Community Extraction in Multilayer Networks with Heterogeneous Community Structure*

James D. Wilson[†‡], John Palowitch[†§], Shankar Bhamidi[§], and Andrew B. Nobel[§]

December 14, 2016

Abstract

Multilayer networks are a useful way to capture and model multiple, binary relationships among a fixed group of objects. While community detection has proven to be a useful exploratory technique for the analysis of single-layer networks, the development of community detection methods for multilayer networks is still in its infancy. We propose and investigate a procedure, called Multilayer Extraction, that identifies densely connected vertex-layer sets in multilayer networks. Multilayer Extraction makes use of a significance based score that quantifies the connectivity of an observed vertex-layer set by comparison with a multilayer fixed degree random graph model. Unlike existing detection methods, Multilayer Extraction handles networks with heterogeneous layers where community structure may be different from layer to layer. The procedure is able to capture overlapping communities, and it identifies background vertex-layer pairs that do not belong to any community. We establish large-graph consistency of the vertex-layer set optimizer of our proposed multilayer score under the multilayer stochastic block model. We investigate the performance of Multilayer Extraction empirically on three applications, as well as a test bed of simulations. Our theoretical and numerical evaluations suggest that Multilayer Extraction is an effective exploratory tool for analyzing complex multilayer networks. Publicly available R software for Multilayer Extraction is available at https://github.com/jdwilson4/MultilayerExtraction.

Keywords: community detection, clustering, multiplex networks, score based methods, modularity

*The authors gratefully acknowledge Peter Mucha for helpful discussions and suggestions for this work. The work of JDW was supported in part by NSF grants DMS-1105581, DMS-1310002, and SES grant 1357622. The work of JP was supported in part by NIH/NIMH grant R01-MH101819-01. The work of SB was supported in part by NSF grants DMS-1105581, DMS-1310002, DMS-160683, DMS-161307, and SES grant 1357622. The work of ABN was supported in part by NSF DMS-1310002, NSF DMS-1613072, NIH HG009125-01, NIH MH101819-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[†]JDW and JP contributed equally to the writing of this paper.

[‡]Department of Mathematics and Statistics, University of San Francisco. San Francisco, CA 94117 jdwilson4@usfca.edu

[§]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill. Chapel Hill, NC 27599 palojj@email.unc.edu, bhamidi@email.unc.edu, and nobel@email.unc.edu

1 Introduction

Networks are widely used to represent and analyze the relational structure among interacting units of a complex system. In the simplest case, a network model is an unweighted, undirected graph $G = (V, E)$, where V is a set of vertices that represent the units, or *actors*, of the modeled system, and E is an edge set containing all pairs of vertices $\{u, v\}$ such that actors u and v share a physical or functional relationship. Networks have been successfully applied in a wide array of fields, including the social sciences to study social relationships among individuals [52], biology to study interactions among genes and proteins [1], and neuroscience to study the structure and function of the brain [48].

In many cases, the vertices of a network can be divided into groups (often disjoint) with the property that there are many edges between vertices in the same group, but relatively few edges between vertices in different groups. Vertex groups of this sort are commonly referred to as *communities*. The unsupervised search for communities in a network is known as *community detection*. Community structure has been used to identify functionally relevant groups in gene and protein interaction systems [30; 39], structural brain networks [3], and social networks [37; 24]. As communities are often associated with important structural characteristics of a complex system, community detection is a common first step in the understanding and analysis of networks. The search for communities that optimize a given quantitative performance criterion is typically an NP-hard problem, so in most cases one must rely on approximate algorithms to identify community structure.

The focus of this paper is community detection in *multilayer* networks. Formally, an (m, n)-multilayer network is a collection $\mathbf{G}(m, n) = (G_1, \ldots, G_m)$ of m simple graphs $G_\ell = ([n], E_\ell)$ having common vertex set $[n] = \{1, \ldots, n\}$, where the edge sets E_ℓ may vary from layer to layer. The graph G_ℓ will be referred to as the ℓth *layer* of the network. We assume that the vertices of the multilayer network are registered, in the sense that a fixed vertex $u \in [n]$ refers to the same actor across all layers. Thus the graph G_ℓ reflects the relationships between identified actors $1, \ldots, n$ in circumstance ℓ. There are no edges between vertices in different layers, and the layers are regarded as unordered so that the indices $\ell \in [m]$ do not reflect an underlying spatial or temporal order among the layers.

In general, the actors of a multilayer network may not exhibit the same community structure across all layers. For example in social networks, a group of individuals may be well connected via friendships on Facebook; however, this common group of actors will likely, for example, not work at the same company. In realistic situations such as these, a given vertex community will only be present in a subset of the layers, and different communities may be present in different subsets of layers. We refer to such multilayer systems as heterogeneous as each layer may exhibit noticeably different community structure. Complex and differential relationships between actors will be reflected in the heterogenous behavior of different layers of a multilayer network. In spite of this heterogeneity, existing community detection methods for multilayer networks typically assume that the community structure is the same across all, or a substantial fraction of, the layers.

We develop and investigate a multilayer community detection method called Multilayer Extraction, which efficiently handles multilayer networks with heterogeneous layers. Theoretical and numerical evaluation of our method reveals that Multilayer Extraction is an effective exploratory tool for analyzing complex multilayer networks. Our contributions to

the current literature of statistical analysis of multilayer networks are threefold

1. We develop a testing-based algorithm for identifying densely connected vertex-layer sets (B, L), where $B \subseteq [n]$ is a set of vertices and $L \subseteq [m]$ is a set of layers such that the vertices in B are densely connected across the layers in L. The strength of the connections in (B, L) is measured by a local modularity score derived from a null random network model that is based on the degree sequence of the observed multilayer network. Identified communities can have overlapping vertex or layer sets, and some vertex-layer pairs may not belong to any community. Vertex-layer pairs that are not assigned to any community are interpreted as background as they are not strongly connected to any other. Overlap and background are common features of real networks that can have deleterious effects on partition based methods [28; 53; 54]. The Multilayer Extraction procedure directly addresses community heterogeneity in multilayer networks.

2. We assess the consistency of Multilayer Extraction under a multilayer generalization of the stochastic 2 block model from [47; 51] for multilayer networks, which we call the multilayer stochastic block model (MSBM). The MSBM is a generative model that characterizes assortative behavior of pre-specified vertex-layer communities in a multilayer network. We are able to show that under the MSBM, the number of mis-clustered vertices and layers from the vertex-layer community that maximizes our proposed significance score vanishes to zero with high probability as the number of vertices tends to infinity. There has been considerable work in the area of consistency analysis for single-layer networks (e.g. [55]); however, to the best of our knowledge, we are the first to address the joint optimality properties for *both* vertices and layers. Furthermore, we provide complete and explicit expressions of all error bounds in the proof since we anticipate future analyses where the number of layers is allowed to grow with the size of the network. Our proof involves a novel inductive argument, which, to our knowledge, has not been employed elsewhere.

3. We apply Multilayer Extraction to three diverse multilayer networks, including social networks, arXiv citation networks and airline transportation networks. Our findings reveal important insights about these complex relational systems beyond the capabilities of existing detection methods. We further compare and contrast Multilayer Extraction with contemporary methods, and highlight the advantages of our approach over single layer and aggregate alternatives.

1.1 Related Work

Multilayer network models have been applied to a variety of problems, including modeling and analysis of air transportation routes [10], studying individuals with multiple sociometric relations [20; 21], and analyzing relationships between social interactions and economic exchange [19]. Kivelä et al. [27] and Boccaletti et al. [8] provide two recent reviews of multilayer networks. We note that $\mathbf{G}(m, n)$ is also sometimes referred to as a *multiplex* network.

While there is a large and growing literature concerning community detection in standard, single-layer, networks [22; 34; 43], the development of community detection methods for

multilayer networks is still relatively new. One common approach to multilayer community detection is to project the multilayer network in some fashion onto a single-layer network and then identify communities in the single layer network [5; 45]. A second common approach to multilayer community detection is to apply a standard detection method to each layer of the observed network separately [2; 6]. The first approach fails to account for layer-specific community structure and may give an oversimplified or incomplete summary of the community structure of the multilayer network; the second approach does not enable one to leverage or identify common structure between layers.

