increases, and it finally disappears at some critical temperature in the community detection problem. The system experiences a transition from spin-glass-like to a normal state, and we could define this transition temperature T_g as a typical glass transition for the community detection system.

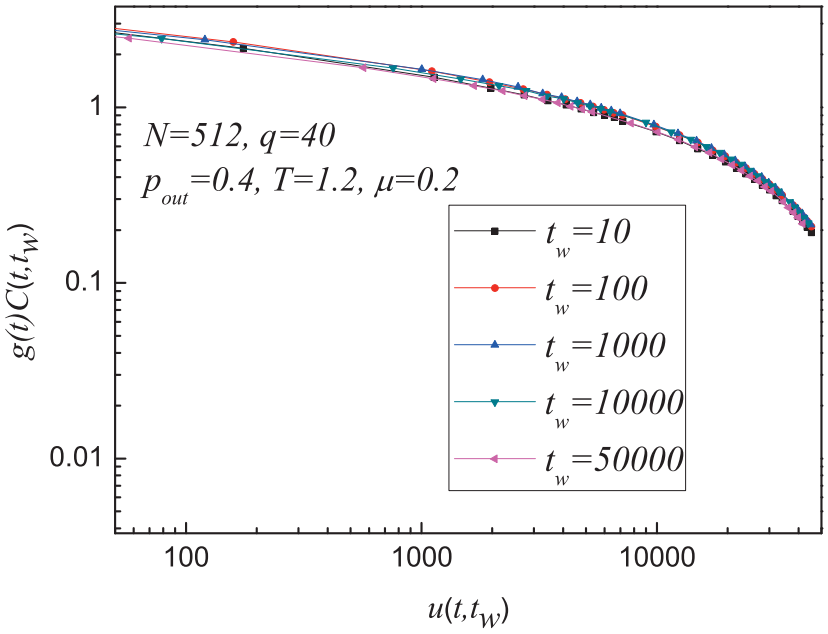


Figure 4.5: The collapsed curves for the correlation function at different waiting times t_w for a system with $N = 512$ nodes, $q = 40$ communities, and $p_{out} = 0.4$. The heat bath temperature is $T = 1.2$. The y-axis is $g(t)C(t_w, t)$ where $g(t) = 6 - \log_{10}(t)$. The x-axis is $u(t_w, t) = \frac{1}{1-\mu}[(t + t_w)^{1-\mu} - t_w^{1-\mu}]$ where $\mu = 0.2$. Note that the value of μ is smaller than the most common value, which indicates that it is more difficult to reach equilibrium in our system.

4.4.2 Time correlation function

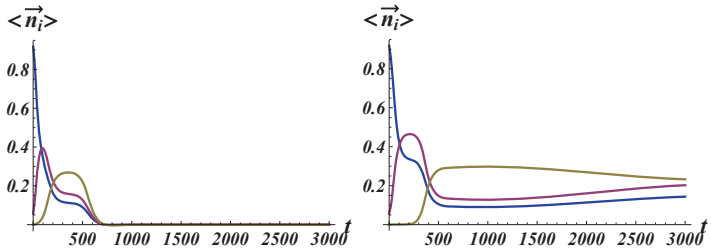
Furthermore, the two-time autocorrelation function, which is defined as

$$C(t_w, t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma_i(t_w), \sigma_i(t_w+t)}, \tag{4.1}$$

can be also used to explore the spin-glass-like behavior. $C(t_w, t)$ denotes the autocorrelation between time t_w and $t + t_w$, N is the number of nodes, $\sigma_i(t_w)$ denotes the community membership for node i at time t_w.

We use the HBA starting from a symmetric initial state and calculate the autocorrelation in Eq. (4.1) for different waiting times t_w and temperatures T. Each correlation curve with longer waiting time lies above those with shorter waiting times, and all the curves (with different waiting times) are non-zero for a long period of simulation time indicating that the correlation function also demonstrates a memory effect in the studied system [96]. Moreover, we can predict the long time behavior of $C(t_w, t)$ by fitting the curves using a commonly-used equation in Fig. 4.5 [100, 101].

If we apply the HBA starting from different initial configurations at low temperature, all the correlation curves with different initializations separate from each other even up to t as large as $t = 10000$ [96]. Then as T increases, all the curves start moving towards, and finally overlap, each other. The temperature at which the different initial configurations overlap indicates when the respective systems start losing any memory of their initial configurations, and it directly relates to the glass transition temperature T_g in the hysteresis loop for the same system. This further establishes the existence of spin glass transition in the community detection problem.



(a) $p_{out} = 0.2$ (b) $p_{out} = 0.3$

Figure 4.6: Plots of node trajectories $\langle \vec{n_i} \rangle$ as a function of time t (number of algorithm steps). The tested system has $N = 24$ nodes, $q = 4$ communities, and is solved at $T = 0.05$ using a HBA. Node i is picked randomly from the 24 nodes. $p_{out} = 0.2$ is below the critical transition in noise for these network parameters in panel (a), and $p_{out} = 0.3$ is above the transition in panel (b). Note that panel (a) shows a perfect dynamical solution for node i where panel (b) indicates an incorrect solution attempt on average.

4.5 Dynamic transition

We also analytically study the related dynamical transition. To describe the dynamical process, we need to calculate the trajectory (of community memberships) for each node as a function of time. Specifically, we replace the delta function $\delta(\sigma_i, \sigma_j)$ in Eq. (2.2) by a product $\vec{n_i} \cdot \vec{n_j}$ where $\vec{n_i}$ and $\vec{n_j}$ are the vertices of a regular $(q-1)$-dimensional simplex which satisfy the equation

$$\vec{n_i} \cdot \vec{n_j} = \left[1 + \frac{1}{q-1}\right] \delta_{ij} - \frac{1}{q-1}, \tag{4.2}$$

where δ_{ij} is the Kronecker delta. By this transformation, we can transfer the Hamiltonian in Eq. (2.2) to an Ising model form,

$$H = -\sum_{ij} C_{ij}\vec{n_i}\cdot\vec{n_j}, \tag{4.3}$$

where C_{ij} is the interaction weight. We then use the Hubbard-Stratonovich transformation by introducing a scalar auxiliary field $\vec{\eta}$ to derive the effective Hamiltonian

$$\beta \times H_{eff} = \sum_{i\neq j} \ln\left(e^{\vec{\eta_i}A'^{-1}_{ij}\vec{\eta_j}}\cdot Tr_{\vec{n_i}}e^{\vec{n_i}\vec{\eta_i}}\right) \tag{4.4}$$

in which $A'_{ij} = (1+\gamma)A_{ij} - \gamma$.

The dynamical equation for a node moving under the effective field is

$$\begin{aligned}\frac{d\vec{\eta_i}}{dt} &= -\frac{\delta H_{eff}}{\delta\vec{\eta_i}} \\ &= -\beta^{-1}\sum_j A'^{-1}_{ij}\vec{\eta_j} + \beta^{-1}\frac{\sum_{\vec{n_j}} e^{-\vec{n_j}\cdot\vec{\eta_i}}\cdot(-\vec{n_j})}{\sum_{\vec{n_j}} e^{-\vec{n_j}\cdot\vec{\eta_i}}}\end{aligned} \tag{4.5}$$

We can solve this dynamical relation to obtain a plot of the auxiliary field $\vec{\eta}$ as a function of time

$$\begin{aligned}\langle\vec{n_i}\rangle &= \frac{\sum_{\vec{n_i}}\vec{n_i}\cdot e^{-\beta H^i_{eff}}}{\sum_{\vec{n_i}} e^{-\beta H^i_{eff}}} \\ &= \frac{\sum_{\vec{n_i}}\vec{n_i}\cdot\exp(-\vec{\eta_i}A'^{-1}_{ij}\vec{\eta_j})\cdot Tr_{\vec{n_i}}\exp(-\vec{n_i}\cdot\vec{\eta_i})}{\sum_{\vec{n_i}}\exp(-\vec{\eta_i}A'^{-1}_{ij}\vec{\eta_j})\cdot Tr_{\vec{n_i}}\exp(-\vec{n_i}\cdot\vec{\eta_i})}.\end{aligned} \tag{4.6}$$

Substituting $\vec{\eta}$ in Eq. (4.5) into Eq. (4.6), we can determine the trajectory of the nodes as shown in Fig. 4.6.

The node trajectories $\langle\vec{n_i}\rangle$ always converge to zero in low noise, and they diverge in high noise in Fig. 4.6. The result demonstrates exactly the same kind of phase

transition where the system moves from ergodicity to a breakdown of ergodicity. The vectors $\vec{n_i}$'s are symmetric, so an ergodic system would make the average go to zero, while a non-ergodic system could allow the average to be non-zero. The transition point at which the dynamical process turns from convergence to divergence is consistent with the one found by the static method above.

4.6 Conclusion

In conclusion: (1) We study the energy as the function of the density of intercommunity edges (noise) and temperature of a heat bath solver for a community detection problem. From these data we detect a rapid phase transition from an easy-to-solve to a hard-to-solve problem. (2) The algorithm's solution time for the system as a function of noise exhibits the same phase transition, and the critical points determined by both transitions are the same for a given network. (3) In the hard-to-solve region, a network shows a memory effect, which is a sign of a breakdown of ergodicity. (4) We discussed three different examples of the memory effect, from which we can extract the glass transition temperature. (5) We study an effective Hamiltonian as an analytic function of thermodynamic variables. We then develop the node's trajectory equation (in terms of an evolving community membership) from it. (6) By plotting the curve of node trajectories as a function of time, we found that the trajectory shows a dynamic transition from convergence to divergence as the noise increases, and the transition point matches well with the one found in the static case.

Chapter 5

Characterizing amorphous structures

5.1 Introduction

Amorphous materials often possess desirable properties relative to their respective crystalline counterparts. For example, amorphous materials in general possess industrial processing and preparation advantages [102, 103], greater solubility of pharmaceuticals [104] and other advantages [102, 105]. Metallic glasses can be stronger than their respective crystalline structure due to fewer realized material defects, and they can possess other interesting electrical, chemical, and magnetic properties [103].

In perfect crystals, the natural system scales are evident by the regular ordering of the lattice. The fundamental unit cells of a crystal typically involve several atoms that are replicated in a simple pattern to span the entire system. There are no

intermediate scale structures that transition the system from the atomic scale of the lattice up to the complete single crystal. Identifying the basic periodic unit cells is vital to the understanding of all crystalline solids, and it is the simplicity of the revealed structure that enables scientists to understand the behavior of these solids in great detail. Early on, the existence of specific unit cell structures were postulated to exist in crystals based on the sharp facets and other macroscopic properties of large crystals. The only natural structure is that of the basic unit cell which replicates itself everywhere (including on the largest scales).

There are other more complex systems in which new structures appear on additional intermediate scales between the atomic-scale and the macro-scale of the system. A simple example is that of a crystal composed of different domains. These distinct domains provide a natural definition of intermediate range structure for these systems even though the basic periodic unit cells are essentially unchanged in each domain. Although basic ordered materials form a fundamental pillar of modern technology (e.g., the transistor was made possible by an understanding of the electronic properties of ordered periodic crystals with introduced impurities), there are many other systems whose understanding is extremely important but are lacking due to the complexity of their structure. In recent years, scientific exploration has endeavored to understand a vast array of such complex materials that do not have a simple theoretical starting point. Such systems range in scope from *structural glasses to complex electronic states.*

Some of the oldest complex materials are glasses which are still not well under-

stood even after millenia. Liquids that are rapidly cooled ("supercooled") below their melting temperature cannot crystallize and instead become "frozen" into an amorphous state. On supercooling, liquids may veer towards local low energy structures, such as icosahedral structures observed in metallic glasses [106, 107], before being quenched into an amorphous state. Because of the lack of a simple crystalline reference, the structure of glasses is notoriously difficult to quantify beyond the very local scales.

The most familiar and oldest technological glasses are the common silicate glasses. More modern glasses include phosphate glasses (biomedical applications), semiconductor chalcogenide glasses (optical recording media), and aforementioned metallic glasses. In addition to their technological applications, some fundamental problems in physics involve a better understanding the nature of the glass transition and the character and evolution of amorphous structures.

Many theories of glasses rely on the hypothesis of natural structures in the glass [108, 109, 110, 111]. Actually finding such structures in a general way has been more elusive. How then does one best characterize the most "natural" structures in amorphous systems? Here, we attempt to provide a general framework to answer this question with specific applications to two model glass formers.

5.2 Background

Existing work in the pursuit of understanding the glass transition is vast, spanning decades and affecting many fields of science and engineering. Glass transitions can be characterized by a number of different, but related, criteria [112]: the viscosity and relaxation times can increase by many orders of magnitude with little change in usual measures of order or quantities that accompany known phase transitions (although recent work [113] may imply a stronger relationship than was previously thought). Glasses demonstrate short range order (SRO) and medium range order (MRO) structures, but no long range order exists. From an energy landscape perspective, the number of metastable energy states increases dramatically through the glass transition [114, 115]. Some other works include a direct analysis of the potential energy landscape of a glass (not in a graph theoretic mold) are found in [116, 114, 115, 117, 118, 119, 120].

Given the broad appearance of a glass-related states in matter, different frameworks have been explored to work towards a "universal" characterization of the glass transition, such as geometry based frustration [108, 109, 111, 110] and the appearance of topological defects as well as possibly related kinetic constraints [111, 110, 121, 122, 123].

There is a rigorous proof that a growing length scale must accompany the diverging relaxation times of glass [124]. Some evidence has been found for a growing correlation length [125, 126, 127, 128]. Correlation lengths have also been studied in terms of

"point-to-set" correlations [129, 130].

In metallic glasses, early work to ascertain local structural, as opposed to mesoscopic, features in monatomic systems used a dense random packing model [131]. It was later established that such structures are better represented by an efficient cluster packing (ECP) model [132, 133, 134]. SRO features were thought to pivot on the existence of local icosahedral structures centered around solute atoms. Various idealized SRO configurations were presented in [135]. Schenk *et al.* experimentally verified icosahedral short range order (ISRO) in undercooled liquids [106]. Kelton *et al.* were the first to experimentally establish a connection between ISRO and the nucleation barrier [107]. Later work further established the importance of ISRO in glasses [136, 137, 138].

Many structural characterizations are oriented toward static viewpoint of the system, but some dynamical features have also been examined. Analysis of "free volume" (unoccupied space between atoms) fluctuations [134] have been used. Shear stress calculations investigate dynamical processes in glass forming materials [119, 139]. Dynamic heterogeneities involving cooperative motion of structures in a glass have also been studied [140, 141, 142, 143].

Characterizations of SRO and MRO structures have been proposed or analyzed in various settings for low [144] and high [138] solute concentrations, binary systems [144, 138], or multicomponent systems [133].

Some methods of characterizing local structures include Voronoi tesselation [145, 123, 144], Honeycutt-Andersen indices [146], and bond orientation ordering parame-

ters [147]. These local measures center on an atom or a given link and, by definition, are restricted from detecting more complex longer range general structures.

Experimental means to directly measure MRO structures are given in [148, 149]. Some potential MRO clusters were examined [120, 144, 138]. Some approaches to understand MRO use pattern matching to idealized MRO structures often constructed as agglomerations of perfectly ordered SRO features.

Our unbiased structure characterization method extends *multiresolution* ideas [37] in network science to complex materials. Any complex physical system may be expressed as a network composed of nodes that code basic units of interest (e.g., atoms, electrons, etc.). Weighted links capture the strength of the interactions between the different nodes or experimentally determined correlations (e.g., covariance or partial correlation contributions to the structure factor). After casting the system as a network, we then search for "communities" of nodes (*i.e.*, clusters of atoms) that are more tightly linked to each other than to nodes in other clusters [17].

Our multiresolution method extends the idea of community detection to quantitatively identify the "best" scale (or *scales*) for a complex physical system. Our approach *does not* rely on intuition or a knowledge of expected "important" features. Rather, it quantitatively estimates the best scale(s) through information-theory-based correlations, such as the variation of information (VI) [86] or normalized mutual information (NMI), among different solutions. We may imagine that these solvers are a group of people all assigned the task of examining the structure of the same network In essence, different copies of the community detection problem are given to indepen-

dent solvers ("replicas"). If many of these solvers strongly agree regarding certain features of the solution, then these aspects are more likely to be correct.

Extrema in NMI or VI (and sometimes information *plateaus*, see Appendix I) between the independent solvers indicate the best scales for the network. Multiple extrema can indicate *multiple relevant length scales* and different time and length scales appear in, e.g., α and β relaxation processes of structural glass formers. The analysis presented here uses a fixed time separation between replicas, but we may further analyze a range of time separations between replicas which would allow us to ascertain the most relevant *time* scales of the system.

While the strength of the replica correlations is related to the relative atomic positions by design, one distinction between our work and some other established studies of local structures in glasses is that our analysis is not looking strictly at the positional structure. Rather, it evaluates structures in terms of the potential energies (i.e., the internal binding energies of the clusters, see also [120]). A disadvantage of this approach is that it does not apply directly to model glass formers that use repulsive-only or hard-sphere potentials. Advantages of our approach are: As mentioned previously, it is not restricted to searching for expected structural features. Detected features can be on essentially any size scale of the model system. Our method can encapsulate weights that represent general statistical (pair of higher order) stress (or other) correlation functions.

Our approach provides a perspective different from the "point-to-set" [150, 151, 124] and other methods [126]. The point-to-set method examines the overlap between

configurations in a given volume (a "cavity") in an equilibrated system and compares those to configurations in the same cavity of the equilibrated system in which the boundary of the cavity was held fixed. Physically, it probes how probable it is to have a particular configuration within a disk or ball of a particular diameter given the boundary conditions. If many states exist inside some sphere of some fixed radius, a change in the boundary conditions will not significantly alter the cluster distribution. Conversely, if the sphere radius is smaller than the natural correlation length, then the number of configurations compatible with the boundary will be small and the overlap will be large. A different approach is in Ref. [126] which examines the distribution of structures inside a given volume to identify the correlation length. The method examines whether the distribution of configurations inside the volume occurs with a random frequency (when the linear scale of the volume is *larger* than the correlation length) or not (when the linear scale of the volume is *smaller* than the correlation length).

Our method looks does not look for overlap at different scales for a multitude of configurations nor their frequency. Rather, the pertinent structures are revealed by the information theory extrema between different copies of the entire system. We do not need to directly tabulate possible configurations and their occurance probabilities nor examine the system in restricted volumes.

Furthermore, the basic structures that we find *may be used as the natural units in a renormalization group type analysis* where clusters are replaced by single nodes and an effective energy can be written that entails interactions between the different clusters

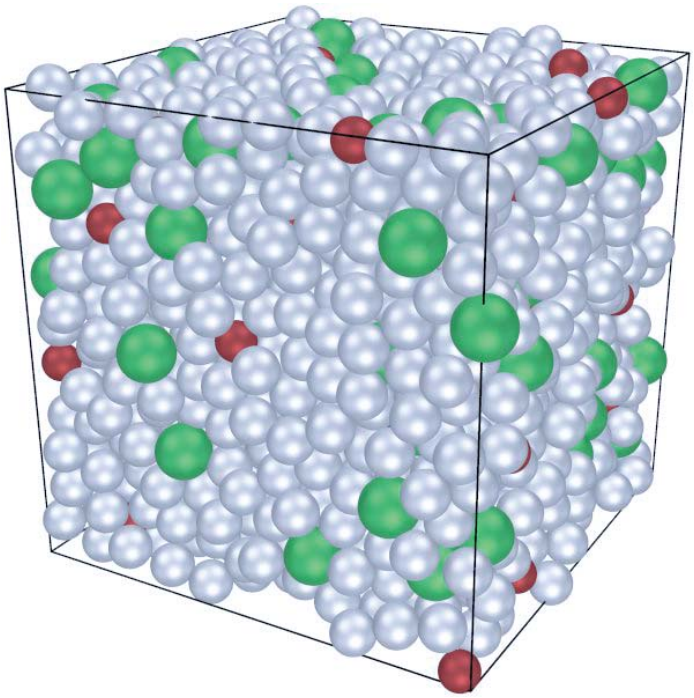


Figure 5.1: A depiction of our simulated model glass former with three components "A", "B", and "C" with mixture ratios of 88%, 7%, and 5%, respectively. The $N = 1600$ atoms are simulated via IMD [152] in cube of approximately 31 Å in size with periodic boundary conditions. The identities of the atoms are C (red), A (silver), B (green) in order of increasing diameters.

alone. Finding the basic units is not trivial in amorphous systems such as glasses or other disordered systems. In these systems, there is no symmetry and no obvious knowledge of how to optimally partition the system as we go up in scale. Physically, the community detection algorithm seeks to find a permutation that renders the interaction maximally block diagonal and sparse with minimal interactions between the blocks.

	a_0	a_1	a_2	a_3	a_4	a_5
AA	*	*	*	*	*	*
AB	1.92	17.4	6.09	3.05	-4.68	3.48
AC	2.38	8.96	-14.9	3.11	-3.88	4.38
BB	*	*	*	*	*	*
BC	1.88	8.00	-3.42	2.53	-1.25	3.00
CC	*	*	*	*	*	*

Table 5.1: Fit parameters for Eq. (5.1) obtained from fitting configuration forces and energies to *ab initio* data. The same-species (*) data is replaced by a suggested potential derived from generalized pseudo-potential theory [155] (see also Appendix H).

5.3 Simulations of model glasses

We examine a model glass former derived from a three-component AlYFe metallic glass [153] which we designate as "A", "B", and "C" in mixture ratios of 88%, 7%, and 5%, respectively. The presence of the different components B and C assists in the formation of a glassy state [154].

5.3.1 Ternary model glass former

As depicted in Fig. 5.1, one system that we examine is derived from a three-component AlYFe metallic glass. The system t is a model glass former with components designated as "A", "B", and "C" in mixture ratios of 88%, 7%, 5%, respectively. We use classical molecular dynamics (MD) [152] to simulate the system dynamics. For this, we need accurate effective pair potentials that portray the pairwise interactions

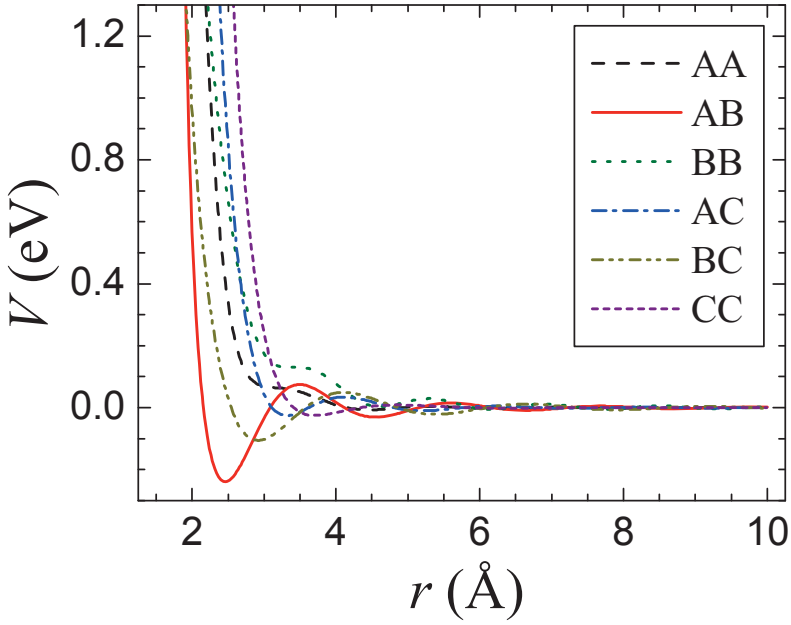


Figure 5.2: A plot of the model potentials for our three-component model glass former (see Fig. 5.1). We indicate the atomic types by "A", "B", and "C" which are included with mixture ratios of 88%, 7% and, 5%, respectively. The units are given for a specific candidate atomic realization (AlYFe) discussed in the text. The same-species data uses a suggested potential derived from generalized pseudo-potential theory [155] (see also Appendix H).

between the atoms in the system. Our model potential energy function is [156]

$$\phi(r) = \left(\frac{a_0}{r}\right)^{a_1} + \frac{a_2}{r^{a_5}} \cos\left(a_3 r + a_4\right) \tag{5.1}$$

where it incorporates a realistic weak long range interaction. r is the distance between the centers of two atoms. Table 5.1 summarizes the parameter values a_i which depend on the specific types for a pair of interacting atoms, and Fig. 5.2 shows the respective potential plots.

The interaction parameters a_i were determined [156] by fitting configuration forces

and energies to *ab initio* data [157]. The same-species model interactions are finally replaced by that suggested by generalized pseudo-potential theory (GPT) [155]. As depicted in Fig. 5.1, we simulate $N = 1600$ atoms in a cubic system approximately 31 Å in size using periodic boundary conditions. This width is approximately twice the size of any suspected MRO structures.

The system is initialized at a temperature of $T = 1500$ K and allowed to equilibrate for a long time using a constant number of atoms, a constant volume, and a constant energy (NVE). After allowing for system equilibration, we save s high temperature configurations separated by a fixed period of simulation time. Prior to cooling, the length scales in the system are changed by 1% to account for the increase in density as a result of cooling since we choose to cool the system in an NVT ensemble to control the temperature. The system is then rapidly quenched to $T = 300$ K, and it is allowed to equilibrate in this mostly frozen state in an NVE ensemble. We again save s separate low temperature configurations separated by a long period of simulation time.

5.3.2 Lennard-Jones glass

We additionally test the ubiquitous Lennard-Jones (LJ) potential using the Kob-Andersen (KA) 80:20 binary liquid [158] which lies in the glass-forming mixture region [159]. The potential is

$$\phi_{\alpha\beta}(r) = 4\epsilon_{\alpha\beta}\left[\left(\frac{\sigma_{\alpha\beta}}{r}\right)^6 - \left(\frac{\sigma_{\alpha\beta}}{r}\right)^{12}\right] \tag{5.2}$$

where α or β designate one of two atomic types A and B. Specifically, in accord with KA we set the dimensionless units $\epsilon_{AA} = 1.0$, $\epsilon_{AB} = 0.50$, $\epsilon_{BB} = 1.5$, $\sigma_{AA} = 1.0$, $\sigma_{AB} = 0.88$, and $\sigma_{BB} = 0.80$.

As in the ternary glassy system above, we use MD [152] to simulate a LJ system of $N = 2000$ atoms. The system is initialized at a temperature of $T = 5$ (using energy units where the Boltzmann constant $k_B = 1$) and allowed to evolve for a long time. We save s high temperature configurations separated by 1000 time steps selected so that the configurations are separated by times that are of the order of the caging time (see KA [158]). The time step size is $\Delta t = 0.0069$ in LJ time units. Then the system is rapidly quenched to $T = 0.01$ which is well below the glass transition temperature of the KA-LJ system. The system is allowed to run in this mostly frozen state, and we save s low temperature configurations separated by 1000 steps of simulation time.

5.4 Multiresolution clustering on amorphous materials

Our idea is to apply, for the first time, multiresolution network analysis methods to ascertain pertinent structures in complex amorphous materials. Using concepts developed in Chapter 2, we define a model network by means of direct physical analogies. We define a node as a single atom. Edges, and their corresponding weights, are directly defined by the associated pair-wise potential energy. Specifically, we use the interatomic potential energies in Eqs. (5.1) and (5.2). This model of weighted

network edges is physically appealing in that finding the best partition for the network is akin to minimizing the cluster binding energies of the physical system. In principle for this application, we could further generalize the community detection Hamiltonian of Eq. (2.2) to include n-body correlations or interactions.

Our community detection algorithm in Sec. 2.2 partitions the network into communities by assigning a unique cluster membership for each node. Local features in metallic glasses generally exhibit interconnecting short range structures [144]. In our community detection problem, this feature corresponds to allowing "overlapping" node memberships where atoms can be members of more than one local cluster. We incorporate this effect by assigning a node as a secondary member of every community for which it has a negative binding energy in terms our Potts model in Eq. (2.2) (see Appendix G).

5.4.1 Motivation and physical analogies

Since the replicas in the current problem represent time-separated configurations, strong agreement among the replicas corresponds to more consistent physical structures over time which fits the intuitive notion of a well-defined natural structure in a physical system. In our current approach, these configurations are solved independently (as opposed to solving the system in a time-dependent sense), but in principle we could add contributions due to time-dependent relations between the graphs.

Two strengths of our community detection method approach include: The analysis is independent of the type of structures that are being analyzed (structured, amor-

phous solid, and possibly even liquid systems). Because edge assignments are based on relative node positions (through the interaction potential), our method should be robust with respect to translational or rotational motion of solid structures in the system (such as crystal nucleation).

We can write down the general partition function for a community partition with intercommunity interactions which, in the usual language, would correspond to the surface terms of clusters in Random First Order Transition theory (RFOT). Our parameter γ effectively plays the role of scaling the relation between surface and bulk terms in RFOT. A high value of γ corresponds to large surface effects while a small γ corresponds to dominant bulk effects.

In an ideal decomposition into communities, there is no interaction between different communities, and the system is effectively that of an ideal gas of disjoint communities. Stated differently, in the simplest setting in which the Hamiltonian would be block diagonal, the evolution of nodes (atoms) in each community would be decoupled from all other nodes in other communities. In such instances, we may treat each community as a different particle in an ideal gas of non interacting such particles. The general problem is to find (the time dependent) permutation that may render it into a nearly/best possible block diagonal form (on the time scale chosen). Community detection emulates this for graphs

For slow cooling of a liquid which enables crystallization, a first order or critical transition appears in the community detection problem (in the partitioning into disjoint ideal gas particles). A similar transition appears in slowly cooled liquids. For

an infinitely rapid cooling of a liquid, the interactions between the particles are those of a spin-glass, and we expect that a spin-glass transition appears in the community detection problem [160].

5.4.2 Application for model glass formers

We assign edges between the nodes (atoms) with the respective weights based on the empirical pair-potentials given by Eqs. (5.1) and (5.2). Specifically, we calculate the potential energy ϕ_{ij} between each pair of nodes i and j in the system and then shift each value by a constant ϕ_0 to obtain $\phi'_{ij} = \phi_{ij} + \phi_0$ (assuming that $\phi_{ij} \to 0$ as $r \to \infty$). The shift $\phi_0 > 0$ is necessary for the community detection algorithm to properly partition the network of atoms since it provides an objective definition of which interatomic spacings are preferable for a well-defined cluster and which are preferred to be excluded from a cluster.

In our particular application here, we calculate the average potential energy of the system and set $\phi_0 = -\phi_{\text{avg}}$. For use in Eq. (2.2), we define an edge with a weight $a_{ij} = -\phi'_{ij}$ between nodes i and j if $\phi'_{ij} < 0$, and we weight any missing links (or "repulsive edges") by $b_{ij} = \phi'_{ij}$ if $\phi'_{ij} \geq 0$. We then solve both model systems over a large range of γ using $s = 12$ replicas and $t = 10$ optimization trials per replica.

While $\phi_0 = -\phi_{\text{avg}}$ is an intuitive shift that accomplishes the goal of an objective cluster definition here, it is not an appropriate shift for some problems. For example, using $\phi_0 = -\phi_{\text{avg}}$ turns out to be problematic in some cases for lattice models. In a general application, we could examine many potential shifts ϕ_0 and look for extrema

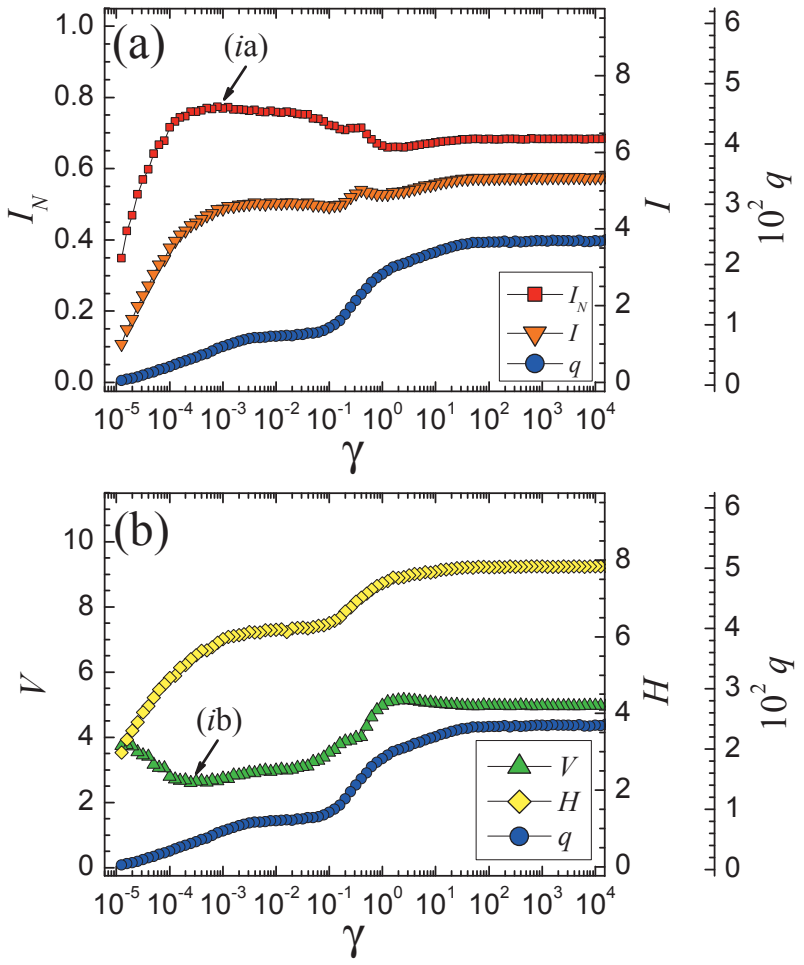


Figure 5.3: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The ternary model system contains 1600 atoms in a mixture of 88% type A, 5% of type B, and 7% of type C with a simulation temperature of $T = 300$ K which is well *below* the melting temperature for this system. This system shows a strongly correlated set of replica partitions as evidenced by the information extrema at (i) in both panels at $\gamma \simeq 0.001$. An example of the best system *partition* is seen in Fig. 5.5, and some sample clusters including overlapping nodes are depicted in Figs. 5.6 and 5.7.

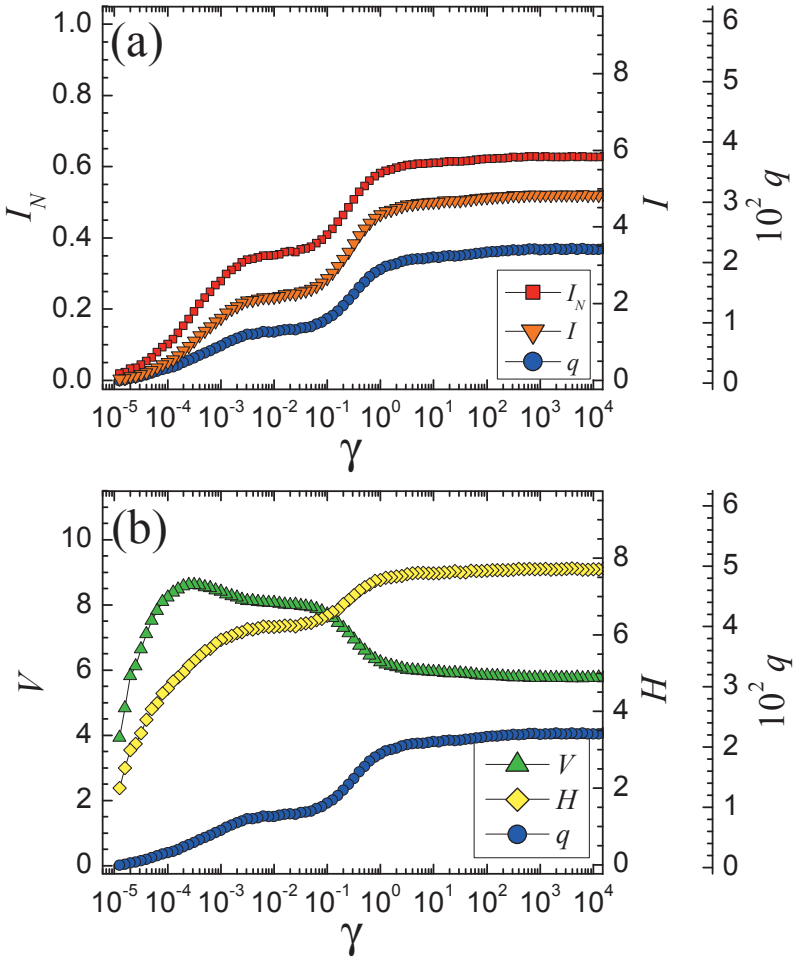


Figure 5.4: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The ternary model system contains 1600 atoms in a mixture of 88% type A, 5% of type B, and 7% of type C with a simulation temperature of $T = 1500$ K which is well *above* the melting temperature for this system. At this temperature, there is no resolution where the replicas are strongly correlated. See Fig. 5.3 for the corresponding low temperature case where the replicas are much more highly correlated at $\gamma \simeq 0.001$.

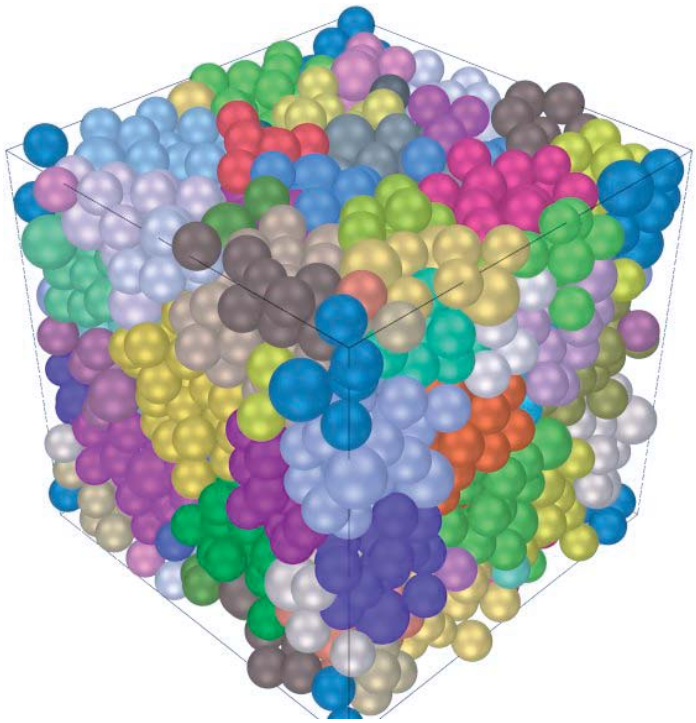


Figure 5.5: A depiction of the full *partitioned* system where unique cluster memberships are depicted as distinct colors (best viewed in color). The atomic identities are B, A, C in order of increasing diameters. Overlapping nodes (multiple memberships per node) are subsequently added to these communities to determine the best interlocking system clusters.

in the information measures V or I_N as a function of both γ in Eq. (2.2) and ϕ_0.

In addition to the tested systems below, we applied the algorithm to various test cases including square, triangular, and cubic lattice structures (see Appendix I). The algorithm is able to correctly identify the natural leading order scales (plaquettes and composites of plaquets as "cascades" in the information theory correlations). We tested identifying natural features and domain boundaries within a 2D Ising lattice (see Appendix J). Further testing (see Appendix K) involved two-dimensional defects

(dislocations, interstitials, etc.) and domain walls (not depicted) in a lattice. Defects in triangular lattices occurred most frequently near cluster boundaries.