In addition to the methods above, there have also been several generalizations of single-layer methods to multilayer networks. For example, Holland et al. [26] and Paul and Chen [40] introduce multilayer generalizations of the standard stochastic block model from Wang and Wong [51] and Snijders and Nowicki [47]. However, these generative models require the community structure to be the same across layers. Paul and Chen [41] describe a class of null models for multilayer community detection based on the configuration and expected degree model. We utilize a similar model in our consideration. Stanley et al. [49] considered the clustering of layers of multilayer networks based on recurring community structure throughout the network. Mucha et al. [32] first extended the notion of modularity to multilayer networks, and De Domenico et al. [15] generalized the map equation, which measures the description length of a random walk on a partition of vertices, to multilayer networks. De Domenico et al. [16] discuss a generalization of the multilayer method in Mucha et al. [32] using tensor decompositions. Approximate optimization of either multilayer modularity or the map equation can be carried out by applying single network algorithms to a large graph formed by concatenating the layers of the observed graph. The communities that are identified by these methods form a partition of $[n] \times [m]$. By contrast, Multilayer Extraction identifies densely connected vertex-layer collections, directly addressing multilayer networks with heterogeneous layers.

1.2 Overview of the Paper

In the next section we describe the null multilayer random graph model and the scoring of vertex-layer sets. In Section 3 we present and prove theoretical results regarding the asymptotic consistency properties of our proposed score for multilayer networks. Section 5 provides a detailed description of the proposed Multilayer Extraction procedure. We apply Multilayer Extraction to three real-world multilayer networks and compare and contrast its performance with existing community detection methods in Section 6. In Section 7 we evaluate the performance of Multilayer Extraction on a test bed of simulated multilayer networks. We conclude the main paper with a discussion of future research directions in Section 8. The Appendix is divided into three sections. In Appendix A, we prove supporting lemmas contributing to the results given in Section 3. In Appendix B, we discuss competing methods to Multilayer Extraction. In Appendix C, we give the complete details of our simulation framework.

2 Scoring a Vertex-Layer Group

Seeking a vertex partition that optimizes, or approximately optimizes, an appropriate score function is a standard approach to single layer community detection. Prominent examples of score-based approaches include modularity maximization [36], likelihood maximization for a stochastic block model [51], as well as minimization of the conductance of a partition [12]. Rather than scoring a partition of the available network, Multilayer Extraction makes use of a significance based score that is applicable to individual vertex-layer sets. Below, we describe the multilayer null model, and then the proposed score.

2.1 The Null Model

Our significance-based score for vertex-layer sets in multilayer networks relies on the comparison of an observed multilayer network with a null multilayer network model. Let $\mathbf{G}(m, n)$ be an observed (m, n)-multilayer network. For each layer $\ell \in [m]$ and pair of vertices $u, v \in [n]$, let

$$x_\ell(u, v) = \mathbb{I}(\{u, v\} \in E_\ell)$$

indicate the presence or absence of an edge between u and v in layer ℓ of $\mathbf{G}(m, n)$. The *degree* of a vertex $u \in [n]$ in layer ℓ, denoted by $d_\ell(u)$, is the number of edges incident on u in G_ℓ. Formally,

$$d_\ell(u) \;=\; \sum_{v \in [n]} x_\ell(u, v).$$

The *degree sequence* of layer ℓ is the vector $\boldsymbol{d}_\ell = (d_\ell(1), \ldots, d_\ell(n))$ of degrees in that layer; the degree sequence of $\mathbf{G}(m, n)$ is the list $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m)$ containing the degree sequence of each layer in the network.

Let $\mathcal{G}(m, n)$ denote the family of all (m, n)-multilayer networks. Given the degree sequence $\boldsymbol{d}$ of the observed network $\mathbf{G}(m, n)$, we define a multilayer configuration model and an associated probability measure $\mathbb{P}_{\boldsymbol{d}}$ on $\mathcal{G}(m, n)$, as follows. In layer G_1, each node is given $d_1(u)$ half-stubs. Pairs of these edge stubs are then chosen uniformly at random, to form edges until all half-stubs are exhausted (disallowing self-loops and multiple edges). This process is done for every subsequent layer $G_2, \ldots, G_m$ independently, using the corresponding degree sequence from each layer.

In the multilayer network model described above, each layer is distributed according to the configuration model, first introduced by [9] and [4]. The probability of an edge between nodes u and v in layer ℓ depends only on the degree sequence $\boldsymbol{d}_\ell$ of the observed graph G_ℓ. The distribution $\mathbb{P}_{\boldsymbol{d}}$ has two complementary properties that make it useful for identifying communities in an observed multilayer network: (i) it preserves the degree structure of the observed network; and (ii) subject to this restriction, edges are assigned at random, without regard to the higher order connectivity structure of the network. Because of these characteristics, the configuration model has long been taken as the appropriate null model against which to judge the quality of a proposed community partition.

The configuration model is the null model which motivates the modularity score of a partition in a network [34; 36]. Consider a single-layer observed network $\mathbf{G}(n)$ with n nodes and degree sequence $\boldsymbol{d}$. For fixed $K > 0$, let $c_u \in [K]$ be the community assignment of node

u. The modularity score of the partition associated with the assignment $c_1, \ldots, c_n$ is defined as

$$M(c_1, \ldots, c_n;\ \mathbf{G}(n)) := \frac{1}{2|E|} \sum_{i \in [K]} \sum_{u < v \in [n]} \left(x(u,v) - \frac{d(u)d(v)}{\sum_{w \in [n]} d(w)} \right) \mathbb{I}(c_u = c_v = i). \tag{1}$$

Above, the ratio $\frac{d(u)d(v)}{\sum_{w \in [n]} d(w)}$ is the approximate expected number of edges between u and v under the configuration model. If the partition $c_1, \ldots, c_n$ represents communities with a large observed intra-edge count relative to what is expected under the configuration model, it receives a high modularity score. The identification of the communities that (approximately) maximize the modularity of a partition is among the most common techniques for community detection in networks.

2.2 Multilayer Extraction Score

Rather than scoring a partition, the Multilayer Extraction method scores individual vertex-layer sets. We define a multilayer node score that is based on the single-layer modularity score (1) and amenable to iterative maximization. We first define a local *set* modularity for a collection of vertices $B \subseteq [n]$ in the layer $\ell \in [m]$:

$$Q_\ell(B) := \frac{1}{n\binom{|B|}{2}^{1/2}} \sum_{u,v \in B : u < v} \left(x_\ell(u,v) - \frac{d_\ell(u)d_\ell(v)}{\sum_{w \in [n]} d_\ell(w)} \right) \tag{2}$$

The scaling term in the equation above is related to the total number of vertices in the network and the total number of possible edges between the vertices in B. This score is one version of the various set-modularities considered in [18], and is reminiscent of the *local* modularity score introduced in [13].

Our Multilayer Extraction procedure seeks communities that are *assortative* across layers, in the sense that $Q_\ell(B)$ is large and positive for each $\ell \in L$. In light of this, we define the *multilayer set score* as

$$H(B, L) := \frac{1}{|L|} \left(\sum_{\ell \in L} Q_\ell(B)_+ \right)^2, \tag{3}$$

where Q_+ denotes the positive part of Q. Generally speaking, the score acts as a yardstick with which one can measure the connection strength of a vertex-layer set. Large values of the score signify densely connected communities.

We note that the multilayer score $H(B, L)$ is reminiscent of a chi-squared test-statistic computed from $|L|$ samples. That is, under appropriate regularity assumptions on $Q_\ell(B)$, the score in (3) will be approximately chi-squared with one degree of freedom.

3 Consistency Analysis

Existing community detection methods differ widely in their underlying performance criteria, as well as the algorithms employed to identify communities that (approximately) optimize these criteria. As such, it can be difficult to compare the effectiveness of one community detection method with another, and it has become common to evaluate community detection methods through a theoretical study of their asymptotic consistency. The first work in this direction was that of [7], who investigated the consistency of score-based community detection methods in single-layer networks with a large number of vertices. Consistency analysis has since been used to evaluate a number of community detection methods, including spectral clustering [46], modularity-based methods [17; 31; 33], likelihood-based methods [55; 11], and aggregate community detection methods for multilayer networks with homogeneous layers [25].