We also tested static configurations [161] for the ternary model glass system where each replica is a model of the same configuration. There we detected structures in both low and high temperatures where the high temperature "structures" are more fragile (that is, harder to solve in the clustering problem). This corresponds to identifying relevant transient features in a dense liquid.

5.4.3 Ternary model glass results

In Figs. 5.3 and 5.4, panels (a) and (b) show the data for the replica information correlations over a range of network resolutions. The lower temperature system at $T = 300$ K in Fig. 5.3(a) shows a peak NMI at (ia) with a corresponding VI minimum at (ib). Figure 5.5 depicts example of the full system *partition*, and Figs. 5.6 and 5.7 depict samples of the best clusters at $\gamma_{best} \simeq 0.001$ where we include overlapping node memberships (the replicas correlations are calculated on partitions as in Fig. 5.5). The correponding $T = 1500$ K high temperature solutions have a much lower NMI at $\gamma_{best} \simeq 0.001$ indicating significantly worse agreement among replicas. That is, one would expect that the high temperature system $T = 1500$ K is in a liquid state, so any observed features are not dynamically stable across all replicas (snapshots of the system over time). At $T = 300$ K, the best structures have consistent cluster sizes that are exclusively MRO.

The plateau regions for $\gamma > 10$ are similar to the LJ plot in Fig. 5.8, but in this

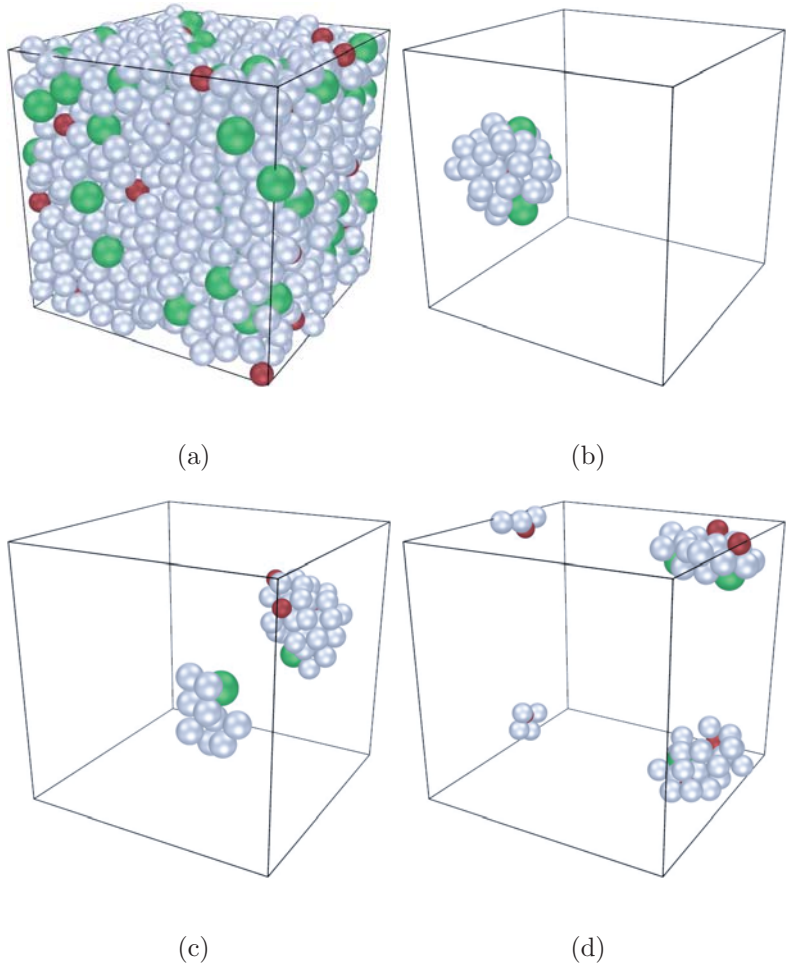


Figure 5.6: Panel (a) is the full system cube, and panels (b) – (d) show three sample clusters (one cluster per box) within the simulation boundaries using periodic boundary conditions. Note that the algorithm identifies structures beyond immediate short range neighbors.

system the NMI plateau is lower. In the high temperature case in Fig. 5.4, there are additional "almost-plateaus" for the range $0.001 \lesssim \gamma \lesssim 0.1$. These plateaus represent a region of structural transition, but we are not concerned with them because the replica correlations are very low.

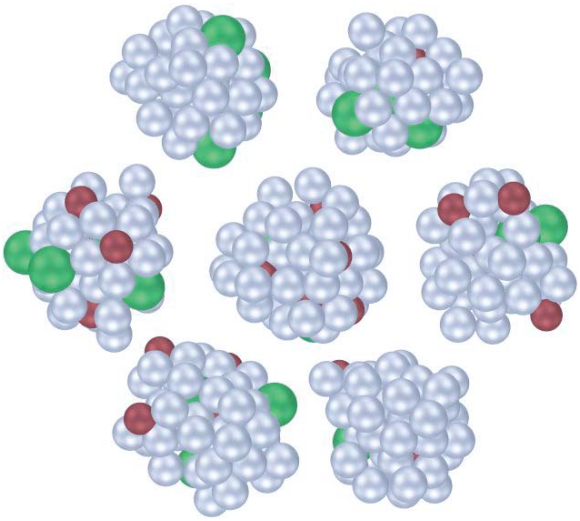


Figure 5.7: A depiction of some of the best clusters for the peak replica correlation at feature (i) in Fig. 5.3. These clusters include overlapping node membership assignments where each node is required to have an overall negative binding energy to the other nodes in the cluster. The atomic identities are C (red), A (silver), B (green) in order of increasing diameters.

5.4.4 Binary Lennard-Jones glass results

In Figs. 5.8 and 5.9, panels (a) and (b) show the data for the multiresolution replica correlations for the simulated LJ system. The lower temperature system at $T = 5$ (in energy units) in Fig. 5.3(a) shows a plateau in NMI at (ia) with a corresponding VI plateau at (ib) which are the local extrema ($V = 0$ at $\gamma \simeq 0.15$ is a trivial solution of a single cluster in this problem). Figure 5.10 depicts a sample of the best clusters, including overlapping node memberships, at resolution (i) for $\gamma_{best} \simeq 10^4$. The corresponding higher temperature solutions at $\gamma_{best} \simeq 10^4$ have a lower NMI (worse agreement among replicas). Our identified structures for this LJ model

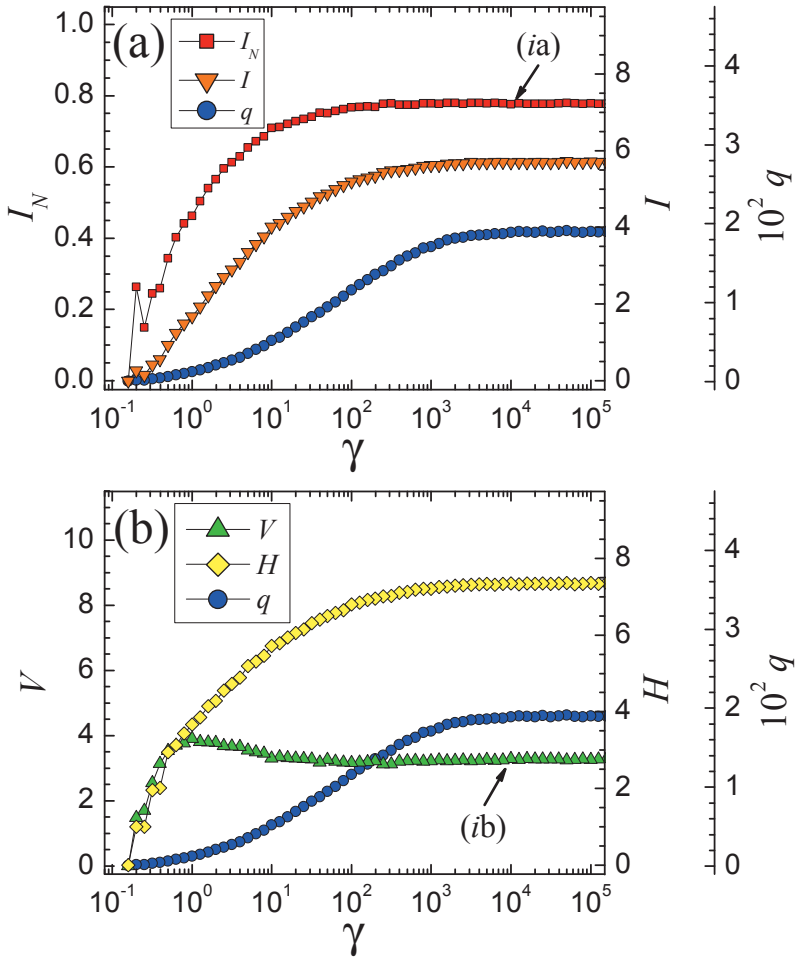


Figure 5.8: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The LJ system contains 2000 atoms in a mixture of 80% type A and 20% type B (Kob-Andersen binary LJ system [158]) with a simulation temperature of $T = 0.01$ (energy units) which is well *below* the glass transition of $T_c \simeq 0.5$ for this system. This system shows a somewhat strongly correlated set of replica partitions as evidenced by the information extrema at (ia,b) in panels (a) and (b). A set of sample clusters for the best resolution at $\gamma = 10^4$ is depicted in Fig. 5.10.

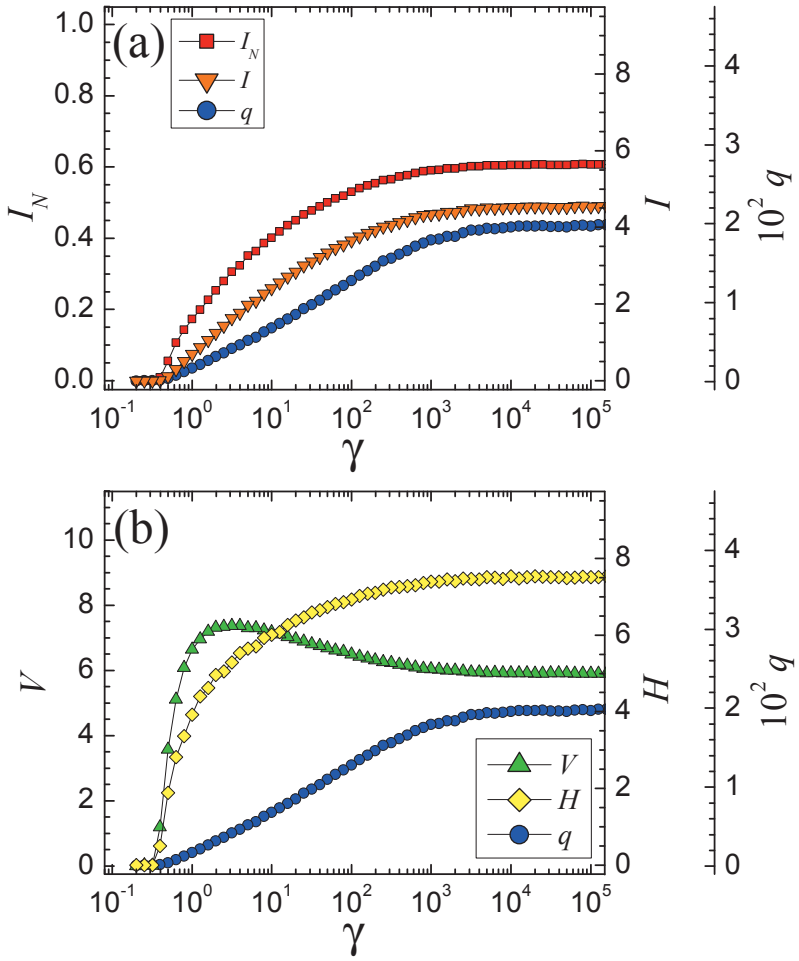


Figure 5.9: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The LJ system contains 2000 atoms in a mixture of 80% type A and 20% type B (Kob-Andersen binary LJ system [158]) with a simulation temperature of $T = 5$ (energy units) which is well *above* the glass transition of $T_c \simeq 0.5$ for this system. At this temperature, the replicas are significantly less correlated than the corresponding low temperature case in Fig. 5.8.

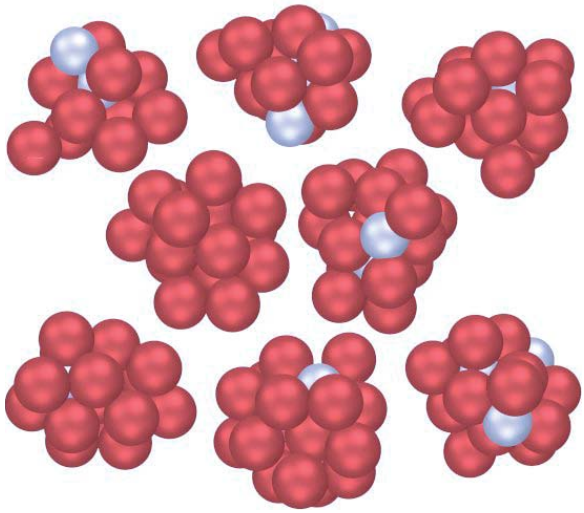


Figure 5.10: A depiction of some of the best clusters for the peak replica correlation at feature (i) in Fig. 5.8. These clusters include overlapping node membership assignments where each node is required to have a overall negative binding energy to the other nodes in the cluster. The atomic identities are B (silver) and A (red) in order of increasing diameters.

system are consistent in terms of the cluster sizes and are almost exclusively SRO configurations with simple adjunct-type atoms extending into the low end of MRO size structures.

5.5 Conclusion

Our algorithm utilizes a network theory model of a set of physical configurations. We identify more cohesive and consistent SRO or MRO structures at temperatures below the glass transition (or in a solid amorphous state) in two different model glasses. Our analysis evaluates structures in terms the potential energies (i.e., the internal binding

energies of the clusters). This approach differs from some other methods of structural analysis that look strictly at the relative atomic positions. When present, the lack of convergence to the same exact clusters, with only the appearance of a similar cluster distribution, suggests high configurational entropy. Our approach identifies MRO as the dominant feature of our ternary model glass former with no strongly defined SRO. In contrast, the LJ system shows a largely SRO structure with adjunct atoms that create near-MRO structures.

Our method is a new and very general approach to determining the natural local and mesoscopic structures of amorphous or other complex physical systems. We do not bias the expected configurations in any way other than to require an attractive model interatomic potential (and/or higher order correlations, in principle) and a set of dynamic configurations from which to define the model networks. The information extrema and/or information plateaus give the different pertinent length scales (lattice scales and correlation lengths) of the system in an unbiased unified way with no prejudice as to what correlation functions should be deemed important.

Compounding the changes in structure that we find by analyzing the atomic system at different temperatures and minimizing the energy function to determine the optimal division into clusters, there are also entropic effects. The distribution of optimal partitions becomes wider and less pronounced also due to these effects as the temperature increases. We also remark that when solving the system of Eq. (2.2) at non-zero temperature for a given network (i.e., atomic configuration that is held fixed), entropic effects can, on their own, lead to a transition as the temperature is

increased [160].

On a lattice, plateaus in information theory correlation steps correspond to a cascade of structures starting from the smallest dyads of nodes, to basic plaquette structures (square, triangle, etc.), and growing ever larger (two joined plaquettes etc.). In Ising spin systems at different temperatures on a square lattice, the domains of "+" and "−" spins are separated from one another by domain walls. The information plateaus correspond similarly to the cascade of small plaquette structures found on the lattice itself (*i.e.*, the single plaquette, two joined plaquettes etc.) up to a cutoff scale set by the domain wall. This is sensible since no clear structure is found beyond the domain length scale.

The largest fluctuations occur at the boundaries between different domains. These domain walls are directly attained by the extrema (those corresponding to the *maximum* in VI). Physically, they correspond to the scales at which the largest fluctuations occur where the large fluctuations lead to poor information theory correlations between the different replicas. Figure J3 corresponds to a sample depiction of the system at the *maximum* VI. Correlation lengths are thus likely related poor correlations in the information theory measures (*maxima* in VI) which is a subject for further study.

Further work could include: (*i*) Analyze the distinctions between different model glass formers that were observed in Figs. 5.3 – 5.10. In particular, the ternary model metallic glass and binary LJ systems show distinctly different features in the multiresolution plots in Figs. 5.3, 5.4, 5.8, and 5.9. (*ii*) These plots provide a characterization of the behavior of the glass formers over a range of system scales, so we could pur-

sue this characterization further to see if it relates to the glass formability. (*iii*) We can investigate any relation between "phase transitions" in the community detection problem [160] (which exist apart from any associated atomic network model) and glass or other transitions known to exist in physical systems. (*iv*) Examine the extremal information correlations as a function of the potential shift ϕ_0 *and* the model weight γ which will provide a more comprehensive analysis of the "best" structures. (*v*) High values of γ roughly correspond to small scale features, and conversely, low values γ roughly correspond to larger features up to the size of the system. Thus, we speculate that there may be a relation between α and β relaxation times and the behavior of our replica correlations in the different regions. (*vi*) We can use the same method to detect general spatio-temporal structures beyond time correlations using the corresponding classical action S from which the equations of motion follow. This action replaces the use of the energy at fixed times. The system will evolve in space-time to minimize the action, and we can represent the system as a network in $(D+1)$ space-time dimensions where D is the number of spatial dimensions. We then employ our algorithm to find the communities in space-time. A $(D+1)$ dimensional action for elastic media including a "dualized form" for the study of defects was given in Refs. [162, 122].

Bibliography

[1] Leon Danon, Albert Díaz-Guilera, and Alex Arenas. The effect of size heterogeneity on community identification in complex networks. *J. Stat. Mech.: Theory Exp.*, 11(11):P11010, 2006.

[2] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure of complex networks. *New J. Phys.*, 11:033015, Mar 2009.

[3] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826, 2002.

[4] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature (London)*, 435:814–818, 2005.

[5] Aaron Clauset. Finding local community structure in networks. *Phys. Rev. E*, 72:026132, Aug 2005.

[6] Vincent D. Blondel, Jean-Loup Guillaume, Renaude Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.*, 10(10):P10008, Oct 2008.

[7] J. J. Xu and H. Chen. Crimenet explorer: A framework for criminal network knowledge discovery. *ACM Trans. Inf. Sys. Secur.*, 23:201–226, 2005.

[8] Naoki Masuda. Immunization of networks with community structure. *New J. Phys.*, 11:123018, Dec 2009.

[9] Roger Guimerà and Luśs A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature (London)*, 433:895–900, 2005.

[10] M. E. J. Newman. The physics of networks. *Phys. Today*, 61(11):33–38, 2008.

[11] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002.

[12] Santo Fortunato and Claudio Castellano. in *Encyclopedia of Complexity and Systems Science,* edited by R. A. Meyers. Springer, 2009.

[13] Santo Fortunato. Community detection in graphs. *Phys. Rep.*, 486:75–174, Jan 2010.

[14] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[15] Jörg Reichardt and Stefan Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Phys. Rev. Lett.*, 93:218701, 2004.

[16] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, 2006.

[17] Peter Ronhovde and Zohar Nussinov. Local resolution-limit-free potts model for community detection. *Phys. Rev. E*, 81:046114, Apr 2010.

[18] M. B. Hastings. Community detection as an inference problem. *Phys. Rev. E*, 74:035102(R), 2006.

[19] Jussi M. Kumpula, Mikko Kivelä, Kimmo Kaski, and Jari Saramäki. A sequential algorithm for fast clique percolation. *Phys. Rev. E*, 78:026109, 2008.

[20] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[21] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.*, 105(4):1118–1123, Jan 2008.

[22] Martin Rosvall and Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7327–7331, May 2007.

[23] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structure in large-scale networks. *Phys. Rev. E*, 76(11):036106, Sep 2007.