3.1 The Multilayer Stochastic Block Model

We assess the consistency of Multilayer Extraction under the multilayer stochastic block model (MSBM) with two vertex communities, defined as a probability distribution $\mathbb{P}_{m,n} = \mathbb{P}_{m,n}(\cdot|\boldsymbol{P}, \pi_1, \pi_2)$ on the family of undirected multilayer networks with m layers, n vertices and 2 communities. The distribution is fully characterized by (i) containment probabilities $\pi_1, \pi_2 > 0$, which satisfy $\pi_1 + \pi_2 = 1$, and (ii) a sequence of symmetric 2×2 matrices $\boldsymbol{P} = \{P_1, \ldots, P_m\}$ where $P_\ell = \{P_\ell(i,j)\}$ with entries $P_\ell(i,j) \in (0,1)$. Under the distribution $\mathbb{P}_{m,n}$, a random multilayer network $\widehat{\boldsymbol{G}}(m,n)$ is generated using two simple steps:

1. A subset of $\lceil \pi_1 n \rceil$ vertices are placed in community 1, and remaining vertices are placed in community 2. Each vertex u in community j is assigned a community label $c_u = j$.

2. An edge is placed between nodes $u, v \in [n]$ in layer $\ell \in [m]$ with probability $P_\ell(c_u, c_v)$, independently from pair to pair and across layers, and no self-loops are allowed.

For a fixed n and m, the community labels $\mathbf{c}_n = (c_1, \ldots, c_n)$ are chosen once and only once, and the community labels are the same across each layer of $\widehat{\boldsymbol{G}}(m,n)$. On the other hand, the inner and intra community connection probabilities (and hence the assortativity) can be different from layer to layer, introducing heterogeneity among the layers. Note that when $m = 1$, the MSBM reduces to the (single-layer) stochastic block model from Wang and Wong [51].

3.2 Consistency of the Score

We evaluate the consistency of the Multilayer Extraction score under the MSBM described above. Our first result addresses the vertex set maximizer of the score given a fixed layer set $L \subseteq [m]$. Our second result (Theorem 3 in Section 3.2.1) leverages the former to analyze the global maximizer of the score across layers and vertex sets. Explicitly, consider a multilayer network $\widehat{\boldsymbol{G}}(m,n)$ with distribution under the multilayer stochastic block model

$\mathbb{P}_{m,n} = \mathbb{P}_{m,n}(\boldsymbol{P}, \pi_1, \pi_2)$. For a fixed vertex set $B \subseteq [n]$ and layer set $L \subseteq [m]$, define the random score by

$$\widehat{H}(B, L) := \frac{1}{|L|} \left(\sum_{\ell \in L} \widehat{Q}_\ell(B)_+ \right)^2,$$

where $\widehat{Q}_\ell(B)$ is the set-modularity of B in layer ℓ under $\mathbb{P}_{m,n}$. Our main results address the behavior of $\widehat{H}(B, L)$ under various assuptions on the parameters of the MSBM.

Toward the first result, for a fixed layer set $L \subseteq [m]$, let $\widehat{B}^{(n)}_{opt}(L)$ denote the node set that maximizes $\widehat{H}(B, L)$ (if more than one set does, any may be chosen arbitrarily). To define the notion of a "misclassified" node, for any two sets $B_1, B_2 \subseteq [n]$ let $d_h(B_1, B_2)$ denote the Hamming distance (rigorously defined as the cardinality of the symmetric difference between B_1 and B_2). We then define the number of misclassified nodes by a set B by

$$\texttt{Error}(B) := d_h(B, C_1) \wedge d_h(B, C_2).$$

Note that this definition accounts for arbitrary labeling of the two communities. As the nodes and community assignments are registered across layers, neither d_h nor `Error` depend on the choice of L. Before stating the main theorem, we define a few quantities that will be used throughout its statement and proof:

Definition 1. *Let "*det*" denote matrix determinant. For a fixed layer set $L \subseteq [m]$, define*

$$\delta_\ell := \det P_\ell \qquad \delta(L) := \min_{\ell \in [L]} \delta_\ell \qquad \pi := (\pi_1, \pi_2)^t \qquad \kappa_\ell := \pi^T P_\ell \pi \qquad \kappa(L) := \min_{\ell \in [L]} \kappa_\ell \tag{4}$$

We now state the fixed-layer-set consistency result:

Theorem 2. *Fix m and let $\{\widehat{\boldsymbol{G}}(m, n)\}_{n>1}$ be a sequence of multilayer stochastic 2 block models where $\widehat{\boldsymbol{G}}(m, n)$ is a random graph with m layers and n nodes generated under $\mathbb{P}_{m,n}(\cdot|\boldsymbol{P}, \pi_1, \pi_2)$. Assume $\pi_1 \leqslant \pi_2$, and that π_1, π_2, and $\mathbb{P}$ do not change with n. Fix a layer set $L \subseteq [m]$. If $\delta(L) > 0$ then there exist constants $A, \eta > 0$ depending on π_1 and $\delta(L)$ such that for all fixed $\varepsilon \in (0, \eta)$,*

$$\mathbb{P}_{m,n} \left(\texttt{Error} \left(\widehat{B}^{(n)}_{opt}(L) \right) < An^\varepsilon \log n \right) \geqslant 1 - \exp \left\{ -\frac{\kappa(L)^2 \varepsilon}{32} n^\varepsilon (\log n)^{2-\varepsilon} + \log 4|L| \right\} \tag{5}$$

for large enough n.

Note that an immediate corollary of Theorem 2 is that for any $\varepsilon \in (0, 1)$,

$$\mathbb{P}_{m,n} \left(\texttt{Error} \left(\widehat{B}^{(n)}_{opt}(L) \right) < n^\varepsilon \log n \right) \to 1 \quad \text{as} \quad n \to \infty.$$

Therefore, the constants A and η play a role only in bounding the convergence rate of the probability.

The proof of Theorem 2 is given in Section 4.1. We note that the assumption that $\pi_1 \leqslant \pi_2$ is made without loss of generality, since the community labels are arbitrary. When $m = 1$, Theorem 2 implies asymptotic $n \to \infty$ consistency in the (single-layer) stochastic

block model. In this case, the condition that $\delta_\ell = P_\ell(1,1)P_\ell(2,2) - P_\ell(1,2)^2 > 0$ is a natural requirement on the inner community edge density of a block model. This condition appears in a variety of consistency analyses, including the evaluation of modularity [55]. When $m > 1$, Theorem 2 implies the vertex set that maximizes $H(B, L)$ will have asymptotically vanishing error with high probability, given that L is a fixed layer set with *all* layers satisfying $\delta_\ell > 0$.

3.2.1 Consistency of the joint optimizer

Theorem 2 does not address the *joint* optimizer of the score across all vertex-layer pairs. First, we point out that for a fixed $B \subseteq [n]$, the limiting behavior of the score $\widehat{H}(B, L)$ depends on $L \subseteq [m]$ through the layer-wise determinants $\{\delta_\ell : \ell \in [n]\}$ and the scaling constant $\frac{1}{|L|}$ inherent to $H(B, L)$, as defined in equation (3). Let $\gamma : \mathbb{N}^+ \mapsto \mathbb{R}^+$ be a non-decreasing function of $|L|$. Define

$$H_\gamma(B, L) := \frac{1}{\gamma(|L|)} \left(\sum_{\ell \in L} Q_\ell(B)_+ \right)^2. \tag{6}$$

and let $\widehat{H}_\gamma(B, L)$ be the corresponding random version of this score under an MSBM. We analyze the joint node-set optimizer of H_γ under some representative choices of γ, an analysis which will ultimately motivate the choice $\gamma(|L|) = |L|$.

We first provide an illustrative example. Consider a MSBM with $m > 1$ layers having the following structure: the first layer has positive determinant, and all $m - 1$ remaining layers have determinant equal to 0. Note that $\delta_1 > 0$ implies that the first layer has ground-truth assortative community structure, and that $\delta_\ell = 0$ for all $\ell > 1$ implies that the remaining layers are (independent) Erdos-Renyi random graphs. In this case, the desired global optimizer of $H_\gamma(B, L)$ is community 1 (or 2) and the first layer. However, setting $\gamma(|L|) \equiv 1$ (effectively ignoring the scaling of H) will ensure that, in fact, the *entire* layer set is optimal, since $Q_\ell(B)_+ \geqslant 0$ by definition. It follows that setting $\gamma(|L|)$ to increase (strictly) in $|L|$, which introduces a penalty on the size of the layer set, is desirable.

For a fixed scaling function γ, define the global joint optimizer of $\widehat{H}(B, L)$ by

$$\left(\widehat{L}^{(n)}_{opt}, \widehat{B}^{(n)}_{opt}\right) := \underset{2^{[n]} \times 2^{[m]}}{\arg\max}\ \widehat{H}_\gamma(B, L) \tag{7}$$

Note that $\left(\widehat{L}^{(n)}_{opt}, \widehat{B}^{(n)}_{opt}\right)$ is random, and may contain multiple elements of $2^{[m]} \times 2^{[n]}$. The next theorem addresses the behavior of $\left(\widehat{L}^{(n)}_{opt}, \widehat{B}^{(n)}_{opt}\right)$ under the MSBM for various choices of $\gamma(|L|)$, and shows that setting $\gamma(|L|) = |L|$ is desirable for consistency.

Theorem 3. *Fix m and let $\{\widehat{\boldsymbol{G}}(m, n)\}_{n>1}$ be a sequence of multilayer stochastic 2 block models where $\widehat{\boldsymbol{G}}(m, n)$ is a random graph with m layers and n nodes generated under $\mathbb{P}_{m,n}(\cdot|\boldsymbol{P}, \pi_1, \pi_2)$. Assume $\pi_1 \leqslant \pi_2$, and that π_1, π_2, and $\mathbb{P}$ do not change with n. Fix $0 = \delta^{(0)} < \delta^{(1)} < 1$. Suppose the layer set $[m]$ is split according to $[m] = \cup_{i=0,1} L_i$, where $\delta_\ell = \delta^{(i)}$ for all $\ell \in L_i$. Then for any $\varepsilon > 0$, the following hold:*

(a) Let $\widehat{L}^+ := \{\ell : \widehat{Q}_\ell(\widehat{B}_{opt}^{(n)}) > 0\}$. *If* $\gamma(|L|) \equiv 1$, *then for all* $n > 1$, $\widehat{L}_{opt}^{(n)} = \widehat{L}^+$, *and*

$$\mathbb{P}_{m,n}\left(\texttt{Error}\left(\widehat{B}_{opt}^{(n)}\right) < n^{\varepsilon}\log n\right) \to 1 \quad as \quad n \to \infty$$

(b) If $\gamma(|L|) = |L|$,

$$\mathbb{P}_{m,n}\left(\widehat{L}_{opt}^{(n)} = L_1,\ \texttt{Error}\left(\widehat{B}_{opt}^{(n)}\right) < n^{\varepsilon}\log n\right) \to 1 \quad as \quad n \to \infty$$

(c) If $\gamma(|L|) = |L|^2$,

$$\mathbb{P}_{m,n}\left(\widehat{L}_{opt}^{(n)} \subseteq 2^{L_1},\ \texttt{Error}\left(\widehat{B}_{opt}^{(n)}\right) < n^{\varepsilon}\log n\right) \to 1 \quad as \quad n \to \infty$$

The proof of Theorem 3 is given in Section 4.2. Part (a) implies that setting $\gamma(|L|) \equiv 1$ ensures that the optimal layer set will be, simply, all layers with positive modularity, thereby making this an undesirable choice for the function γ. Part (c) says that if $\gamma(|L|) = |L|^2$, the layer set with the highest *average* layer-wise modularity will be optimal (with high probability as $n \to \infty$), which means that all subsets of L_1 are asymptotically equivalent with respect to $\widehat{H}(B, L)$ (with high probability). By part (b), if $\gamma(|L|) = |L|$, then L_1 is the unique asymptotic maximizer of the population score (with high probability). Therefore, $\gamma(|L|) = |L|$ is the most desirable choice of scaling.

4 Proofs

In this section we prove the theoretical results given in Section 3. The majority of the section is devoted to a detailed proof of Theorem 2 and supporting lemmas. This is followed by the proof of Theorem 3, of which we give only a sketch, as many of the results and techniques contributing to the proof of Theorem 2 can be re-used.

4.1 Proof of Theorem 2, and Supporting Lemmas

We prove Theorem 2 via a number of supporting lemmas. We begin with some notation:

Definition 4. *For a fixed vertex set* $B \subseteq [n]$ *define*

$$\rho_n(B) = \frac{|B \cap C_{1,n}|}{|B|}, \quad s_n(B) = \frac{|B|}{n}, \quad v_n(B) := (\rho_n(B), 1 - \rho_n(B)) \tag{8}$$

We will at times suppress dependence on n *and* B *in the above expressions.*

Definition 5. *Define the* ***population*** *normalized modularity of a set* B *in layer* ℓ *by*

$$q_\ell(B) := \frac{s_n(B)}{\sqrt{2}}\left(v_n(B)^t P_\ell v_n(B) - \frac{(v_n(B)^t P_\ell \pi)^2}{\kappa_\ell}\right) \tag{9}$$

Define the ***population*** *score function* $H_*(\cdot, L) : 2^{[n]} \mapsto \mathbb{R}$ *by*

$$H_*(B, L) = |L|^{-1} \left(\sum_{\ell \in [L]} q_\ell(B) \right)^2 \tag{10}$$

Throughout the results in this section, we assume that $L \subseteq [m]$ is a fixed layer set (as in the statement of Theorem 2). We will therefore, at times, suppress the dependence on L from $\delta(L)$ and $\kappa(L)$ (from Definition 1).

4.1.1 Sketch of the Proof of Theorem 2

The proof of Theorem 2 is involved and broken into many lemmas. In this section, we give a rough sketch of the argument, as follows. The lemmas in this section establish that:

1. $C_{1,n}$ maximizes the population score $H_*(\cdot, L)$ (Lemmas 6 and 7).
2. For large enough sets $B \subseteq [n]$, the random score $\widehat{H}(B, L)$ is bounded in probability around the population score $H_*(B, L)$ (Lemmas 9 and 12).
3. **Inductive Step**: For fixed $k > 1$, assume that $d_h(\widehat{B}^{(n)}_{opt}(L), C_{1,n})/n = O_p(b_{n,k})$, where larger k makes $b_{n,k}$ of smaller order. Then, based on concentration properties of the score, in fact $d_h(\widehat{B}^{(n)}_{opt}(L), C_{1,n})/n = O_p(b_{n,k+1})$ (Lemma 13).
4. There exists a constant η such that for any $\varepsilon \in (0, \eta)$, $d_h(\widehat{B}^{(n)}_{opt}(L), C_{1,n})/n = O_p(n^\varepsilon \log n)$ (Theorem 2). This result is shown using the Inductive Step.

4.1.2 Supporting lemmas for the Proof of Theorem 2

Lemma 6. *Define* $\phi(L) := (|L|^{-1} \sum_\ell \frac{\det P_\ell}{2\kappa_\ell})^2$. *Then:*

1. *For any* $B \subseteq [n]$, $q_\ell(B) = \frac{s_n(B)}{\sqrt{2}}(\pi_1 - \rho_n(B))^2 \cdot \frac{\det P_\ell}{2\kappa_\ell}$, *and therefore*

$$H_*(B, L) = |L|\phi(L)\frac{s_n(B)^2}{2}(\pi_1 - \rho_n(B))^4$$

2. $\delta(L)^2 \leqslant \phi(L) \leqslant \frac{1}{\pi_1^2 \delta(L)^2}$ *and therefore* $H_*(C_{1,n}, L) \geqslant |L|\frac{\pi_1^2}{2}(1 - \pi_1^4)\delta(L)^2$

Lemma 7. *Fix any* $n > 1$. *Define*

$$\mathcal{R}(t) := \begin{cases} \{B \subseteq [n] : |s(B) - \pi_1| \vee [1 - \rho(B)] \leqslant t\}, & \pi_1 < \pi_2 \\ \{B \subseteq [n] : |s(B) - \pi_1| \vee \rho(B) \leqslant [1 - \rho(B)] \leqslant t\}, & \pi_1 = \pi_2 \end{cases}$$

Then there exists a constant $a > 0$ *depending just on* π_1 *such that for sufficiently small* t, $B \notin \mathcal{R}(t)$ *implies* $H_*(B, L) < H_*(C_{1,n}, L) - a|L|\phi(L)t$.