[24] Michael J. Barber and John W. Clark. Detecting network communities by propagating labels under constraints. *Phys. Rev. E*, 80:026129, 2009.

[25] V. Gudkov, V. Montealegre, S. Nussinov, and Z. Nussinov. Community detection in complex networks by dynamical simplex evolution. *Phys. Rev. E*, 78:016113, 2008.

[26] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda. Detecting complex network modularity by dynamical clustering. *Phys. Rev. E*, 75:045102(R), 2007.

[27] Aaron Clauset, M. E. J. Newman, and Christopher Moore. in *Proceedings of the* 23^{rd} *International Conference on Machine Learning*. Association of Computing Machinery, New York, 2006.

[28] Brian Karrer, Elizaveta Levina, and M. E. J. Newman. Robustness of community structure in networks. *Phys. Rev. E*, 77:046119, 2008.

[29] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5):056131, 2004.

[30] Roger Guimerà, Marta Sales-Pardo, and Luis A. Nunes Amaral. Module identification in bipartite and directed networks. *Phys. Rev. E*, 76:036102, September 2007.

[31] Nan Du, Bin Wu, Bai Wang, and Yi Wang. Overlapping community detection in bipartite networks. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, page 176, Amsterdam, Apr 2008. IOS Press.

[32] Overlapping nodes are shared between multiple communities.

[33] A. Arenas, A Fernández, and S. Gómez. Analysis of the structure of complex networks at different resolution levels. *New J. Phys.*, 10:053039, May 2008.

[34] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution and multi-resolution methods in complex networks. *Fluct. Noise Lett.*, 7:L209–L214, 2007.

[35] Tapio Heimo, Jussi M. Kumpula, Kimmo Kaski, and Jari Saramäki. Detecting modules in dense weighted networks with the potts method. *J. Stat. Mech.: Theory Exp.*, 8(8):P08007, 2008.

[36] Daniel J. Fenn, Mason A. Porter, Mark McDonald, Stacy Williams, Neil F. Johnson, and Nick S. Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007-2008 credit crisis. *Chaos*, 19:033119, Aug 2009.

[37] Peter Ronhovde and Zohar Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Phys. Rev. E*, 80(1):016109, Jul 2009.

[38] Jie Zhang, Kai Zhang, Xiao ke Xu, Chi K Tse, and Michael Small. Seeding the kernels in graphs: toward multi-resolution community analysis. *New J. Phys.*, 11:113003, Nov 2009.

[39] Xue-Qi Cheng and Hua-Wei Shen. Uncovering the community structure associated with the diffusion dynamics on networks. *J. Stat. Mech.: Theory Exp.*, 2010(04):P04024, Apr 2010.

[40] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnella. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328:876–878, May 2010.

[41] Marcelo Blatt, Shai Wiseman, and Eytan Domany. Superparamagnetic clustering of data. *Phys. Rev. Lett.*, 76(18):3251–3554, 1996.

[42] Hristo N. Djijev. A scalable multilevel algorithm for graph clustering and community structure detection. In *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006, Revised Papers*, volume 4936, pages 117–128, Berlin, Heidelberg, 2007. Springer-Verlag.

[43] I Ispolatov, I Mazo, and A Yuryev. Finding mesoscopic communities in sparse networks. *J. Stat. Mech.: Theory Exp.*, 09(09):P09014, 2006.

[44] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proc. Natl. Aca. Sci. U.S.A.*, 104:36–41, 2007.

[45] Jussi M. Kumpula, Jari Saramäki, Kimmo Kaski, and J. Kertész. Limited resolution in complex network community detection with potts model approach. *Euro. Phys. J. B*, 56:41–45, 2007.

[46] We adopt the term 'multiresolution' as used in [34] to indicate that this algorithm is not limited to hierarchical structures.

[47] Erzsébet Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and Albert-Lászloó Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[48] Marta Sales-Pardo, Roger Guimerà, André A. Moreira, and Luis A. Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proc. Natl. Aca. Sci. U.S.A.*, 104:15224–15229, September 2007.

[49] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[50] Erzsébet Ravasz and Albert-Lászloó Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, 2003.

[51] Luciano da F. Costa and Roberto Fernandes Silva Andrade. What are the best concentric descriptors for complex networks. *New J. Phys.*, 9:311, Sep 2007.

[52] The resolution limit causes a community division of affected methods to be roughly constrained by the graph's own global parameters. For modularity, the number of communities q scales as $\sqrt{L}$ on average [44] where L is the number of edges in the graph. This behavior is distinct from the utility of varying *graph-independent* model weights to examine system structure at different community edge densities. Ref. [33] recently applied this resolution 'limit' in a novel manner to examine the structure of a system over a range of resolutions, but the model used here is independent of this additional global constraint.

[53] The Potts model in [18] also implicitly uses a version of a missing edge penalty, but the (accurate) implementation of the model is defined with a mean-field approximation.

[54] Peter Ronhovde and Zohar Nussinov. An improved potts model applied to community detection. *e-print arXiv:0803.2548v1*, Apr 2008.

[55] V. A. Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Phys. Rev. E*, 80:036115, Sep 2009.

[56] 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. 2.4.2, the model would use $\gamma_{ER} \equiv \gamma_{RB} p = 1/2$ regardless of the system size or level of noise. See Appendix B for more discussion and Secs. 2.5.5 and 2.5.5 for differences between the models on weighted graphs.

[57] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *J. Stat. Mech.: Theory Exp.*, 9(9):P09008, 2005.

[58] Andreas Noack and Randolf Rotta. Multi-level algorithms for modularity clustering. In J. Vahrenhold, editor, *Experimental Algorithms*, volume 5526, pages 257–268. Springer-Verlag Berlin, Heidelberg, Jun 2009.

[59] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Phys. Rev. E*, 80:056117, 2009.

[60] Aaron Clauset, M. E. J. Newman, and Christopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Jun 2004.

[61] Systems were solved on AMD Opteron computers at 2.2 – 2.8 GHz with up to 48 GB of random access memory.

[62] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69(6):066133, Jun 2004.

[63] In Fig. 2.1, 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.* [163] 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. 2.4.2, for example, we use a random initial state with $q_0 \simeq q$, but the accuracy *decreases* compared to a solution beginning with $q_0 = N$.

[64] Data for the constructed system in Fig. 2.2 can be found at http://physics.wustl.edu/zohar/communitydetection/.

[65] We use a log scale for the model weight γ because of its relation to the minimum community edge density $p_{\min}$ in Eq. (2.3). 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.

[66] 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. 2.4.2, if one randomly selects successive pairs of nodes when adding new edges (such as is done in the PLOD algorithm [164]), nodes with smaller degrees will "fill up" first leaving nodes with larger degrees to be increasingly connected to each other as a group.

[67] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78:046110, 2008.

[68] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. The performance of modularity maximization in practical contexts. *e-print arXiv:0910.0165*, Oct 2009.

[69] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 101(9):2658–2663, Mar 2004.

[70] Stefanie Muff, Francesco Rao, and Amedeo Caflisch. Local modularity measure for network clusterizations. *Phys. Rev. E*, 72:056107, Nov 2005.

[71] A. Arenas, J. Duch, A. Fernández, and S. Goméz. Size reduction of complex networks preserving modularity. *New J. Phys.*, 9:176, June 2007.

[72] W. W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33:452, 1977.

[73] Additional work during each replica solution (Step 2 of the multiresolution algorithm) includes two optimizations. First, as in [54] each replica can utilize multiple trials t to mitigate the effects of local energy minima by selecting the lowest energy trial for that replica. Typically $t \simeq 4$, but it can be higher $t \simeq 20$ for difficult systems. Second, after each replica solution is completed, the algorithm merges clusters that will lower the energy. This effect arises in part because the algorithm evolves the system by moving one node at a time to a new community based on the largest energy decrease for that node on

the current iteration. Some edge configurations (particularly heavily weighted graphs with $\gamma \ll 1$) can hinder the process of merging clusters if it is performed one node at a time. The number of merges is indefinite, but it is generally small enough to not significantly alter the overall performance of the algorithm. The additional cost is estimated to be $O(NZ \log Z \log q)$ except for $\gamma \ll 1$ where q is small. These optimizations are optional in the sense that the overall algorithm will still work but the distinction between neighboring resolutions will be lower.

[74] The scaling dependence of β is a complicated function $\beta \equiv \beta(N, q, Z_{in}, Z_{out}, \gamma)$ where N is the number of nodes, q is the number of communities, γ is the Potts model weight, and Z_{in} and Z_{out} are the average number of interior and exterior connected edges for each node with respect to its own community. A typical average scaling is $\beta \lesssim 0.3$. Empirically for large graphs, worst case behavior is seen in localized regions with an intermediate to high Z_{out}/Z_{in} ratio where there is significant system confusion (where any algorithm would have difficulty). In these cases $\beta \simeq 9$, and the system is in transition to a structure that is more difficult to solve. The transition effect is localized and is not representative of the global scaling. Afterward, β drops to a value more representative of the average scaling. As the system becomes increasingly difficult, bordering on incoherent, the convergence rate can actually accelerate as the algorithm is rapidly trapped by local minima. The rapid transition is representative of general transitions between typical-easy and rare-hard problems and constitutes an analog of the

singular transition point the k-SAT problem. We will report on this effect in an upcoming publication.

[75] For the NMI, VI, and I calculations, we are generally comparing configurations that are somewhat similar. The confusion matrix can have at most N non-zero entries; therefore, we use a pseudo-sparse matrix representation to calculate these measures usually in $O(N)$. Worst case (pathological) behavior for initialization is $O(Nq)$. The calculation cost can be as fast as $O(q)$ for strongly correlated systems. The total cost for Step 5 of the algorithm will typically scale as $O(Nr^2 \log N)$ where the user has control over the number of replicas r and $\log N$ is the estimated scaling for the required number of resolutions.

[76] Ana L. N. Fred and Anil K. Jain. Robust data clustering. *2003 Proc. IEEE Comp. Soc. Conf. on Comp. Vis. Pattern Recog.*, 2:128–33, 2003.

[77] We can eliminate the scale factor of $\log Z$ for any unweighted graph. However, here we wish to demonstrate the full weighted scaling, so almost all of our unweighted graph examples include the $\log Z$ scale factor.

[78] Smaller systems were solved on a single processor of an Intel Core 2 Quad Q6600 at 2.4 GHz with 2GB RAM. Trials in excess of $O(10^8)$ edges were solved on a single processor of an 8x Opteron 8220 at 2.8GHz with 32GB RAM.

[79] We plot the *average* q as opposed to q based on an optimal partition as used in [33] in order to be consistent with our use of averaged information measures.

[80] David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten, and Steve M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav. Ecol. and Sociobiol.*, 54:396–405, 2003.

[81] David Lusseau. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Proc. R. Soc. London, Ser. B (Suppl.)*, 270:S186, 2003.

[82] David Lusseau. Evidence for social role in a dolphin social network. *Evol. Ecol.*, 21(3):357–366, 2007.

[83] K. Read. Cultures of the central highlands, new guinea. *Southwest. J. of Anthrop.*, 10:1–43, 1954.

[84] Per Hage and Frank Harary. *Structural Models in Anthropology*, pages 40, 56–60. Cambridge University Press, 1983.

[85] Different internal community densities are best resolved at different values of γ (*i.e.*, different *resolutions*). As the number of nodes N decreases in the benchmark [67], the density of inter-community edges between two communities increases relative to the internal community densities. Together, they cause the range of the solvable resolutions (see the solution at (*i*a,b) in Fig. 3.11) to decrease and eventually restrict an accurate or intended solution.

[86] Marina Meilă. Comparing clusterings — an information based distance. *J. Multivariate Anal.*, 98:873, 2007.

[87] M. Mézard, G. Parisi, and R. Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297:812–815, Aug 2002.

[88] Florent Krzakala and Lenka Zdeborová. Phase transitions and computational difficulty in random constraint satisfaction problems. *Journal of Physics: Conference Series*, 95(1):012012, 2008.

[89] Rémi Monasson, Riccardo Zecchina, Scott Kirkpatrick, Bart Selman, and Lidror Troyansky. Determining computational complexity from characteristic 'phase transitions'. *Nature*, 400:133–137, Jul 1999.

[90] Tad Hogg, Bernardo A. Huberman, and Colin P. Williams. Phase transitions and the search problem. *Artificial Intelligence*, 81(1-2):1–15, Mar 1996.

[91] M. Bayati, C. Borgs, A. Braunstein, J. Chayes, A. Ramezanpour, and R. Zecchina. Statistical mechanics of steiner trees. *Phys. Rev. Lett.*, 101(3):037208, Jul 2008.

[92] Jie Zhou and Haijun Zhou. Ground-state entropy of the random vertex-cover problem. *Phys. Rev. E*, 79(2):020103, Feb 2009.

[93] Peter Cheeseman, Bob Kanefsky, and William M. Taylor. Where the *really* hard problems are. In *Conference On Artificial Intelligence archive. Proceedings of*

the 12th international joint conference on Artificial Intelligence, volume 1, pages 331–337, 1991.

[94] Hidetoshi Nishimori and K. Y. Michael Wong. Statistical mechanics of image restoration and error-correcting codes. *Phys. Rev. E*, 60(1):132–144, Jul 1999.

[95] J. Reichardt. Structure in complex networks. In *Experimental Algorithms*, volume 766. Springer-Verlag Berlin, Heidelberg, 2009.

[96] Dandan Hu, Peter Ronhovde, and Zohar Nussinov. Phase transition and memory effects in the community detection problem. *(in preparation)*, 2010.

[97] K. Jonason, E. Vincent, J. Hammann, J.-P. Bouchaud, and P. Nordblad. Memory and chaos effects in spin glasses. *Phys. Rev. Lett.*, 81(15):3243–3246, Oct 1998.

[98] K. Jonason, P. Nordblad, E. Vincent, J. Hammann, and J.-P. Bouchaud. Memory interference effects in spin glasses. *Eur. Phys. J. B*, 13(1):99–105, 2000.

[99] David V. Foster, Jacob G. Foster, Maya Paczuski, and Peter Grassberger. Clustering phase transitions and hysteresis: Pitfalls in constructing network ensembles. *e-print arXiv:0911.2055*, Nov 2009.

[100] D. A. Stariolo, M. A. Montemurro, and F. A. Tamarit. Aging dynamics of $\mathsf{\pmj}$ edwards-anderson spin glasses. *Eur. Phys. J. B*, 32(3):361–367, Apr 2003.

[101] L. Berthier and A. P. Young. Aging dynamics of the heisenberg spin glass. *Phys. Rev. B*, 69(18):184423, May 2004.

[102] Richard Zallen. *The Physics of Amorphous Solids*, pages 23–32. John Wiley & Sons, Inc., 1983.

[103] A. L. Greer and E. Ma. Bulk metallic glasses: At the cutting edge of metals research. *MRS Bulletin*, 32:611–619, 2007.

[104] Bruno C. Hancock and Michael Parks. What is the true solubility advantage of amorphous pharmaceuticals. *Pharmaceutical Research*, 17(4):397–404, 2000.

[105] Mark Telford. The case for bulk metallic glass. *Materials Today*, 7(3):36–43, March 2004.

[106] T. Schenk, D. Holland-Moritz, V. Simonet, R. Bellissent, and D. M. Herlach. Icosahedral short-range order in deeply undercooled metallic melts. *Phys. Rev. Lett.*, 89(7):075507, Aug 2002.

[107] K. F. Kelton, G. W. Lee, A. K. Gangopadhyay, R. W. Hyers, T. J. Rathz, J. R. Rogers, M. B. Robinson, and D. S. Robinson. First x-ray scattering studies on electrostatically levitated metallic liquids: Demonstrated influence of local icosahedral order on the nucleation barrier. *Phys. Rev. Lett.*, 90(19):195504, May 2003.

[108] David R. Nelson. *Geometrical Frustration*. Cambridge University Press, Cambridge, 1999.

[109] J. F. Sadoc and R. Mosseri. *Geometrical Frustration*. Cambridge University Press, Cambridge, 1999.

[110] G. Tarjus, S. A. Kivelson, Z. Nussinov, and P. Viot. The frustration-based approach of supercooled liquids and the glass transition: a review and critical assessment. *J Phys.: Condens. Matter*, 17:R1143–R1182, 2005.

[111] Zohar Nussinov. Avoided phase transitions and glassy dynamics in geometrically frustrated systems and non-abelian theories. *Phys. Rev. B*, 69:014208, 2004.

[112] Vassiliy Lubchenko and Peter G. Wolynes. Theory of structural glasses and supercooled liquids. *Annu. Rev. Phys. Chem.*, 58:235–266, 2007.

[113] Hajime Tanaka, Takeshi Kawasaki, Hiroshi Shintani, and Keiji Watanabe. Critical-like behaviour of glass-forming liquids. *Nature Materials*, 9:324–331, Apr 2010.

[114] L. Angelani, Giorgio Parisi, G. Ruocco, and G. Viliani. Potential energy landscape and long-time dynamics in a simple model glass. *Phys. Rev. E*, 61(2):1681–1691, Feb 2000.

[115] Giorgio Parisi. The physics of the glass transition. *Physica A*, 280:115–124, Apr 2000.

[116] Srikanth Sastry, Pablo G. Debenedetti, and Frank H. Stillinger. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature*, 393:554–557, 1998.

[117] Pablo G. Debenedetti and Frank H. Stillinger. Supercooled liquids and the glass transition. *Nature*, 410:259–267, Mar 2001.

[118] Vassiliy Lubchenko and Peter G. Wolynes. Theory of aging in structural glasses. *J. Chem. Phys.*, 121(7):2852–2865, Aug 2004.

[119] William L. Johnson, Marios D. Demetriou, John S. Harmon, Mary L. Lind, and Konrad Samwer. Rheology and ultrasonic properties of metallic glass-forming liquids: A potential energy landscape perspective. *MRS Bulletin*, 32:644–650, Aug 2007.

[120] Jonathan P. K. Doye, David J. Wales, Fredrik H. M. Zetterling, and Mikhail Dzugutov. The favored cluster structures of model glass formers. *J. Chem. Phys.*, 118(6):2792–2799, 2008.

[121] F. Ritort and P. Sollich. Glassy dynamics of kinetically constrained models. *Adv. Phys.*, 52(4):219–342, 2003.

[122] V. Cvetkovic, Z. Nussinov, and J. Zaanen. Topological kinematic constraints: dislocations and the glide principle. *Philosophical Magazine*, 86(20):2995–3020, Jul 2006.

[123] E. Aharonov, E. Bouchbinder, H. G. E. Hentschel, V. Ilyin, N. Makedonska, I. Procaccia, and N. Schupper. Direct identification of the glass transition: Growing length scale and the onset of plasticity. *Euro. Phys. Lett.*, 77:56002, 2007.

[124] Andrea Montanari and Guilhem Semerjian. Rigorous inequalities between length and time scales in glassy systems. *J. Stat. Phys.*, 125(1):23–54, Oct 2006.

[125] Majid Mosayebi, Emanuela Del Gado, Patrick Ilg, and Hans Christian Ottinger. Probing a critical length scale at the glass transition. *Phys. Rev. Lett.*, 104:205704, May 2010.

[126] Jorge Kurchan and Dov Levine. Correlation length for amorphous systems. *e-print arXiv:0904.4850*, Apr 2009.

[127] Ludovic Berthier, Giulio Biroli, Jean-Philippe Bouchaud, L. Cipelletti, D. El Masri, D. L'Hôte, F. Ladieu, and M. Pierno. Direct experimental evidence of a growing length scale accompanying the glass transition. *Science*, 310(5755):1797–1800, Dec 2005.

[128] Smarajit Karmakar, Chandan Dasgupta, and Srikanth Sastry. Growing length and time scales in a glass-forming liquid. *Proc. Natl. Acad. Sci. U.S.A.*, 106(10):3675–3679, Mar 2010.