The proofs of Lemmas 6-7 are given in Appendix A. We now give a general concentration inequality for $\widehat{H}(B,L)$, which shows that for sufficiently large sets $B \subseteq [n]$, $\widehat{H}(B,L)$ is close to the population score $H_*(B,L)$ with high probability. This result is used in the proof of Lemma 12, and its proof is given in Appendix A. We first give the following definition:

Definition 8. *For fixed $\varepsilon > 0$ and $n > 1$, define $\mathcal{B}_n(\varepsilon) := \{B \subseteq [n] : |B| \geqslant n\varepsilon\}$.*

Lemma 9. *Fix $\varepsilon \in (0,1)$. Let κ be as in Definition 1. For each $n > 1$ suppose a collection of node sets $\mathcal{B}_n$ is contained in $\mathcal{B}_n(\varepsilon)$. Then for large enough n,*

$$\mathbb{P}_n\left(\sup_{\mathcal{B}_n}\left(\left|\widehat{H}(B,L) - H_*(B,L)\right|\right) > \frac{4|L|t}{n^2} + \frac{52|L|}{\kappa n}\right) \leqslant 4|L||\mathcal{B}_n|\exp\left(-\kappa^2\frac{\varepsilon t^2}{16n^2}\right)$$

for all $t > 0$.

We now define new notation that will serve the remaining lemmas:

Definition 10. *Let $\gamma_n := \log n/n$, and for any integer $k > 0$, define $b_{n,k} := \gamma_n^{1-\frac{1}{2^k}}$.*

Definition 11. *For any $r \in [0,1]$ and $C \subseteq [n]$, define the r-neighborhood of C by $N(C,r) := \{B \subseteq [n] : d_h(B,C)/n \leqslant r\}$. For all $n > 1$, any constant $A > 0$, and fixed $k > 1$, define*

$$\widetilde{N}_{n,k}(A) := \begin{cases} N\left(C_1, A \cdot b_{n,k-1}\right) \cup N\left(C_2, A \cdot b_{n,k-1}\right), & k > 1 \\ \mathcal{B}_n(A), & k = 1 \end{cases}$$

Lemma 12, stated below, is a concentration inequality for the random variable $\widehat{H}(B,L)$ on particular neighborhoods of C_1:

Lemma 12. *Fix $\varepsilon \in (0, \pi_1)$ and any constant $A > 0$. For $k > 1$ satisfying $1/2^{k-1} < \varepsilon$, we have for sufficiently large n that*

$$\sup_{B \in \widetilde{N}_{n,k}(A)}\left|\widehat{H}(B,L) - H_*(B,L)\right| \leqslant 5|L|b_{n,k} \tag{11}$$

with probability greater than $1 - 2\exp\{-\frac{\kappa^2\varepsilon}{32}n\gamma_n^{1-\varepsilon}\log(n) + \log 4|L|\}$. The conclusion holds with $k = 1$ if $A = \varepsilon$.

The proof of Lemma 12 is given in Appendix A. We now state and prove the key lemma used to drive the induction step in the proof of Theorem 2:

Lemma 13. *Fix $\varepsilon \in (0, \pi_1)$ and an integer $k > 1$ satisfying $1/2^{k-1} < \varepsilon$. Suppose there exist constants $A, b > 0$ such that for large enough n,*

$$\mathbb{P}_n\left(\widehat{B}_{opt}(n) \in \widetilde{N}_{n,k}(A)\right) \geqslant 1 - b\exp\left\{-\frac{\kappa^2\varepsilon}{32}n\gamma_n^{1-\varepsilon}\log n + \log 4|L|\right\} := 1 - b\beta_n(\varepsilon)$$

Then there exists a constant $A' > 0$ depending only on π_1 and δ such that for large enough n, $\mathbb{P}_n\left(\widehat{B}_{opt}(n) \in \widetilde{N}_{n,k+1}(A')\right) \geqslant 1 - (4+b)\beta_n(\varepsilon)$. The conclusion holds for $k = 1$ if $A = \varepsilon$.

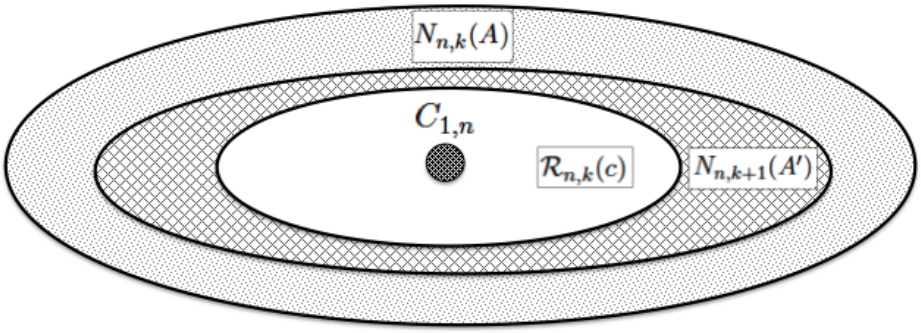

Figure 1: Illustration of relationship between collections of node sets.

Proof. Assume $\pi_1 < \pi_2$; the following argument may be easily adapted to the case where $\pi_1 = \pi_2$, which we explain at the end. Recall $b_{n,k}$ from Definition 10. For $c > 0$, define

$$\mathcal{R}_{n,k}(c) := \{B \subset [n] : |s(B) - \pi_1| \vee [1 - \rho(B)] \leqslant c \cdot b_{n,k}\},$$

Note that sets $B \in \mathcal{R}_{n,k}(c)$ have bounded Hamming distance from $C_{1,n}$, as shown by the following derivation. Writing $s = s(B)$ and $\rho = \rho(B)$, for all $B \in \mathcal{R}_{n,k}(c)$ we have

$$\begin{aligned} n^{-1}|d_h(B, C_{1,n})| &= n^{-1}\big(|B \setminus C_{1,n}| + |C_{1,n} \setminus B|\big) = n^{-1}\big(|B| - |B \cap C_{1,n}| + |C_{1,n}| - |B \cap C_{1,n}|\big) \\ &= s + \pi_1 - 2\rho s \leqslant s + (s + c \cdot b_{n,k}) - 2\,(1 - c \cdot b_{n,k})\, s \\ &= c \cdot b_{n,k} + 2sc \cdot b_{n,k} \leqslant 3c \cdot b_{n,k} \end{aligned} \tag{12}$$

Therefore, $\mathcal{R}_{n,k}(c) \subseteq N(C_{1,n}, A' \cdot b_{n,k}) \subset \widetilde{N}_{n,k+1}(A')$ with $A' = 3c$.

We have assumed $\widehat{B}^{(n)}_{opt}(L) \in \widetilde{N}_{n,k}(A)$ with high probability; our aim is to show $\widehat{B}^{(n)}_{opt}(L) \in \widetilde{N}_{n,k+1}(A')$. Since $\mathcal{R}_{n,k}(c) \subseteq \widetilde{N}_{n,k+1}(A')$, it is sufficient to show that $\widehat{B}^{(n)}_{opt}(L) \notin \widetilde{N}_{n,k}(A) \cap \mathcal{R}_{n,k}(c)^c$ with high probability. This is illustrated by figure 1: since $\widehat{B}^{(n)}_{opt}(L)$ is inside the outer oval (with high probability), it is sufficient to show that it cannot be outside the inner oval. To this end, it is enough to show that, with high probability, $\widehat{H}(B, L) < \widehat{H}(C_{1,n}, L)$ for all sets B in $\widetilde{N}_{n,k}(A) \cap \mathcal{R}_{n,k}(c)^c$. Note that by Lemma 12,

$$\sup_{B \in \widetilde{N}_{n,k}(A)} \widehat{H}(B, L) < H_*(B, L) + 5|L|b_{n,k} \tag{13}$$

for large enough n, with probability at least $1 - 2\beta_n(\varepsilon)$. Next, since $cb_{n,k} \to 0$ as $n \to \infty$, by Lemma 7 there exists a constant $a > 0$ depending just on π_1 such that for large enough n, $B \in \mathcal{R}_{n,k}(c)^c$ implies $H_*(B, L) < H_*(C_{1,n}) - a|L|\phi(L)cb_{n,k}$. Applying Lemma 12 again, we also have $H_*(C_{1,n}, L) < \widehat{H}(C_{1,n}) + 5|L|b_{n,k}$ with probability at least $1 - 2\beta_n(\varepsilon)$. Furthermore,