[129] Giulio Biroli, Jean-Philippe Bouchaud, Andrea Cavagna, Tomás S. Grigera, and Paolo Verrocchio. Thermodynamic signature of growing amorphous order in glass-forming liquids. *Nature Physics Letters*, 4:771–775, Oct 2008.

[130] Giulio Biroli and Jean-Philippe Bouchaud. The random first-order transition theory of glasses: a critical assessment. *e-print arXiv:0912.2542*, Dec 2009.

[131] J. D. Bernal. Geometry of the structure of monatomic liquids. *Nature*, 185:68–70, 1960.

[132] D. B. Miracle, W. S. Sanders, and O. N. Senkov. The influence of efficient atomic packing on the constitution of metallic glasses. *Philos. Mag.*, 83(20):2409–2428, Jul 2003.

[133] Daniel B. Miracle. A structural model for metallic glasses. *Nature Materials*, 3:697–702, Oct 2004.

[134] Daniel B. Miracle, Takeshi Egami, Katharine M. Flores, and Kenneth F. Kelton. Structural aspects of metallic glasses. *MRS Bulletin*, 32:629–634, Aug 2007.

[135] Daniel B. Miracle, Eric A. Lord, and Srinivasa Ranganathan. Candidate atomic cluster configurations in metallic glass structures. *Materials Transactions*, 47(7):1737–1742, Jul 2006.

[136] W. K. Luo, H. W. Sheng, F. M. Alamgir, J. M. Bai, J. H. He, and E. Ma. Icosahedral short-range order in amorphous alloys. *Phys. Rev. Lett.*, 92(14):145502, Apr 2004.

[137] P. Ganesh and M. Widom. Ab initio simulations of geometrical frustration in supercooled liquid fe and fe-based metallic glass. *Phys. Rev. B*, 77:014205, Jan 2008.

[138] Y. T. Shen, T. H. Kim, A. K. Gangopadhyay, and K. F. Kelton. Icosahedral order, frustration, and the glass transition: Evidence from time-dependent nucleation and supercooled liquid structure studies. *Phys. Rev. Lett.*, 102:057801, Feb 2009.

[139] Emanuela Del Gado, Patrick Ilg, Martin Kröger, and Hans Christian Öttinger. Nonaffine deformation of inherent structure as a static signature of cooperativity in supercooled liquids. *Phys. Rev. Lett.*, 101:095501, Aug 2008.

[140] Walter Kob, Claudio Donati, Steven J. Plimpton, Peter H. Poole, and Sharon C. Glotzer. Dynamical heterogeneities in a supercooled lennard-jones liquid. *Phys. Rev. Lett.*, 79(15):2827–2830, 1997.

[141] Eric R. Weeks, J. C. Crocker, Andrew C. Levitt, Andrew Schofield, and D. A. Weitz. Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. *Science*, 287(5453):627–631, Jan 2000.

[142] Asaph Widmer-Cooper, Peter Harrowell, and H. Fynewever. How reproducible are dynamic heterogeneities in a supercooled liquid? *Phys. Rev. Lett.*, 93(13):135701, Sep 2004.

[143] Jacob D. Stevenson, Jörg Schmalian, and Peter G. Woylnes. The shapes of cooperatively rearranging regions in glass-forming liquids. *Nature Physics*, 2:268–274, Apr 2006.

[144] H. W. Sheng, W. K. Luo, F. M. Alamgir, J. M. Bai, and E. Ma. Atomic packing and short-to-medium-range order in metallic glasses. *Nature*, 439:419–425, Jan 2006.

[145] J. L. Finney. Random packings and the structure of simple liquids. i. the geometry of random close packing. *Proc. R. Soc. London, Ser. A*, 319(1539):479–493, Nov 1970.

[146] J. Dana Honeycutt and Hans C. Andersen. Molecular dynamics study of melting and freezing of small lennard-jones clusters. *J. Phys. Chem.*, 91(19):4950–4963, Sep 1987.

[147] Paul J. Steinhardt, David R. Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Phys. Rev. B*, 28(2):784–805, 1983.

[148] T. C. Hufnagel and S. Brennan. Short-and medium-range order in $(Zr_{70}Cu_{20}Ni_{10})_{90-x}Ta_xAl_{10}$ bulk amorphous alloys. *Phys. Rev. B*, 67:014203, 2003.

[149] M. M. J. Treacy, J. M. Gibson, L. Fan, D. J. Paterson, and I. McNulty. Fluctuation microscopy: a probe of medium range order. *Rep. Prog. Phys.*, 68:2899–2944, 2005.

[150] Marc Mézard and Andrea Montanari. Reconstruction on trees and spin glass transition. *J. Stat. Phys.*, 124(6):1317–1350, Sep 2006.

[151] S. Franz and Andrea Montanari. Analytic determination of dynamical and mosaic length scales in a kac glass model. *J. Phys. A: Math. Theor.*, 40(11):F251–F257, Feb 2007.

[152] J. Stadler, R. Mikulla, and H. R. Trebin. Imd: A software package for molecular dynamics studies on parallel computers. *International Journal of Modern Physics C*, 8(5):1131–1140, Jun 1997.

[153] K. K. Sahu, N. A. Mauro, L. Longstreth-Spoor, D. Saha, Z. Nussinov, M. K. Miller, and K. F. Kelton. Phase separation mediated devitrification of $al_88y_7fe_5$ glasses. *Acta Materialia*, 58(12):4199–4206, Jul 2010.

[154] T. Egami. Atomistic mechanism of bulk metallic glass formation. *J. Non-Cryst. Solids*, 317:30–33, 2003.

[155] John A. Moriarty and Mike Widom. First-principles interatomic potentials for transition-metal aluminides: Theory and trends across the 3*d* series. *Phys. Rev. B*, 56(13):7905–7917, Oct 1997.

[156] Marek Mihalkovič, C. L. Henley, M. Widom, and P. Ganesh. Empirical oscillating potentials for alloys from ab-initio fits. *e-print arXiv:cond-mat.mtrl-sci/0802.2926*, Feb 2008.

[157] VASP website: http://cms.mpi.univie.ac.at/vasp/.

[158] Walter Kob and Hans C. Andersen. Testing mode-coupling theory for a supercooled binary lennard-jones mixture: The van hove correlation function. *Phys. Rev. E*, 51(5):4626–4641, 1995.

[159] L.-C. Valdes, F. Affouard, M. Descamps, and J. Habasaki. Mixing effects in glass-forming lennard-jones mixtures. *J. Chem. Phys.*, 130:154505, 2009.

[160] Dandan Hu, Peter Ronhovde, and Zohar Nussinov. Phase transition in the community detection problem: spin-glass type and dynamic perspectives. *e-print arXiv:1008.2699*, Aug 2010.

[161] Peter Ronhovde, Saurish Chakrabarty, Mousumi Sahu, Kisor K. Sahu, Kenneth F. Kelton, and Zohar Nussinov. Detecting hidden spatial and spatio-temporal structures in glasses and complex systems by multiresolution network clustering. *(in preparation)*, 2010.

[162] J. Zaanen, Z. Nussinov, and S. I. Mukhin. Duality in $2+1d$ quantum elasticity: superconductivity and quantum nematic order. *Annals of Physics*, 310(1):181–260, Mar 2004.

[163] Sanjeev Chauhan, Michelle Girvan, and Edward Ott. Spectral properties of networks with community structure. *Phys. Rev. E*, 80:056114, Nov 2009.

[164] Christopher R. Palmer and J. Gregory Steffan. Generating network topologies that obey power laws. In *Global Telecommunications Conference, 2000. GLOBECOM '00*, volume 1, pages 434–438, San Francisco, CA, 2000. IEEE.

[165] V. Gudkov and V. Montealegre. Analysis of networks using generalized mutual entropies. *Physica A*, 387(2):2620–2630, Jan 2008.

[166] As a specific realization, consider a symmetric adjacency matrix A for a four node graph given by $A_{12} = A_{13} = A_{14} = A_{23} = A_{24} = 1$. The permutation P_{34} that exchanges 3 with 4 leaves A invariant, not invariant under the permutation P_{14}. permutation of i with j invariant. If we have two solutions B and C that place i and j in different communities then $I_N(B, C) \neq 1$ despite the invariance of A under the permutation. Unless one can distinguish between the nodes via external means, community groupings of the four nodes such as (123)(4) and (124)(3) correspond to the same breaking of the lattice (the graph is symmetric under the permutation), but they have a relative mutual information that differs from the self mutual information [$(8 - 3\log 3)/4$ vs $(10 - 6\log 3)/4$ for the two groupings respectively].

Appendix

A Information theory measures

The normalized mutual information I_N and the variation of information V provide methods of comparing one proposed community division to another. In order to define $I_N(A, B)$ or $V(A, B)$ between two partitions A and B, we first ascribe a Shannon entropy $H(A)$ for an arbitrary community partition A. We assign the probability that a given node will fall in community k as $P(k) = n_k/N$, where n_k is the number of nodes in community k and N is the total number of nodes in the system. Then the Shannon entropy is

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

where q_A is the number of communities in partition A.

Mutual information $I(A, B)$ was developed within information theory. It evaluates how similar two data sets are in terms of information contained in both sets of data. The mutual information between two partitions A and B of a graph is calculated by defining a "confusion matrix" for the two community partitions. The confusion matrix specifies how many nodes n_{ij} of community i of partition A are in community

j of partition B. Mutual information $I(A, B)$ is defined as

$$I(A, B) = \sum_{i=1}^{q_A} \sum_{j=1}^{q_B} \frac{n_{ij}}{N} \log \left(\frac{n_{ij} N}{n_i n_j} \right) \tag{A-2}$$

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. An interesting generalized mutual information is also defined in [165]. Danon *et al.* [57] suggested that a normalized variant [76] of mutual information be adapted for use in evaluating similar community partitions. Using Eqs. (A-1) and (A-2), the normalized mutual information $I_N(A, B)$ between partitions A and B is

$$I_N(A, B) = \frac{2I(A, B)}{H(A) + H(B)} \tag{A-3}$$

which can take values in the range $0 \leq I_N(A, B) \leq 1$. Fred and Jain [76] introduced, for computer vision problems, a single resolution application of NMI that we use in our work.

The variation of information [86] is a metric in the formal sense of the term and measures the "distance" in information between two partitions A and B. Using Eqs. (A-1) and (A-2), $V(A, B)$ is calculated by

$$V(A, B) = H(A) + H(B) - 2I(A, B). \tag{A-4}$$

As an information distance, low values of $V(A, B)$ indicate better agreement between partitions A and B. VI has a range $0 \leq V(A, B) \leq \log N$. It is sufficient and even preferable to use the un-normalized version of VI. We utilize both NMI and VI to demonstrate that our approach is not limited to a specific measure.

The mutual information I and Shannon entropy H also play a supplemental role in determining multiresolution structure. For the Shannon entropy H, we average over all replicas using

$$\langle H \rangle = \frac{1}{r} \sum_{A} H(A). \tag{A-5}$$

For I_N, V, and I, we calculate the average of the measures over all pairs of replicas with

$$\langle S \rangle = \frac{2}{r(r-1)} \sum_{A>B} S(A,B) \tag{A-6}$$

where S is any of the information measures and r is the number of replicas. We use base 2 logarithms in all information calculations.

Similarly, higher order cumulants of S can be computed with a (replica symmetrically weighted) probability distribution function that we set to be

$$P(S) = \frac{2}{r(r-1)} \sum_{A>B} \delta\big[S - S(A,B)\big]. \tag{A-7}$$

In Eq. (A-7), $\delta\big[S - S(A,B)\big]$ is the Dirac delta function. For any function f of S, the expectation value of f is

$$\langle f \rangle = \int dS \; P(S) f(S). \tag{A-8}$$

Formally, in our probability distribution of Eq. (A-7), the information measure S plays a role analogous to the overlap parameter in spin-glass problems.

B Resolution limit and the Erdős-Rényi Potts model

For unweighted graphs, the RBER model of Eq. (2.21), 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 [15] 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. 2.3) that supported the approximation of $\gamma_{ER} \propto p$. However, the approximation is not generally applicable. For example, between the systems in Secs. 2.3, 2.4, and 2.6, 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. 2.4.2) 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. (2.21) 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. 2.5.1), but it also appeared to impose a resolution

limit [45]. 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 [15] without the density dependence p.

The second term in Eq. (2.21) 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 [15]), 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. 2.5.2 and 2.5.3. Using Fig. 2.8(a), the original condition for two arbitrary unweighted communities A and B to merge is given by Eq. (2.14). Without p, the new merge condition is

$$l > \gamma_{ER} nm \tag{B-1}$$

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. 2.7, the original merge condition is given by Eq. (2.10). The new condition for two neighboring cliques to merge is

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

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. (B-2) 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. 2.5.1 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. 2.5.5).

C Example noise test solution with the RBCM

In Sec. 2.4.2, we add noise to a strongly defined system to test the accuracy of the RBCM of Eq. (2.7) compared to the APM of Eq. (2.2). A sample system is depicted in Fig. 2.4, and the accuracy results are summarized in Figs. 2.5 and 2.6. For the APM, we solve the system with the model weight $\gamma = 1$ for all graphs. Figure C1 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

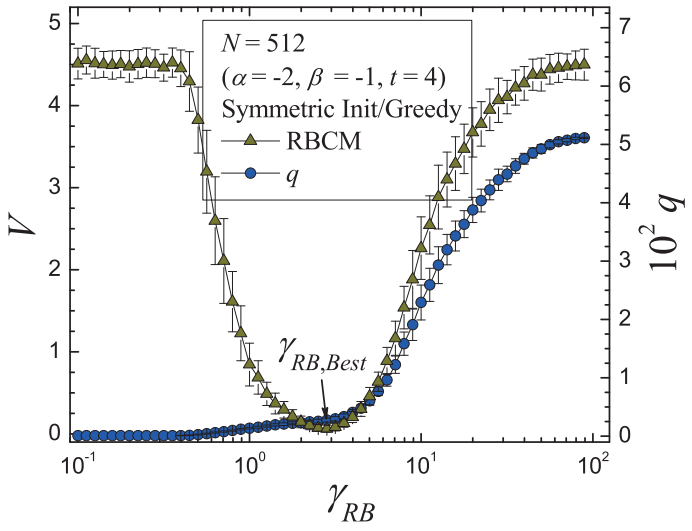


Figure C1: Plot of VI V vs γ_{RB} for the RBCM. In Sec. 2.4.2, we generate a set of strongly defined communities with varying levels of intercommunity noise k_{out}. Using the greedy algorithm in Sec. 2.2, we compare the accuracy of solutions found with the RBCM to our APM. Since the models operate differently, we compare the results in Figs. 2.5 and 2.6 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. 2.4) 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.

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 C1 shows only one *best* answer for the RBCM. We average over 100 graphs for each γ_{RB}, and the best VI average is plotted

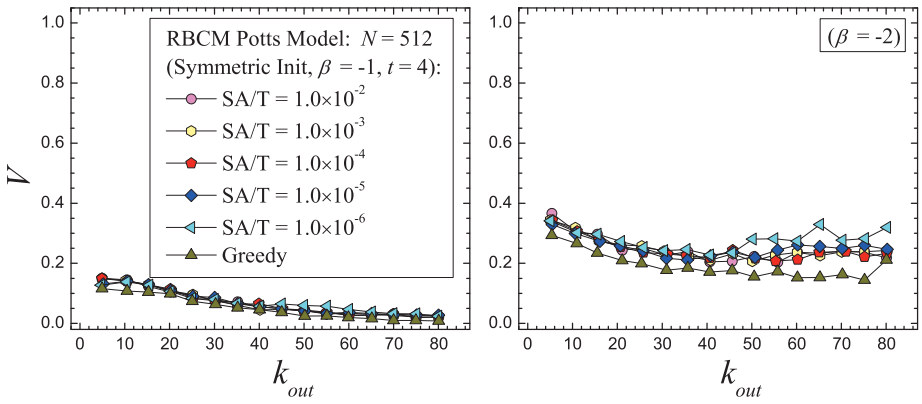


Figure D1: Plot of VI V vs the average external degree k_{out} for the RBCM. In Sec. 2.4.2, we generate a set of strongly defined communities with high levels of intercommunity noise. In Figs. 2.5 and 2.6, we vary the initial average power-law degree $\langle k \rangle_\alpha \simeq k_{out}$ and solve the networks using the greedy algorithm in Sec. 2.2 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.

in Figs. 2.5 and 2.6 with the result for the APM.

D Noise test analysis of SA at different starting temperatures

In Sec. 2.4.2, 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. 2.4.2). We solve each system using the RBCM with the greedy algorithm in Sec. 2.2 and SA both with $t = 4$ trials. In Figs. 2.5 and 2.6, 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. 2.2 using a symmetric initial state $(q_0 = N)$.

Given this counter-intuitive result, in Fig. D1 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 $T_{i+1} = 0.999T_i$ 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.

E Generalization of the information-based replica method

In Sec. 2.2, we may recast the information theory measures used to evaluate the correlation between different replicas for other (non-graph theoretic) optimization problems with general Hamiltonians (or cost functions) $\mathcal{H}$. An alternate form of Eq. (A-2) for the mutual information between replicas i and j is

$$I(i,j) = H(i) + H(j) - H(i,j) \tag{E-1}$$

where $H(i)$, $H(j)$, and $H(i,j)$ denote the entropies of replica i, replica j, and the combined system formed by both replicas, respectively. Instead of using Eq. (A-2), we write the Shannon entropy $H(i,j)$ for the combined replicas i and j which we then apply in Eq. (E-1). For general Hamiltonians $\mathcal{H}$, we replace $H(i)$, $H(j)$, and $H(i,j)$ by a thermodynamic entropy for the respective systems.

In the general case, the thermodynamic entropy $H(i,j)$ of the system formed by the union of replicas i and j is

$$H(i,j) = \frac{\partial}{\partial T}\left\{\beta^{-1}\log\left[\mathrm{Tr}_{i,j}\left(e^{-\beta\mathcal{H}(i)} + e^{-\beta\mathcal{H}(j)}\right)\right]\right\}, \tag{E-2}$$

and the entropy $H(i)$ of system i or j is

$$H(i) = \frac{\partial}{\partial T}\left\{\beta^{-1}\log\left[\mathrm{Tr}_{i}\left(e^{-\beta\mathcal{H}(i)}\right)\right]\right\}. \tag{E-3}$$

$\mathcal{H}(i)$ and $\mathcal{H}(j)$ are the Hamiltonians of replicas i and j, and $\beta = 1/(T\ln 2)$ is the inverse temperature. Within our approach, an ensemble reduces to a finite number of

points (replicas) whose correlations are monitored by information theory measures. This form pertains to the general case in which both i and j pertain to a collection of decoupled copies, and the traces are over all coordinates in replicas i and j.

The standard mutual information of Eq. (A-2) is generally not invariant (as it ideally should be) under the permutation of "identical" nodes (those with an identical neighbor list that are otherwise indistinguishable by other parameters of the system). Specifically, we refer to nodes i and j as identical in a graph if the adjacency matrix A is invariant under the permutation of node i with node j [166]. That is, A commutes with the permutation of nodes i and j, $[P_{ij}, A] = 0$, if nodes i and j are identical. The thermodynamic entropies of Eqs. (E-2) and (E-3) are invariant under permutations of identical nodes because any symmetries, or lack thereof, are fully represented in the system Hamiltonian $\mathcal{H}$.

In the simplest case with only one copy of the system in replica i and one copy in replica j, there is only one term in both i and j; and the designation $\mathrm{Tr}_{i,j}$ becomes redundant (the entropies of i and j are also trivially $H(i) = H(j) = 0$). In a more realistic approximation to thermodynamic quantities, each of the replicas i and j contain a number of independent decoupled copies of the system. Inserting Eqs. (E-2) and (E-3) into Eq. (E-1), we obtain the mutual information between i and j. NMI and VI are then given by Eqs. (A-3) and (A-4). Other information measures $S(i, j)$ between replicas i and j may also be computed. Along similar lines, multi-replica (higher than two) forms may replace the sum over two-replica configurations in Eqs. (3.1) and (E-2).