$\phi(L) \geqslant \delta^2$ by Lemma 6. Applying these inequalities to (13), we get

$$\sup_{B \in \widetilde{N}_{n,k}(A) \cap \mathcal{R}_{n,k}(c)^c} \widehat{H}(B, L) < \widehat{H}(C_{1,n}, L) - a|L|\delta^2 c b_{n,k} + 10|L|b_{n,k} \tag{14}$$

with probability at least $1 - 4\beta_n(\varepsilon)$. With c large enough, (14) implies that $\widehat{H}(B, L) < \widehat{H}(C_{1,n}, L)$ for all $B \in \widetilde{N}_{n,k}(A) \cap \mathcal{R}_{n,k}(c)^c$. This proves the result in the $\pi_1 < \pi_2$ case.

If $\pi_1 = \pi_2$, the argument is almost identical. We instead define $\mathcal{R}_{n,k}(c)$ as

$$\mathcal{R}_{n,k}(c) := \{B \subseteq [n] : |s(B) - \pi_1| \vee \rho(B) \vee [1 - \rho(B)] \leqslant c \cdot b_{n,k}\}.$$

A derivation analogous to that giving inequality (12) yields

$$n^{-1} \left(d_h(B, C_{1,n}) \vee d_h(B, C_{2,n})\right) \leqslant 3c \cdot b_{n,k}$$

which directly implies that $\mathcal{R}_{n,k}(c) \subseteq \widetilde{N}_{n,k+1}(A')$ with $A' = 3c$. The rest of the argument goes through unaltered. ■

4.1.3 Proof of Theorem 2

Recall $Q_\ell(B)$ from Definition 2 and let $\widehat{Q}_\ell(B)$ be its random version under the MSBM, as in Section 3.2. For any $B \subseteq [n]$, we have the inequality

$$\left[\widehat{Q}_\ell(B)\right]_+ \leqslant \frac{Y_\ell(B)}{n\binom{|B|}{2}^{1/2}} \leqslant \frac{\binom{|B|}{2}}{n\binom{|B|}{2}^{1/2}} \leqslant \frac{|B|}{n} \tag{15}$$

This yields the following inequality for $\widehat{H}(B, L)$:

$$\widehat{H}(B, L) = |L|^{-1} \left[\left(\sum_{\ell \in [L]} Q_\ell(B)\right)_+\right]^2 \leqslant |L|^{-1} \left[\sum_{\ell \in [L]} Q_\ell(B)_+\right]^2 \leqslant |L|^{-1} n^{-2} |B|^2 \tag{16}$$

Recall that $\mathcal{B}_n(\varepsilon) := \{B \in 2^{[n]} : |B| \geqslant \varepsilon n\}$. Inequality (16) implies $\widehat{H}(B, L) \leqslant |L|\varepsilon^2$ for all $B \in \mathcal{B}_n(\varepsilon)^c$. By part 2 of Lemma 6, $\phi(L) \geqslant \delta^2$. Therefore, defining $\tau := \frac{\pi_1^2}{2}(1 - \pi_1)^4 \delta^2 / 2$,

$$|L|\tau < |L|\phi(L)\frac{\pi_1^2}{2}(1 - \pi_1)^4 = H_*(B, L)$$

Therefore, for all $B \in \mathcal{B}_n(\varepsilon)^c$, we have $\widehat{H}(B, L) \leqslant |L|\varepsilon^2 < H_*(C_{1,n}, L) - |L|(\tau - \varepsilon^2)$. By Lemma 12, for large enough n we therefore have

$$\sup_{\mathcal{B}_n(\varepsilon)^c} \widehat{H}(B, L) < \widehat{H}(C_{1,n}, L) - |L|(\tau - \varepsilon^2) + 5|L|\gamma_n^{1-\varepsilon} \tag{17}$$

with probability greater than $1 - 2\beta_n(\varepsilon)$, where $\beta_n(\varepsilon) := \exp\{-\frac{\kappa^2 \varepsilon}{32} n \gamma_n^{1-\varepsilon} \log n + \log 4|L|\}$. For any $\varepsilon < \sqrt{\tau}$, inequality (17) implies $\widehat{H}(B, L) < \widehat{H}(C_{1,n}, L)$ for all $B \in \mathcal{B}_n(\varepsilon)$, and

therefore $\widehat{B}_{opt}(n) \in \mathcal{B}_n(\varepsilon)$, with probability at least $1 - 2\beta_n(\varepsilon)$. Note that $\varepsilon < \sqrt{\tau} < \pi_1$, and $N_{n,k}(\varepsilon) = \mathcal{B}_n(\varepsilon)$ by Definition 11. Therefore, the conditions for Lemma 13 with $k = 1$ (and $A = \varepsilon$) are satisfied. For any fixed $\varepsilon \in (0, \eta)$ with $\eta := \sqrt{\tau}$, we may now apply Lemma 13 recursively until $1/2^k \leqslant \varepsilon$. This establishes that for sufficiently large n,

$$\mathbb{P}_n\left(\widehat{B}_{opt}(n) \in \widetilde{N}_{n,k}(A)\right) \geqslant 1 - (2 + 4k)\beta_n(\varepsilon) \tag{18}$$

By definition, $\widehat{B}_{opt}(n) \in \widetilde{N}_{n,k}(A)$ implies that

$$\texttt{Error}(\widehat{B}_{opt}(n)) := \min_{C=C_1,C_2} d_h(\widehat{B}_{opt}(n), C) \leqslant A \cdot n \cdot b_{n,k}. \tag{19}$$

Note that

$$n \cdot b_{n,k} = n\gamma_n^{1-\frac{1}{2^k}} = n \cdot n^{\frac{1}{2^k}-1}(\log n)^{1-\frac{1}{2^k}} < n^{\varepsilon} \log n$$

since $1/2^k \leqslant \varepsilon$. Combined with inequality (18), this completes the proof. ■

4.2 Proof of Theorem 3

To prove part (a), we first note that Theorem 2 implies that on the layer set L_1, for any $\varepsilon > 0$, $\texttt{Error}(\widehat{B}_{opt}^{(n)}) = O_p(n^{\varepsilon} \log n)$. Lemma 6 can be used to show that $H_*(B, L) = 0$ for any $L \subseteq L_0$ and any $B \subseteq [n]$. Using Lemma 9 and taking a union bound over L_0, it is then straightforward to show (using techniques from the proof of Theorem 2) that on the full layer set $[m]$, for any $\varepsilon > 0$, $\texttt{Error}(\widehat{B}_{opt}) = O_p(n^{\varepsilon} \log n)$. Considering now $\widehat{L}_{opt}^{(n)}$, observe that if $\widehat{Q}_{\ell}(B) \leqslant 0$, then $\widehat{H}(B, L) = \widehat{H}(B, L \setminus \{\ell\})$. This immediately implies that $\widehat{L}_{opt}^{(n)} = \widehat{L}^{+}$.

To prove part (b), we note that it is straightforward to show (using Lemma 6) that $H_*(B, L_1) \geqslant H_*(B, L)$ for any $L \subset [m]$, with equality if and only if $L = L_1$. Using Lemma 9 and a union bound over $[m]$ will show that $\widehat{L}_{opt}^{(n)} = L_1$ with high probability. Applying Theorem 2 completes the part. Part (c) is shown similarly, with the application of Lemma 6 showing that for any $L \subseteq L_1$ and $L' \subseteq [m]$, $H_*(B, L) \geqslant H_*(B, L')$, with equality if and only if $L' \subseteq L_1$. ■

5 The Multilayer Extraction Procedure

The Multilayer Extraction procedure is built around three operations: initialization, extraction, and refinement. In the initialization stage, a family of seed vertex sets is specified. Next an iterative extraction procedure (**Extraction**) is applied to each of the seed sets. **Extraction** alternately updates the layers and vertices in a vertex-layer community in a greedy fashion, improving the score at each iteration, until no further improvement to the score is possible. The family of extracted vertex-layer communities is then reduced using the **Refinement** procedure, which ensures that the final collection of communities contains the extracted community with largest score, and that the pairwise overlap between any pair of communities is at most β, where $\beta \in [0, 1]$ is a user-defined parameter. The importance and relevance of this parameter is discussed in Section 5.3.1. We describe the Multilayer Extraction algorithm in more detail below.