We may also reconstruct the information measures using a different physical analogy. The Shannon entropy of Eq. (A-1) is analogous to an ensemble where each of the N nodes corresponds to one point in the ensemble. The communities correspond to q possible states of a single particle with energies $\{E_k\}$ for $k = 1$ to q at a given temperature T such that the same community occupation probabilities are reproduced as $p_k = n_k/N = e^{-\beta E_k}/\sum_{i=1}^{q} e^{-\beta E_i}$ where the inverse temperature is $\beta = 1/(T \ln 2)$. The mutual information I of Eq. (A-2) is equivalent to an ensemble of size N for a two-particle system in which each particle can be in any of q states. The interaction between the two particles is such that it leads to energies $\{E_{ij}\}$ for the two occupied communities i and j. These interactions lead to a relative probability $p_{ij} = n_{ij}/N$ for occupying the two-particle states that is proportional to $e^{-\beta E_{ij}}$. The effective Hamiltonian for the resulting physical system does not directly depend on the identities of the N nodes (although it does not distinguish between "identical" and distinguishable nodes).

One potential limitation of our thermodynamic framework in Eqs. (E-2) and (E-3) is that general, non-graph theoretic, applications may require many copies of the same system. The traces Tr_i, Tr_j need to be calculated on multiple copies of the same system. This is bypassed in the application of mutual information for graph problems because the node number N effectively plays the role of many ensemble points (multiple replica copies) on which the thermodynamic average is to be taken.

F Multiresolution LFR benchmark comments

As discussed in Sec. 3.3, we used the new benchmark problem presented in [67] to test the accuracy of our multiresolution algorithm of Sec. 2.2. Our algorithm attempts to identify all strongly defined resolutions. By design, the benchmark in [67] constructs a "realistic" system with a single intended solution; however, random effects of the graph generation process can also create additional transient, but nevertheless strongly defined resolutions which our algorithm can detect. In implementing the benchmark, we endeavor to automate the identification process to determine the single "best" resolution as intended by the benchmark. We explain two special cases.

The first difficulty is encountered for $\mu \lesssim 0.4$. We can detect multiple resolutions with perfect correlations ($I_N = 1$ and $V = 0$) for resolutions near the intended benchmark solution which occur more frequently as μ decreases. This effect is similar to partition (i) that occurred near partition (ii) in Fig. 3.9. The transitional resolutions are not necessarily meaningless partitions on an individual graph-by-graph basis, but they are artifacts of the randomly generated system and thus vary across the different benchmark graphs. Similar to structure (iia,b) in Fig. F1, the plateaus in the information measures H or I (or the number of clusters q [33]) indicate a more "stable" partition. It is this stable partition that corresponds to the intended solution for the benchmark graph. Thus, when necessary, we use the plateaus to discriminate between the short-lived and the most stable strongly defined partitions in order to determine the single "best" resolution for each benchmark graph.

A second difficulty is shown in Fig. F1 which occurs most frequently in the range of mixing parameter $0.45 \lesssim \mu \lesssim 0.65$. The stable configuration that corresponds to the intended benchmark answer is configuration (*ii*a,b). The low-density, transient, but strongly correlated configuration at (*i*a,b) has a slightly higher NMI correlation. Even a casual visual inspection of the data in Fig. F1 indicates that configuration (*ii*a,b) is the dominant configuration for the graph. Specifically, configuration (*ii*a,b) possesses both very strong NMI and VI correlations ($I_N \simeq 1.0$ and $V \simeq 0.0$) as well as stable and long H, I, and q plateaus, and indeed it corresponds almost exactly to the intended benchmark answer. However, the automated application of the multiresolution algorithm slightly favors configuration (*i*a,b) as the "best" resolution since it has a higher NMI ($\delta I_N \simeq 6.3 \times 10^{-5}$) and a lower VI. (See Secs. 3.2.2 and 3.4 regarding potential problems of using the plateaus in H, I, or q as the primary measure for identifying the "best" resolutions.)

These graphs are the cause of the accuracy perturbations in Figs. 3.12(a) and 3.12(b). They are less frequent for $\beta = 2$ since the community size distribution is more skewed towards smaller communities than for $\beta = 1$. We note that the average accuracy for the perturbations in Figs. 3.12(a) and 3.12(b) is still high at $I_N \simeq 0.96$. In Fig. 3.12, an iteration cap acted as an effective filter for most low-density spikes. We could further improve the automated analyses of such graphs by replacing this filter with moving NMI or VI averages (*i.e.*, each moving average is over the NMI or VI of several nearby resolutions) to exclude resolutions such as the short-lived configuration (*i*).

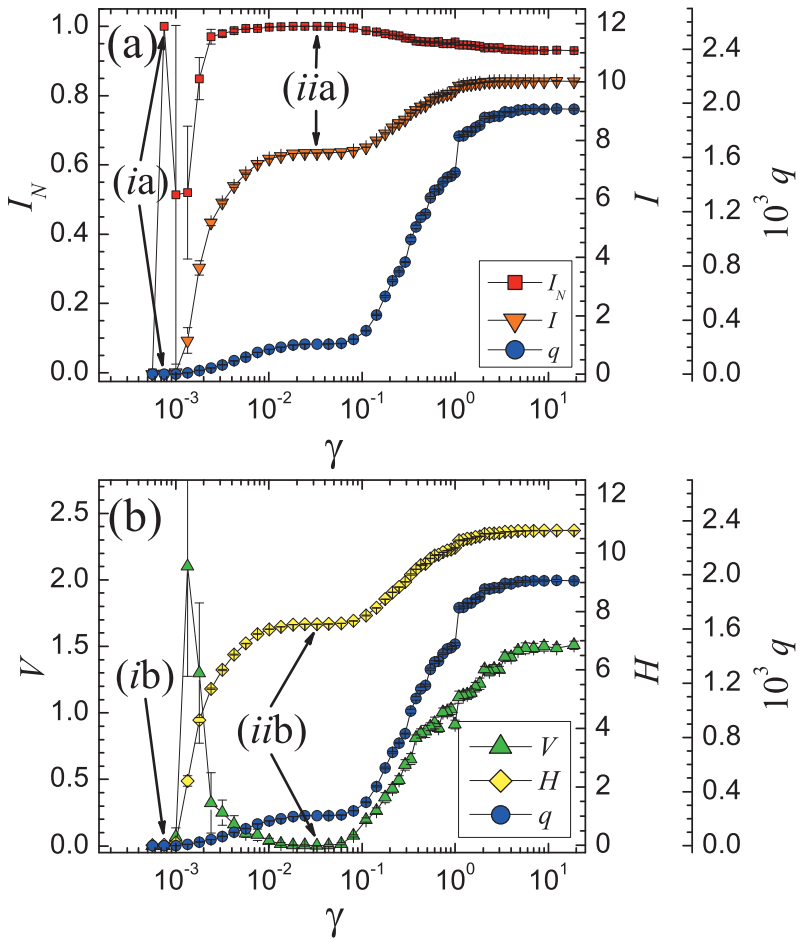


Figure F1: Plot of information measures I_N, V, H, and I, and q (right-offset axes) vs the Potts model weight γ for a realization of the benchmark [67] (see Sec. 3.3). See Fig. 2.3 for a description of the legends and axes. This plot uses $N = 5000$, $\mu = 0.45$, $\alpha = 2$, and $\beta = 1$. We use the algorithm in Sec. 2.2 to attempt to identify the "best" resolution for the graph. For some cases in the benchmark, there exists more than one extremal value of I_N and V where the low-density configuration at (i) has a slightly stronger NMI and VI correlations than the intended benchmark answer at (ii). In this example, a casual inspection indicates that the stable region at (ii) is clearly the "best" partition which corresponds almost exactly to the intended solution. The automated algorithm favors the configuration at (i).

G Overlapping dynamics

In Chapter 5, we wish to account for the possibility of a given atom being connected to more than one physical cluster. For example, in a cubic lattice, each atom participates in the local structure of multiple unit cells. In community detection, this corresponds to allowing "overlapping" community memberships where a node can be a member of more than one community. To accomplish this task, we select the lowest energy replica partition at the best resolution(s) of the model network [*i.e.*, value(s) of γ in Eq. (2.2) corresponding to extrema in I_N or V].

First, we fix the initial node memberships including the number of communities q. We then sequentially iterate through the nodes and each membership for a given node and make changes according to the following: (1) place the node in any additional (non-member) clusters to which it is bound (a negative energy contribution), or (2) remove the node from any member clusters (except for the original membership) in which the current net energy contribution is positive. This process iterates through all nodes as many times as necessary until no node additions or removals are found. The total computational cost is slightly higher than the initial partitioning cost in Sec. 2.2 due to the multiple allowed memberships. See also [2] for another method that allows overlapping multiscale network analysis in a general fashion.

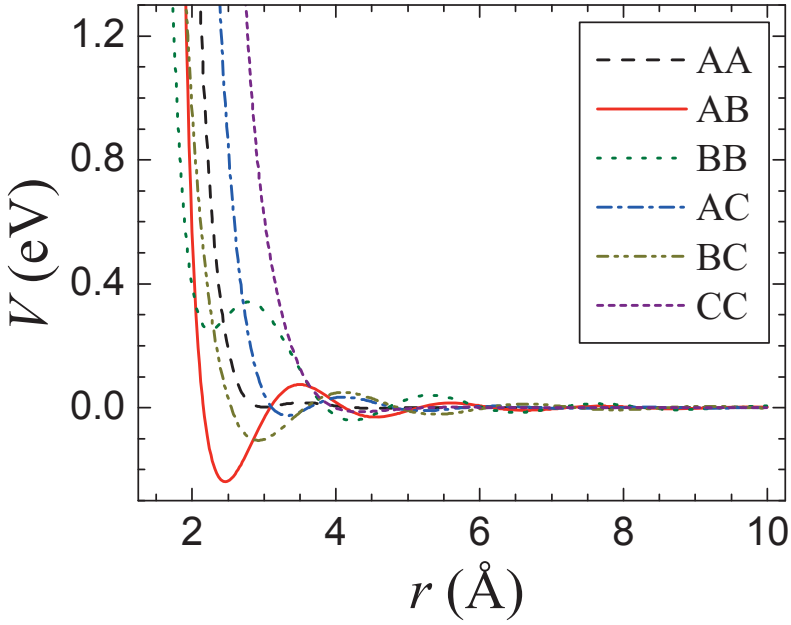


Figure H1: A plot of the model potentials for our three-component model glass former using the fit data in Table H1 in Sec. 5.3.1. We indicate the three atomic types by "A", "B", and "C". The units are given for a specific candidate atomic realization (AlYFe). Here, we use the *ab initio* fit data in Table 5.1 in place of the same-species GPT interactions used in Table 5.1 and Fig. 5.2.

	a_0	a_1	a_2	a_3	a_4	a_5
AA	2.11*	9.49*	-32.3*	3.66*	-10.6*	6.20*
AB	1.92	17.4	6.09	3.05	-4.68	3.48
AC	2.38	8.96	-14.9	3.11	-3.88	4.38
BB	2.01*	4.95*	5.01*	2.74*	-2.26*	3.00*
BC	1.88	8.00	-3.42	2.53	-1.25	3.00
CC	2.75*	15.3*	-6400*	2.38*	-4.69*	8.71*

Table H1: Fit parameters for Eq. (5.1) obtained by fitting configuration forces and energies to *ab initio* data. The same-species (*) data is different from Table 5.1.

H Alternate ternary metallic glass model

We repeat the analysis of Secs. 5.3.1 and 5.4.3, for an alternate ternary model glass system (AlYFe). In particular, we use the best parameter fits for the same-species interactions as opposed to implementing the GPT [155] as used in Figs. 5.3 – 5.7. The lower temperature system at $T = 300$ K in Fig. H2(a) shows a peak NMI at (ia) with a corresponding VI minimum at (ib). Following Sec. 5.4.3, Fig. H4 depicts a sample of the best clusters at $\gamma_{best} \simeq 0.001$ where we include overlapping node memberships (the replicas correlations are calculated on partitions). The corresponding $T = 1500$ K high temperature solutions have a much lower NMI at $\gamma_{best} \simeq 0.001$ indicating very poor agreement among the replicas. At $T = 300$ K, the best structures have consistent cluster sizes that are MRO or a little larger which are generally larger than 1/2 the simulated system width. Therefore, it would beneficial to test the consistency of the clusters in a larger system requiring a substantially longer computational time.

In this system the NMI plateau at $\gamma \gtrsim 100$ is actually higher than the configuration at (i), but the clusters are almost exclusively small ($n \simeq 5$ nodes) and are not completely contiguous. The distinction in the results between the different potential models is likely due to the longer range A-A minimum in the *ab initio* fit data as compared to the GPT minimum.

I Multiresolution application to lattice systems

We define several uniform lattices systems for the purpose of comparing the results to the model glasses where we use relatively small systems for presentation purposes. We would normally model the respective lattices with unweighted networks, but here we wish to be consistent with the analysis in Chapter 5. With this in mind, we further apply a "potential" shift ϕ_0 which corresponds to the negative of the average weight over *all* pairs of nodes (any non-neighbor has a weight of $b_{ij} = 1$).

We allow zero energy moves for the lattice solutions and perform a more strenuous optimization for these solutions. These representative networks result in "imperfect" tilings of the favored local structures due to the constraints imposed by the perfect symmetry in the Hamiltonian and by the local solution dynamics with the evolving community structure. In the depictions, different colors represent distinct clusters (best viewed in color) and edges *between* clusters are made partially transparent as a visualization aid. No overlapping nodes are assigned in the lattice depictions.

I.1 Square lattice

We define a uniform, initially unweighted, square lattice with $N = 400$ nodes. Edges are assigned to each neighbor in the x and y directions with periodic boundary conditions. The "potential" shift is $\phi_0 = 0.979\,95$. We then perform the same multiresolution analysis to the graph as in the previous systems. In Fig. I1, we see that where are three dominant plateaus in the information measures.

The overall multiresolution pattern resembles the LJ system in Fig. 5.4.4, except for the presence of the plateaus in the lattice plot, which suggests that the LJ system is "more ordered" than the metallic glass model. However, a purely random graph [37] also shows a similar pattern, including a plateau, which indicates that there may exist an analogy between purely "random" and perfectly "ordered" systems in our analysis. However, these data are not alone sufficient to be conclusive.

We select a configuration at $\gamma = 60$ corresponding to the center plateau. The lattice is depicted in Fig. I2 with $q = 120$ clusters of 78 squares, 4 triads, and 38 dyads. At this resolution, the square dominates the configuration which shows that our algorithm is able to isolate the basic unit cells of the lattice in a natural fashion.

The plateau for $\gamma \gtrsim 100$ corresponds to essentially all dyads, and the plateau for $\gamma \lesssim 30$ corresponds to a mixture of dyads, squares, and tight six-node configurations (a square plus two adjacent nodes). The lower γ plateau favors the six-node configuration in terms of the cluster energies, but the larger features are even more difficult with which to tile the lattice than for squares.

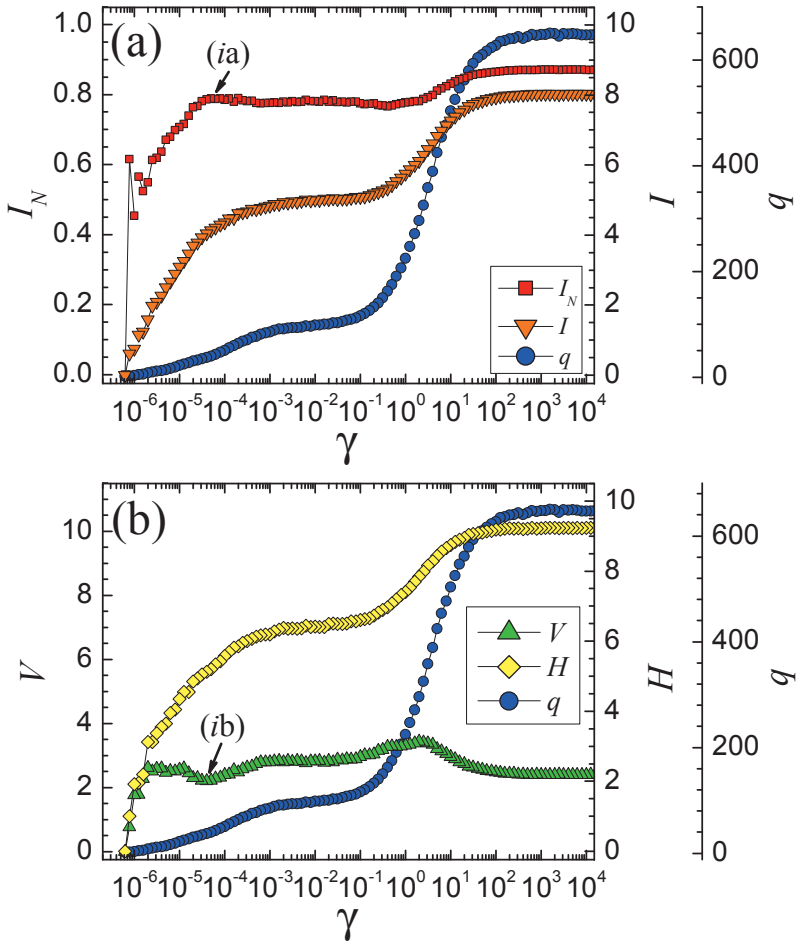


Figure H2: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The ternary model system contains 1600 atoms in a mixture of 88% A, 7% B, and 5% C with a simulation temperature of $T = 300$ K which is well *below* the melting temperature for this system. This alternate system uses parameter fits from Table H1 for the same-species interactions as opposed to the GPT interactions used in Table 5.1 and Figs. 5.3 – 5.7. This system shows a locally preferred resolution at (i) in both panels. A sample cluster for the best resolution at $\gamma \simeq 0.000\,1$ is depicted in Fig. H4. The region for $\gamma \gtrsim 100$ actually has a higher correlation than the local peak (i), but the clusters are very small ($n \simeq 5$ nodes) and somewhat dispersed.

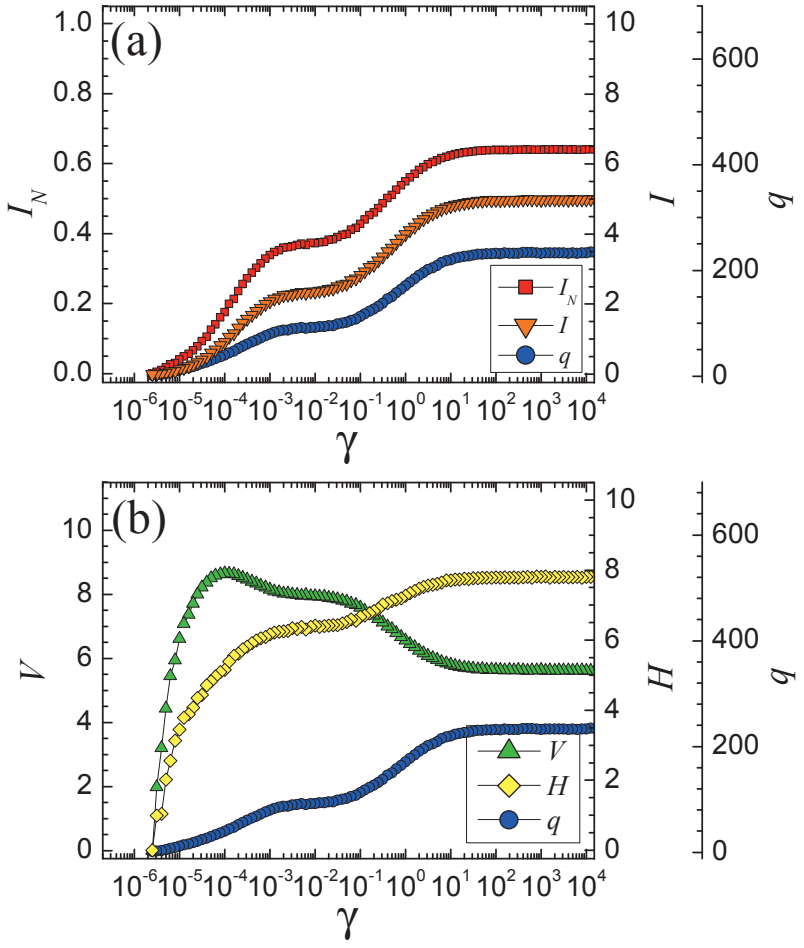


Figure H3: Panels (a) and (b) show the plots of information measures I_N, V, H, and I and the number of clusters q (right-offset axes) versus the Potts model weight γ in Eq. (2.2). The ternary model system contains 1600 atoms in a mixture of 88% A, 5% B, and 7% C with a simulation temperature of $T = 1500$ K which is well *above* the melting temperature of $T_m \simeq 1260$ K for this system. This alternate system uses parameter fits from Table H1 for the same-species interactions as opposed to the GPT interactions used in Table 5.1 and Figs. 5.3 – 5.7. At this temperature, there is no resolution where the replicas are strongly correlated. See Fig. H2 for the corresponding low temperature case where the replicas are much more highly correlated at $\gamma \simeq 0.000\,1$.

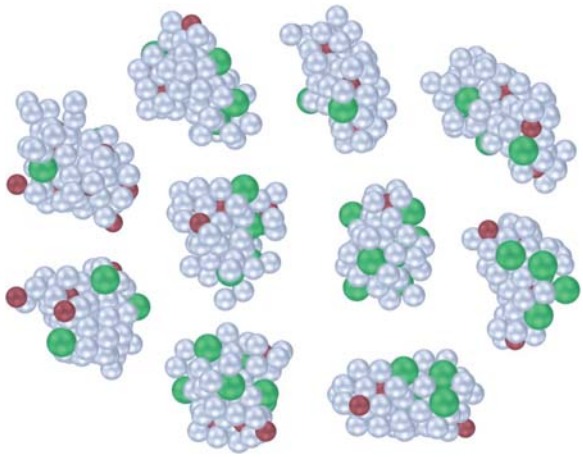


Figure H4: A depiction of some of the best clusters for the peak replica correlation at feature (i) in Fig. 5.3. These clusters include overlapping node membership assignments where each node is required to have an overall negative binding energy to the other nodes in the cluster. The atomic identities are C (red), A (silver), B (green) in order of increasing diameters. These clusters are generally larger than 1/2 the simulated system width; therefore, it would beneficial to test their consistency in a larger simulation (requiring substantially longer computational time).

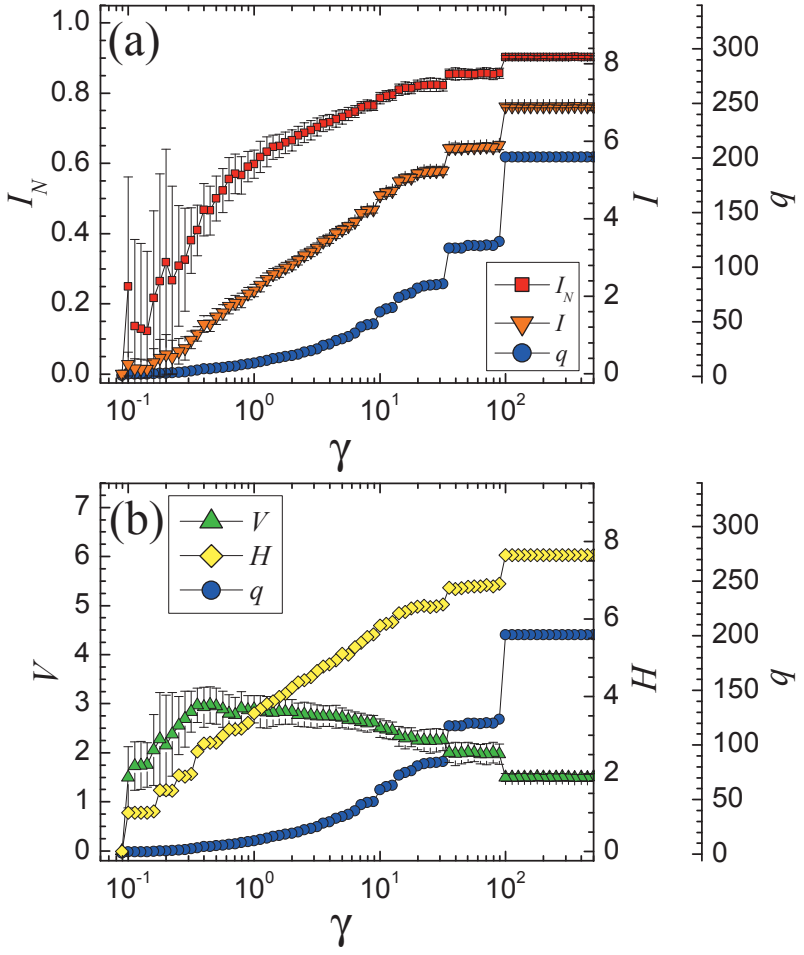


Figure I1: A plot for the multiresolution analysis of a square lattice with periodic boundary conditions. Neighbors edges have an initial weight of $a_{ij} = 1$ and non-neighbors have an initial weight of $b_{ij} = 1$. We further apply a "potential" shift of $\phi_0 = 0.979\,95$ in order to be consistent with our previous analysis on glasses in Chapter 5. There are three distinct plateaus in the information measures. A depiction of the system at $\gamma = 60$ for the center plateau is shown in Fig. I2.

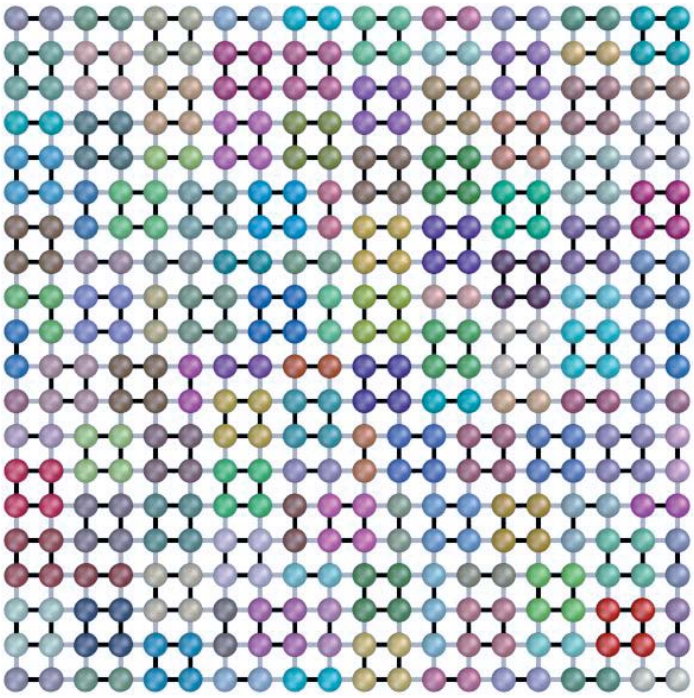


Figure I2: A depiction of a partition of a square lattice with periodic boundary conditions. The corresponding multiresolution plot is seen in Fig. I1. We use the algorithm described in the paper at $\gamma = 60$ to solve the system. To aid in visualization of the clusters, neighbor links *not in the same cluster* are made partially transparent. In this configuration, there were $q = 120$ clusters with 78 squares, 4 triads, and 38 dyads which indicates that square configuration dominates the partition, and it shows how our algorithm can naturally identify the basic unit cells of the square lattice.

I.2 Triangular lattice

Similar to the square lattice we define a uniform, initially unweighted, triangular lattice with $N = 400$ nodes. Edges are assigned to each triangular neighbor using periodic boundary conditions. The "potential" shift is $\phi_0 = 0.969\,925$. We again

perform the same multiresolution analysis to the graph with the results shown in Fig. I3. There are three dominant plateaus in the information measures; and the overall multiresolution pattern resembles both the square lattice and the LJ systems except for the presence of the plateaus here.

We select a configuration at $\gamma = 50$ corresponding to the "left" plateau. The lattice is depicted in Fig. I4 with $q = 104$ clusters of 17 triads, and 86 "diamonds", and 1 five-node cluster. The five node cluster is a result of a preference being given for an isolated node to join a diamond as opposed to forming its own single-node cluster. At this resolution, the diamond configuration dominates.

Two plateaus to the right of $\gamma \gtrsim 65$ are both strongly dominated by triads of nodes. The distinction between the two is that the central plateau favors an isolated node joining a triangle, to form a rare diamond, rather than forming its own single-node cluster. Together, the different plateaus show that our algorithm is able to isolate the basic unit cells of the lattice in a natural fashion.

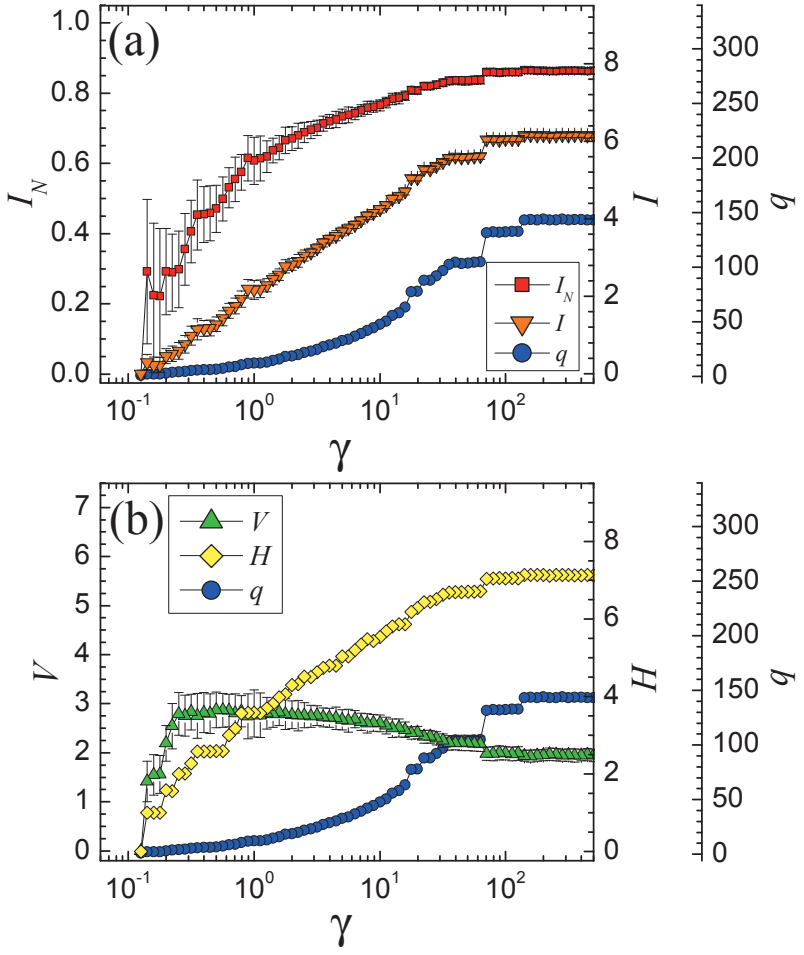


Figure I3: A plot for the multiresolution analysis of a triangular lattice with periodic boundary conditions. Neighbor edges have an initial weight of $a_{ij} = 1$ and non-neighbors have an initial weight of $b_{ij} = 1$. We further apply a "potential" shift of $\phi_0 = 0.969\,925$ in order to be consistent with our previous analysis on glasses in Chapter 5. There are three distinct plateaus in the information measures, but the latter two are closely related (see text). A depiction of the system at $\gamma = 50$ for the left plateau is shown in Fig. I4.

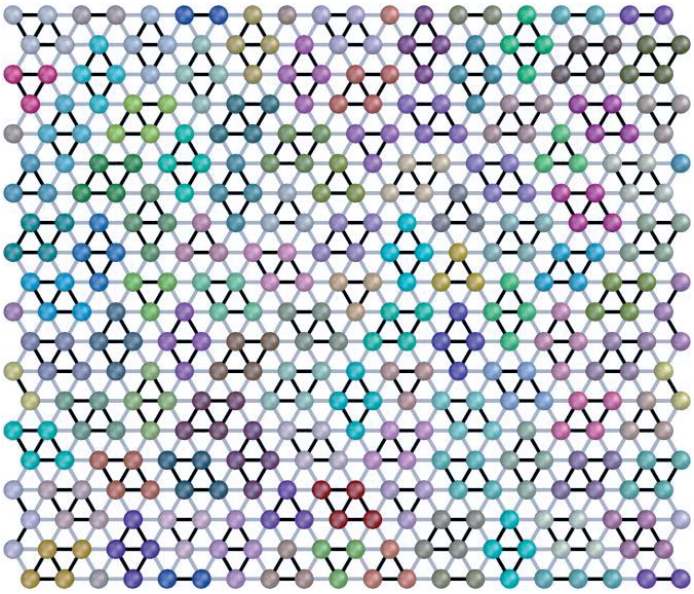


Figure I4: A depiction of a partition of a triangular lattice with periodic boundary conditions. The corresponding multiresolution plot is seen in Fig. I3. We use the algorithm described in the paper at $\gamma = 50$ (the "left" plateau) to solve the system. To aid in visualization of the clusters, neighbor links *not in the same cluster* are made partially transparent. In this configuration, there were $q = 104$ clusters with 17 triads, 86 "diamonds," and 1 five-node configuration (see text) which indicates that the diamond configuration dominates the partition. Our algorithm can naturally identify the different basic unit cells of the lattice.

I.3 Cubic lattice

We further define a uniform, initially unweighted, cubic lattice with $N = 1000$ nodes. Edges are assigned to each neighbor in the x, y, and z directions using periodic boundary conditions. The "potential" shift is $\phi_0 = 0.987\,988$. We perform the same multiresolution analysis to the graph as in the previous systems where the results are summarized in Figs. I5 and I6.

Out of $q = 177$ clusters, we identified 66 squares, 51 six-node configurations (square plus two adjacent nodes), 45 cubes, and 15 other assorted configurations smaller than cubes. At $\gamma \simeq 50$, the cube is the preferred cluster in terms of the cluster energy, but they consist of only slightly more than 25% of the clusters because the large cube configuration is more difficult to identify due to the perfect network symmetry and local constraints imposed by the evolution of the community structure during the algorithm dynamics. The "middle" plateau represents a square-dominated region (in any orientation) with 201 out of 294 clusters are squares, and the "right" plateau consists of dyads of nodes almost exclusively.

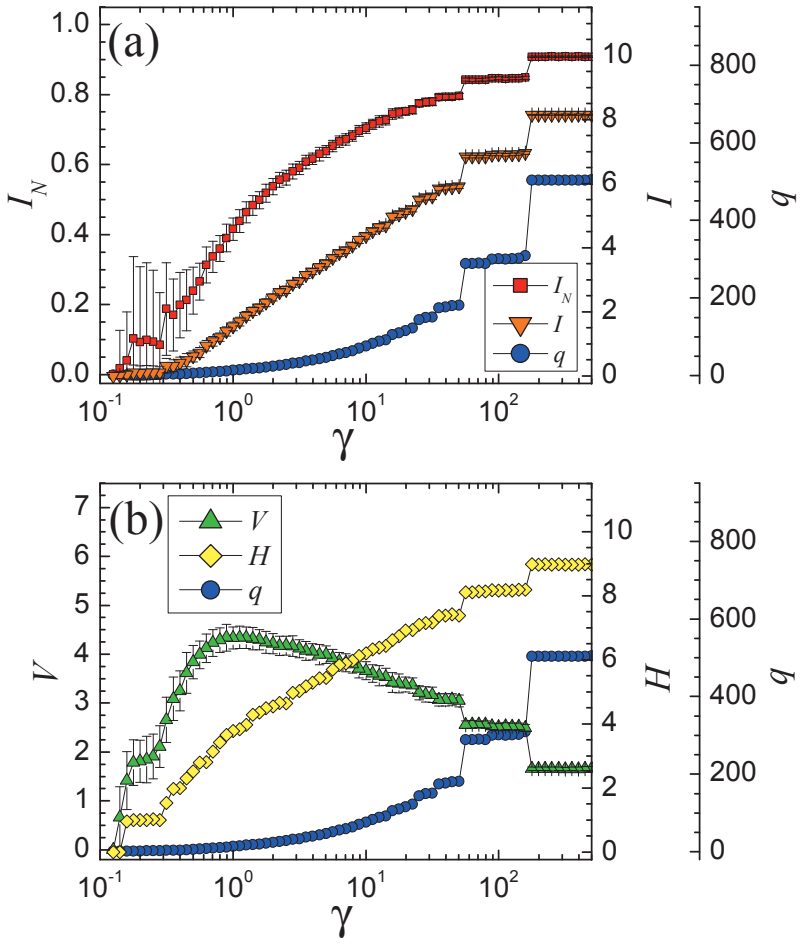


Figure I5: A plot for the multiresolution analysis of a cubic lattice with periodic boundary conditions. Neighbors edges have an initial weight of $a_{ij} = 1$ and non-neighbors have an initial weight of $b_{ij} = 1$. We further apply a "potential" shift of $\phi_0 = 0.987\,988$ in order to be consistent with our previous analysis on glasses in Chapter 5. There are three distinct plateaus in the information measures. The leftmost short plateau is the cube preferred resolution (in terms of cluster energy) with 45 out of 177 clusters are cubic clusters (no configurations larger than cubes are found). A depiction of the system at $\gamma = 50$ for the left plateau is shown in Fig. I6.

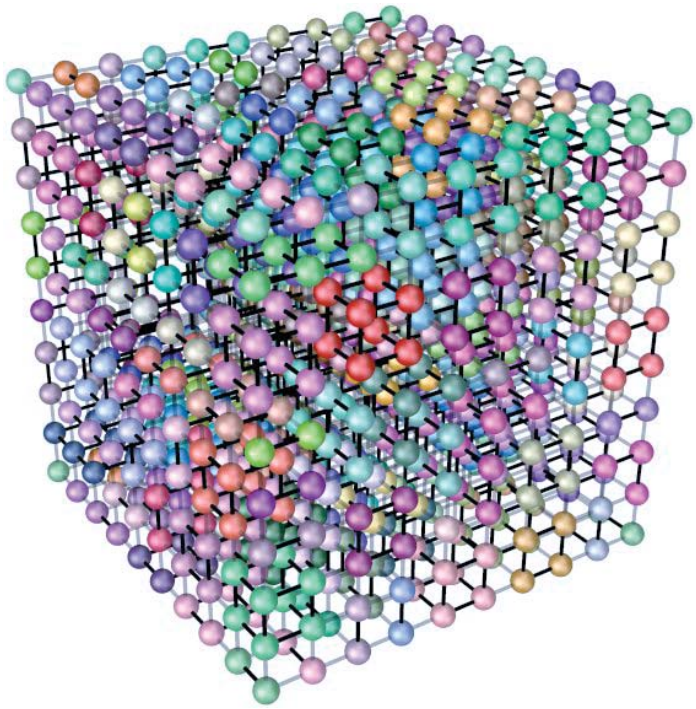


Figure I6: A depiction of a partition of a cubic lattice with periodic boundary conditions. The corresponding multiresolution plot is seen in Fig. I5. We solve the system using the algorithm described in Sec. 2.2 at $\gamma = 50$ (the short "left" plateau). Neighbor links *not in the same cluster* are made partially transparent as a visualization aid. In this configuration, there were $q = 177$ clusters with 66 squares, 51 six-node configurations, 45 cubes, and 15 other assorted configurations. A cubic cluster is the preferred partition (in terms of cluster energy), but it is difficult to fill the system with cubes in practice due to the perfect symmetry of the network and constraints imposed by the evolving community structure.