5.1 Initialization

For each vertex $u \in [n]$ and layer $\ell \in [m]$ let $N(u, \ell) = \{v \in [n] : \{u, v\} \in E_\ell\}$ be the set of vertices connected to u in G_ℓ. We will refer to $N(u, \ell)$ as the neighborhood of u in layer ℓ. Let $\mathcal{B}_0 = \{N(u, \ell), u \in [n], \ell \in [m]\}$ be the family of all vertex neighborhoods in the observed multilayer network $\mathbf{G}(m, n)$. Multilayer Extraction uses the vertex sets in $\mathcal{B}_0$ as seed sets for identifying communities. Our choice of seed sets is motivated by Gleich and Seshadhri [23], who showed that vertex neighborhoods are optimal seed sets for local detection methods seeking communities with low conductance.

5.2 Extraction

Given an initial vertex set, the **Extraction** procedure seeks a vertex-layer community with large score. The algorithm iteratively conducts a *Layer Set Search* followed by a *Vertex Set Search*, and repeats these steps until a vertex-layer set, whose score is a local maximum, is reached. In each step of the procedure, the score of the candidate community strictly increases, and the procedure is stopped once no improvements to the score are possible. These steps are described next.

Layer Set Search: For a fixed vertex set $B \subseteq [n]$, **Extraction** searches for the layer set L that maximizes $H(B, \cdot)$ using a rank ordering of the layers that depends only on B. In particular let $Q_\ell(B)$ be the local set modularity of layer ℓ from (2). Let L_o be the layer set identified in the previous iteration of the algorithm. We will now update the layer set $L_0 \rightsquigarrow L$. This consists of the following three steps:

(i) Order the layers so that $Q_{\ell_1}(B) \geqslant \cdots \geqslant Q_{\ell_m}(B)$.

(ii) Identify the smallest integer k such that $H(B, \{\ell_1, \ldots, \ell_k\}) \geqslant H(B, \{\ell_1, \ldots, \ell_k, \ell_{k+1}\})$. Write $L_p := \{\ell_1, \ldots, \ell_k\}$ for the proposed change in the layer set.

(iii) *If* $H(B, L_p) > H(B, L_o)$ set $L = L_p$. *Else* set $L = L_o$

In the first iteration of the algorithm (where we take $L_o = \emptyset$), we set $L = L_p$ in step (iii) of the search. The selected layer set L_p is a local maximum for the score $H(B, \cdot)$.

Vertex Set Search: Suppose now that we are given a vertex-layer set (B, L). **Extraction** updates B, one vertex at a time, in a greedy fashion, with updates depending on the layer set L and the current vertex set. In detail, for each $u \in [n]$ let

$$\delta_u(B, L) = \begin{cases} H(B/\{u\}, L) - H(B, L) & \text{if } \ u \in B \\ H(B \cup \{u\}, L) - H(B, L) & \text{if } \ u \notin B. \end{cases} \tag{20}$$

Vertex Set Search iteratively updates B using the following steps:

(i) Calculate $\delta_u(B, L)$ for all $u \in [n]$. If $\delta_u(B, L) \leqslant 0$ for all $u \in [n]$, then stop. Otherwise, identify $u^* = \arg\max_{u \in [n]} \delta_u(B, L)$.

(ii) If $u^* \in B$, then remove u^* from B. Otherwise, add u^* to B.

At each iteration of **Extraction**, the score of the updated vertex-layer set strictly increases, and the eventual convergence of this procedure to a local maximum is guaranteed as the possible search space is finite. The resulting local maxima is returned as an extracted community.

5.3 Refinement

Beginning with the n vertex neighborhoods in each layer of the network, the **Extraction** procedure identifies a collection $\mathcal{C}_T = \{(B_t, L_t)\}_{t \in T}$ of at most $m*n$ vertex-layer communities. Given an overlap parameter $\beta \in [0, 1]$, the family $\mathcal{C}_T$ is refined in a greedy fashion, via the **Refinement** procedure, to produce a subfamily $\mathcal{C}_S$, $S \subseteq T$, of high-scoring vertex-layer sets having the property that the overlap between any pair of sets is at most β.

To quantify overlap, we specify a generalized Jaccard match score to measure overlap between two communities. We measure the overlap between two candidate communities (B_q, L_q) and (B_r, L_r) using a generalized Jaccard match score

$$J(q, r) \;=\; \frac{1}{2}\frac{|B_q \cap C_r|}{|B_q \cup C_r|} + \frac{1}{2}\frac{|L_q \cap L_r|}{|L_q \cup L_r|} \tag{21}$$

It is easy to see that $J(q, r)$ is between 0 and 1. Moreover, $J(q, r) = 1$ if and only if $(B_q, L_q) = (B_r, L_r)$ and $J(q, r) = 0$ if and only if (B_q, L_q) and (B_r, L_r) are disjoint. Larger values of $J(\cdot, \cdot)$ indicate more overlap between communities.

In the first step of the procedure, **Refinement** identifies and retains the community (B_s, L_s) in $\mathcal{C}_T$ with the largest score and sets $S = \{s\}$. In the next step, the procedure identifies the community (B_s, L_s) with largest score that satisfies $J(s, s') \leqslant \beta$ for all $s' \in S$. The index s is then added to S. **Refinement** continues expanding S in this way until no further additions to S are possible, namely when for each $s \in T$, there exists an $s' \in S$ such that $J(s, s') > \beta$. The refined collection $C_S = \{B_s, L_s\}_{s \in S}$ is returned.

5.3.1 Choice of β

Many existing community detection algorithms have one or more tunable parameters that control the number and size of the communities they identify [50; 29; 32; 28; 54]. The family of communities output by Multilayer Extraction depends on the overlap parameter $\beta \in [0, 1]$. In practice, the value of β plays an important role in the structure of the vertex-layer communities. For instance, setting $\beta = 0$ will provide vertex-layer communities that are fully disjoint (no overlap between vertices or layers). On the other hand, when $\beta = 1$ the procedure outputs the full set of extracted communities, many of which may be redundant. In exploratory applications, we recommend investigating the identified communities at multiple values of β, as the structure of communities at different resolutions may provide useful insights about the network itself (see for instance Leskovec et al. [29] or Mucha et al. [32]).

Empirically, we observe that the number of communities identified by the Multilayer Extraction procedure is non-decreasing with β, and there is typically a long interval of β values over which the number and identity of communities remains constant. In practice we specify a default value of β by analyzing the number of communities across a grid of

β between 0 and 1 in increments of size 0.01. For fixed i, let $\beta_i = (i-1) * 0.01$ and let $k_i = k(\beta_i)$ denote the number of communities identified at β_i. The default value β' is the smallest β value in the longest stable window, namely

$$\beta' \;=\; \text{smallest } \beta_i \text{ such that } k(\beta_i) = \text{mode}(k_1, \ldots, k_{101})$$

6 Application Study

In this section, we assess the performance and potential utility of the Multilayer Extraction procedure through an empirical case study of three multilayer networks. We compare and contrast the performance of Multilayer Extraction with four benchmark methods: Spectral Clustering [35], Label Propagation [44], Fast and Greedy [14], and Walktrap [42]. Each of these methods have publicly available implementations in the *igraph* package in R and in Python, and each method is a standard single-layer detection method that can handle weighted edges. We apply these methods to both the aggregate (weighted) network computed from the average of the layers in the analyzed multilayer network, and to each layer separately. We note that we do not compare Multilayer Extraction with the multilayer methods described in Section 1.1 due to the fact that the communities identified by those methods are difficult to compare directly with Multilayer Extraction. A more detailed description of the competing methods and their parameter settings is provided in the Appendix. For this analysis and the subsequent analysis in Section 7, we set Multlayer Extraction to identify vertex-layer communities that have a large significance score as specified by equation (2).