J Multiresolution application to a 2D Ising lattice

We define two-dimensional square lattice of Ising spins using the Hamiltonian

$$H = -\sum_{ij} \sigma_i \sigma_j \tag{J-1}$$

for all pairs of neighbor spins i and j. Neighbors with the same spin are assigned an edges with an initial weight of $a_{ij} = 1$, neighbors with opposite spins have no edge and are given an initial weight of $b_{ij} = 1$ (or a *repulsive* edge), and non-neighbors have an initial weight of $a_{ij} = b_{ij} = 0$. There is an unavoidable inherent discontinuity in defining the interactions for the Ising system (only *neighbors* can interact) and the continuous systems in Chapter 5 (*all* pairs of nodes interact). With this caveat, we analyze this Ising system according to the same multiresolution analysis for the systems in Chapter 5. With this in mind, we further apply the average "potential" shift ϕ_0 which varies on each defined network and which corresponds to the negative of the average weight over *all* pairs of nodes (including non-neighbors). We allow zero energy moves for the solution dynamics. No overlapping nodes are assigned in the Ising lattice.

In Fig. J2, different colors represent distinct clusters (best viewed in color) and edges *between* clusters (neighbor spins with the *same sign* but that are in *different* clusters) are depicted in gray. Missing edges are not depicted. In Fig. J1, there are two main plateaus for $\gamma \gtrsim 100$ and a noticeable configuration "shift" at $\gamma \simeq 170$. The "right" plateau for $\gamma \gtrsim 500$ consists largely of same-spin dyads except where a given spin has no matching neighbors. The "middle" plateau for $170 \lesssim \gamma \lesssim 500$ corresponds to a "natural" grouping of medium size clusters.

Here, the plateaus correspond to a cascade of structures starting from the smallest dyads of nodes, to basic plaquette structures (square, triangle, etc.), and growing ever larger (two joined plaquettes etc.). In Ising spin systems at different temperatures

on a square lattice, the domains of "+" and "−" spins are physically separated from one another by domain walls. The plateaus here similarly correspond to the cascade of small plaquettes found on the lattice up to a cutoff scale set by the domain walls.

No clear structure is seen beyond the natural domain length scale, but the domain walls are closely related to the *maximum* in VI (NMI displays a different, non-extremal, behavior where the *standard deviation* is large). Physically, they correspond to the scales at which the largest fluctuations occur where the large fluctuations result in poor information correlations between the replicas. Figure J3 shows a sample depiction of the system at this VI peak. Correlation lengths are thus likely related to poorly correlated replicas best indicated by a VI *maximum*. A more detailed analysis of the suspected relationship is a subject for further study.

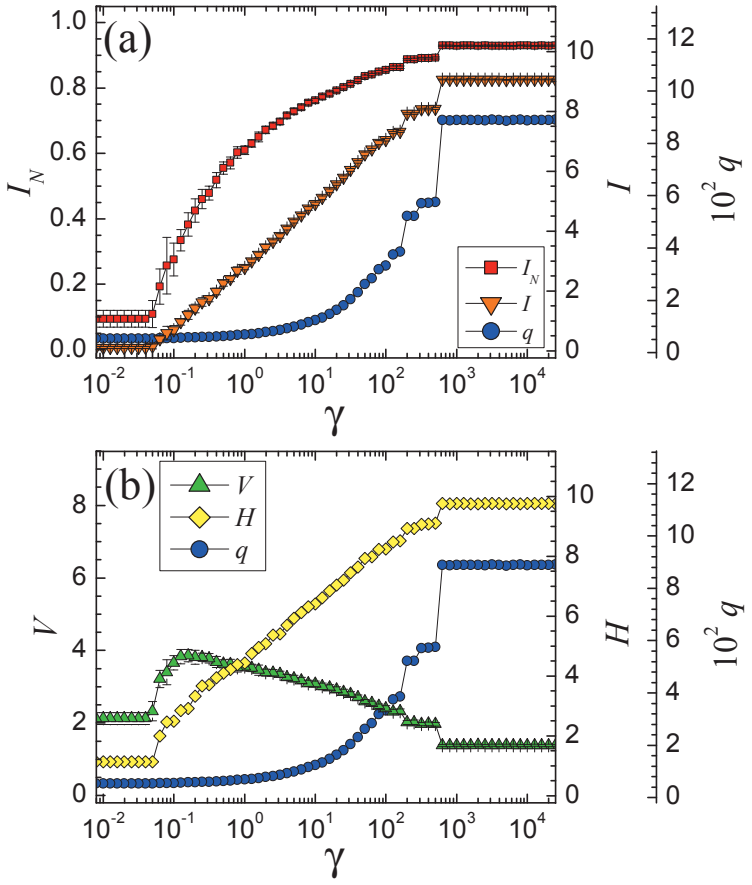


Figure J1: A plot for the multiresolution analysis of a square lattice of Ising spins with periodic boundary conditions at a simulation temperature of $T = 2.269$ (almost exactly at the magnetic transition). Neighbors with the same spin orientation are assigned edges with an initial weight of $a_{ij} = 1$. Neighbors with opposite spin orientations are assigned missing links with weights (or *repulsive* edges) of $b_{ij} = 1$, and non-neighbors have an initial weight of $a_{ij} = bij = 0$. There are two distinct plateaus in the information measures for $\gamma \gtrsim 100$ with a significant configuration shift near $\gamma \simeq 170$. A depiction of the system at $\gamma = 400$ for the center plateau is shown in Fig. J2, and a sample configuration for the VI *peak* at $\gamma \simeq 0.15$ is shown in Fig. J3.

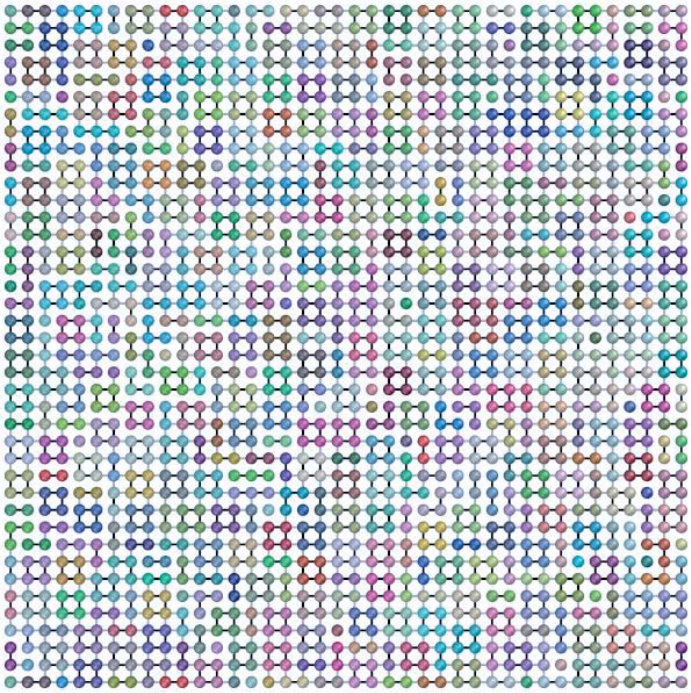


Figure J2: A depiction of a partition of a square lattice of Ising spins with periodic boundary conditions at a simulation temperature of $T = 2.269$. The corresponding multiresolution plot is seen in Fig. J1. We use the algorithm described in the paper at $\gamma = 400$ to solve the system. For presentation purposes, we depict all spins by spheres and links (between same-sign Ising spins) that cross a community boundary in gray. This multiresolution analysis shows the dominant communities are square plaquettes within same-spin domains.

K Multiresolution analysis of 2D LJ lattices with elastic defects

We define a uniform lattice based on a two-dimensional LJ system using the LJ potential in Eq. (5.2) where the lattice is constructed as the ideal ground state of the system using periodic boundary conditions. The main purpose is to check the

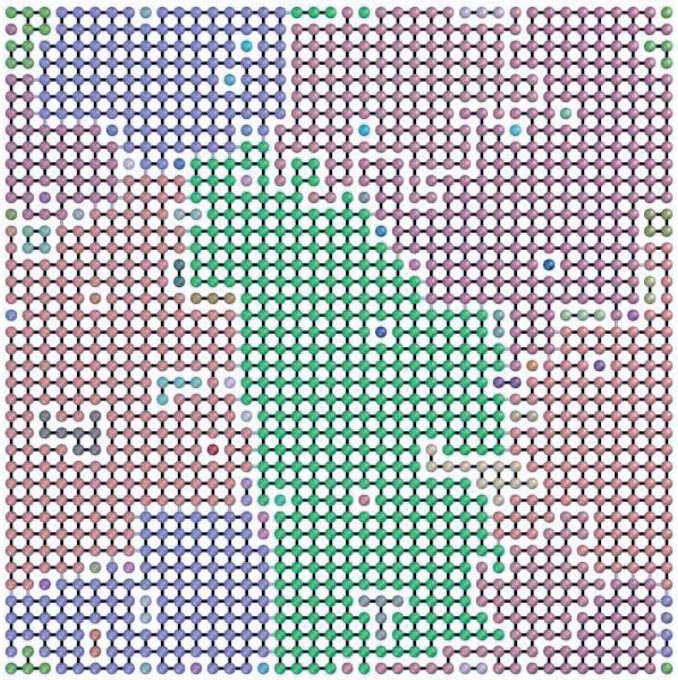


Figure J3: A depiction of a partition of a square lattice of Ising spins with periodic boundary conditions at a simulation temperature of $T = 2.269$. The corresponding multiresolution plot is seen in Fig. J1. We solve the network using the algorithm described in Sec. 2.2 at $\gamma \simeq 0.15$ (the actual partition here used an unshifted lattice), corresponding to the *peak* VI. For presentation purposes, we depict all spins by spheres. Edges (neighbors that have the same Ising spin) that cross a community boundary are depicted in gray. Any identified clusters vary substantially across the replicas since the community solutions are in a state of maximum fluctuation, but the cluster sizes (length) are of the scale of the size of the same spin domains. This qualitatively implies that the poor information correlations in our analysis may be related to the correlation length of the system.

consistency of our multiresolution method in Chapter 5, but the lattice also allows us to perform a preliminary investigation of how our analysis treats lattice defects in

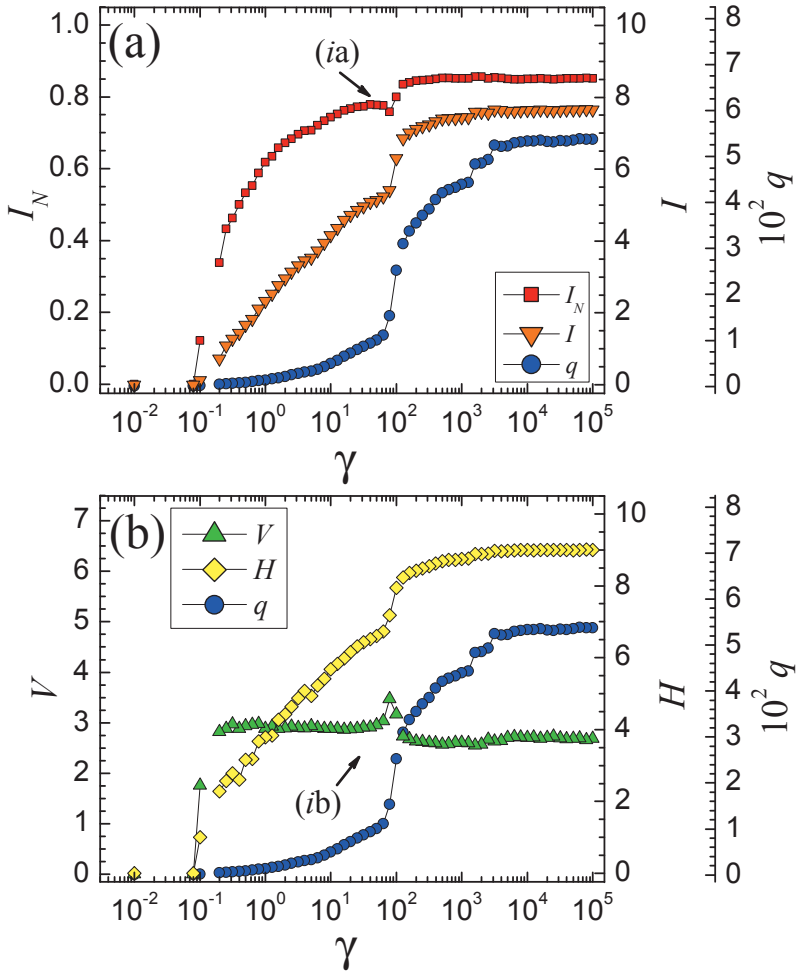


Figure K1: A plot for the multiresolution analysis of a 2D triangular LJ lattice in an ideal ground state using periodic boundary conditions. Edges are weighted according the LJ potential in Eq. (5.2). There are two distinct preferred regions, a small peak in the information measures on the left and a large plateau on the right. The left peak corresponds to the largest possible "natural" clusters for this system. We show a sample depiction at $\gamma \simeq 39$ in Fig. K2.

a controlled setting when some nodes are randomly removed from the system.

Figure K2: A depiction of a partition of a 2D LJ triangular lattice with periodic boundary conditions. The corresponding multiresolution plot is in Fig. K1 where the small peak on the left occurs at $\gamma \simeq 39$. This peak indicates the largest natural clusters for this particular lattice system.

K.1 LJ triangular lattice

Similar to the regular triangular lattice test in Sec. I.2, we define a uniform lattice with $N = 3120$ nodes. Any edges between pairs of nodes are assigned and weighted according to the LJ potential in Eq. (5.2) using periodic boundary conditions. We again perform the same multiresolution analysis on the graph from Chapter 5 where the results are shown in Fig. K1.

We select a configuration at $\gamma \simeq 39$ corresponding to the smaller peak on the

left where the best lattice at this resolution is depicted in Fig. K2. This particular resolution shows the *largest* "natural" clusters that appear with this particular lattice spacing. In general, we would further study a range of lattice spacings to more completely understand this system.

K.2 LJ triangular lattice with defects

Similar to the complete LJ lattice Sec. K.1, we define a uniform triangular lattice with $N = 2943$ nodes where we create random "defects" by removing some nodes. Edges are again assigned and weighted according to the LJ potential in Eq. (5.2) applying periodic boundary conditions. We perform the multiresolution analysis from Chapter 5, on the graph where the results are shown in Fig. K3.

We select a configuration at $\gamma \simeq 31.6$ from the smaller peak on the left which corresponds to the largest natural communities for this system at this specific lattice spacing. An example of the best lattice partition is depicted in Fig. K4. Defects appear to occur most likely *near the boundary of neighboring communities*.

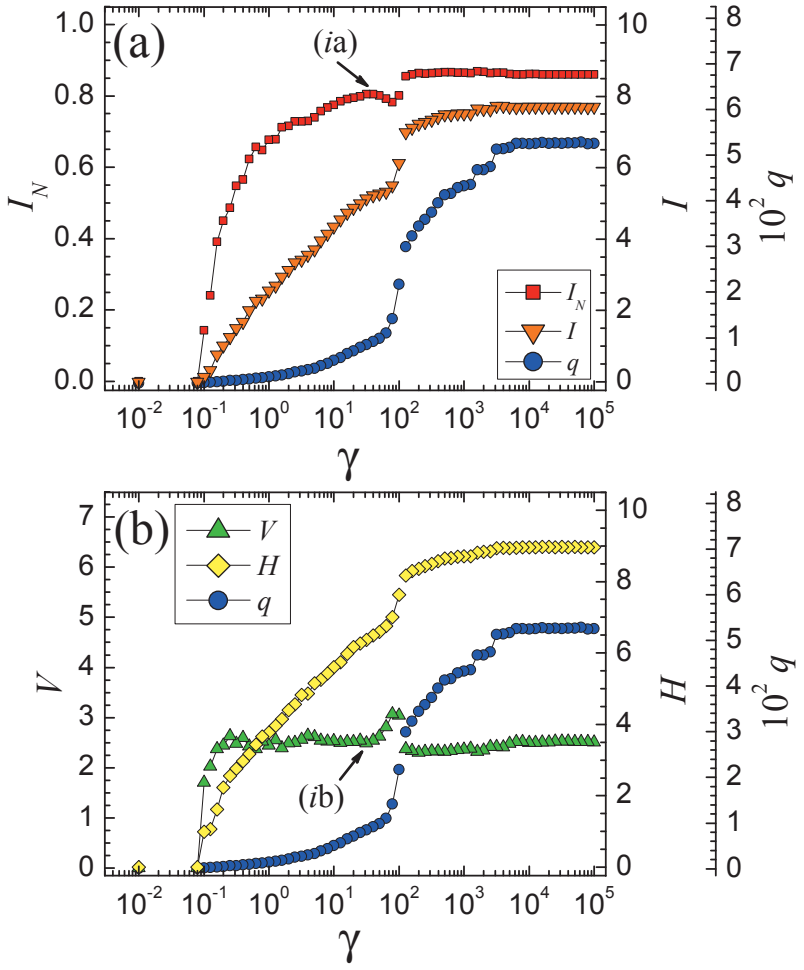


Figure K3: A plot for the multiresolution analysis of a 2D triangular LJ lattice with periodic boundary conditions. Edges are weighted according the LJ potential in Eq. (5.2). As with the complete (no defects) lattice in Fig. K1, there are two preferred regions, a small peak on the left and a large plateau on the right, where the peak here corresponds to the largest possible "natural" clusters. The defects make only a small alteration to the multiresolution plot. A depiction of the system at $\gamma \simeq 31.6$ for the left peak is shown in Fig. K4.

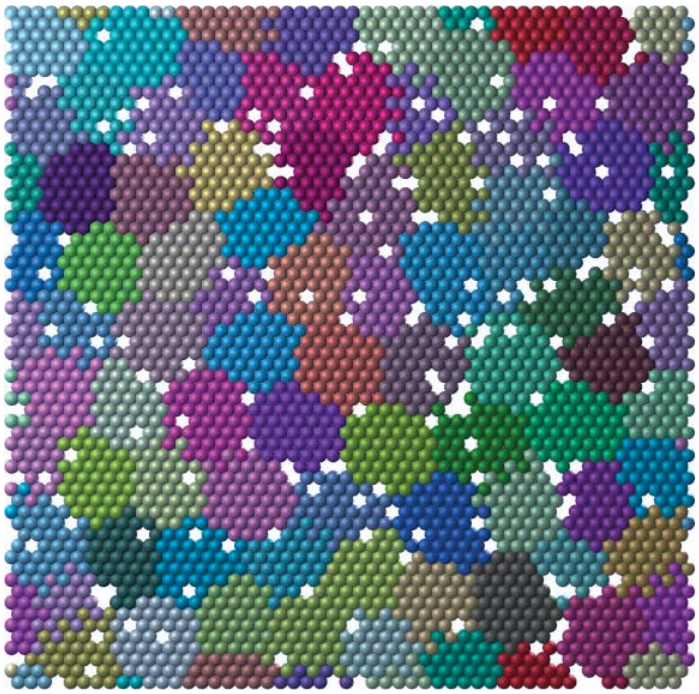


Figure K4: A depiction of a partition of a 2D LJ triangular lattice with periodic boundary conditions. The corresponding multiresolution plot is seen in Fig. K3. We use the algorithm described in the paper at $\gamma \simeq 31.6$ (the left peak) to solve the system. Our algorithm places defects generally near the boundary of the communities in order to minimize the energy cost of the defects in the community assignments.