For each method we calculate a number of quantitative features, including the number and size of the identified communities, as well as the number of identified background vertices. We also evaluate the similarity of communities identified by each method. As aggregate and layer-by-layer methods do not provide informative layer information, we compare the vertex sets identified by each of the competing methods with those identified by Multilayer Extraction. To this end, for two vertex sets $\mathcal{B}, \mathcal{C} \subseteq 2^{[n]}$ define the coverage of $\mathcal{B}$ by $\mathcal{C}$ as

$$C(\mathcal{B}; \mathcal{C}) = |\mathcal{B}|^{-1} \sum_{B \in \mathcal{B}} \max_{C \in \mathcal{C}} \left(\frac{|B \cap C|}{|B \cup C|} \right) \tag{22}$$

The coverage $C(\mathcal{B}; \mathcal{C})$ quantifies the extent to which vertex sets in $\mathcal{B}$ are contained *in* $\mathcal{C}$. In general, $C(\mathcal{B}; \mathcal{C}) \neq C(\mathcal{C}; \mathcal{B})$. The coverage value $C(\mathcal{B}; \mathcal{C})$ is between 0 and 1, with $C(\mathcal{B}; \mathcal{C}) = 1$ if and only if $\mathcal{B}$ is a subset of $\mathcal{C}$.

We investigate three multilayer networks of various size, sparsity, and relational types: a social network from an Austrailian Computer Science department [25]; an air transportation network of European airlines [10]; and a collaboration network of network science authors on arXiv.org [15]. The size and edge density of each network is summarized in Table 1.

Figure 2 provides a summary of the quantitative features of the communities identified by each method. The coverage between each family of identified communities is shown in Figure 3. For ease of discussion, we will abbreviate Multilayer Extraction by M-E in the next two sections of the manuscript, where we evaluate the performance of the method on real and simulated data.

Network	# Layers	# Vertices	Total # Edges
AU-CS	5	61	620
EU Air Transport	36	450	3588
arXiv	13	14489	59026

Table 1: Summary of the real multilayer networks in our study.

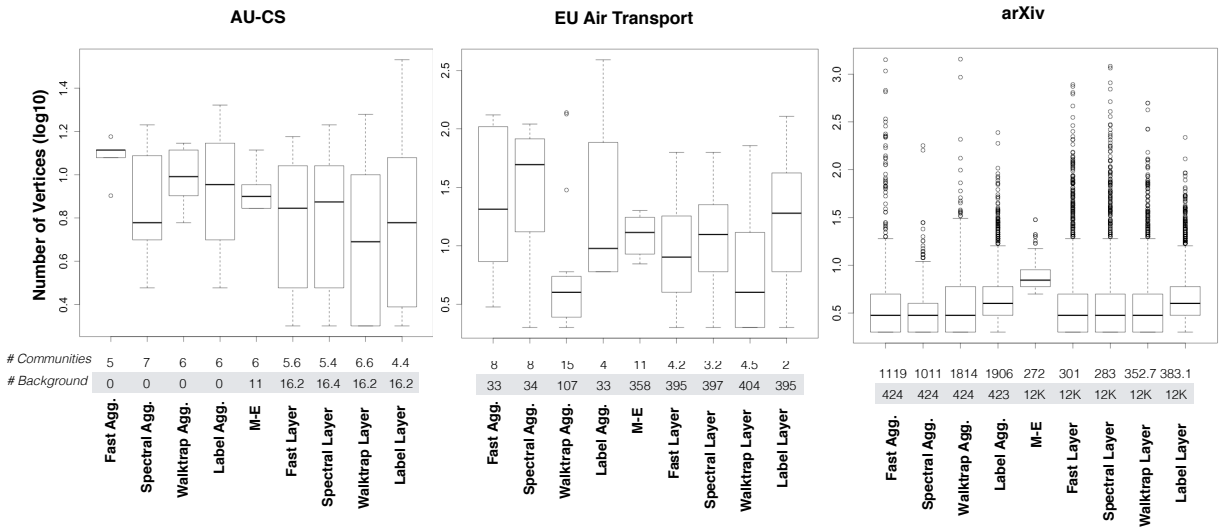

Figure 2: Quantitative summary of the identified communities in each of the three real applications. Boxplots represent the number of vertices ($\log_{10}$) in identified communities for each method. For the layer-by-layer methods, we report the average number of communities and background vertices for each layer.

6.1 AU-CS Network

The AU-CS network describes online and offline relationships of 61 employees of a Computer Science research department in Australia. The vertices of the network represent the employees in the department. The layers of the network represent five different relationships among the employees: *Facebook*, *leisure*, *work*, *co-authorship*, and *lunch*.

Results

M-E identified 4 communities, which are illustrated in Figure 4 **a**. These communities reveal several interesting patterns in the network. Both the *work* and *lunch* layers were contained in all four of the identified communities, reflecting a natural co-occurrence of work and lunch interactions among the employees. Furthermore, two of the identified communities contained the *leisure* and *Facebook* layers, both of which are social activities that sometimes extend beyond work and lunch. These interpretable, and perhaps expected, features of this social network were directly identified by M-E. None of the competing methods were able to identify these types of layer features.

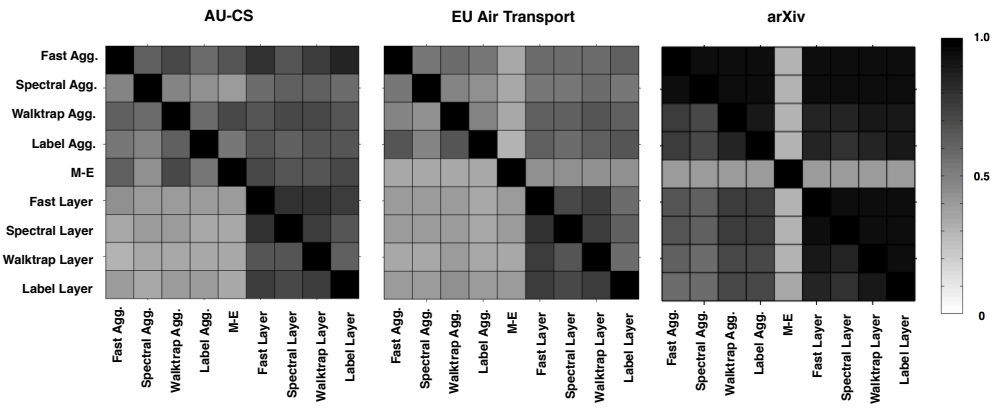

Figure 3: A heat map illustrating the coverage of communities identified by each of the methods applied to the real multilayer networks. Coverage is assessed using the measure in (21).

All of the methods identify a similar number of communities (ranging from 4 to 7). The 37 background vertices identified by M-E were sparsely connected, having two or fewer connections in 3 of the layers. The vertex sets identified by M-E were similar to those identified by the single-layer methods. Furthermore, Figure 3 reveals that the communities identified by the aggregate approaches are in fact well contained in the family identified by M-E (average coverage = 0.78). In summary, the vertex sets identified by M-E reflect both the aggregate and separate layer community structure of the network, and the layer sets reveal important features about the social relationships among the employees that could not have been easily identified using existing methods.

6.2 European Air Transportation Network

Vertices in the European Air Transportation Network represent 450 airports in Europe and layers represent 37 different airlines. An edge in layer j is present between two airports if airline j traveled a direct flight between the two airports on June 1st, 2011. Each airline belongs to one of five classes: *major* (18); *low-cost* (10); *regional* (6); *cargo* (2); and *other* (1). A multiplex visualization of this network is shown in Figure 5.

Results

M-E identified 11 small communities (mean number of vertices = 11.82, mean number of layers = 3.73). Both the single-layer methods and M-E identified on the order of 400 background vertices ($\approx 89\%$). This suggests that the airlines follow distinct routes containing a small number of unique airports. Aggregate detection approaches identified a similar number of communities as M-E. However, the sizes of the identified communities were twice as large on average (mean = 24.10), and the aggregate methods identified far fewer background vertices.

The layer sets of the extracted M-E communities are illustrated in Figure 4 **b**, which shows that the layers of each community are closely associated with airline classes. Indeed,

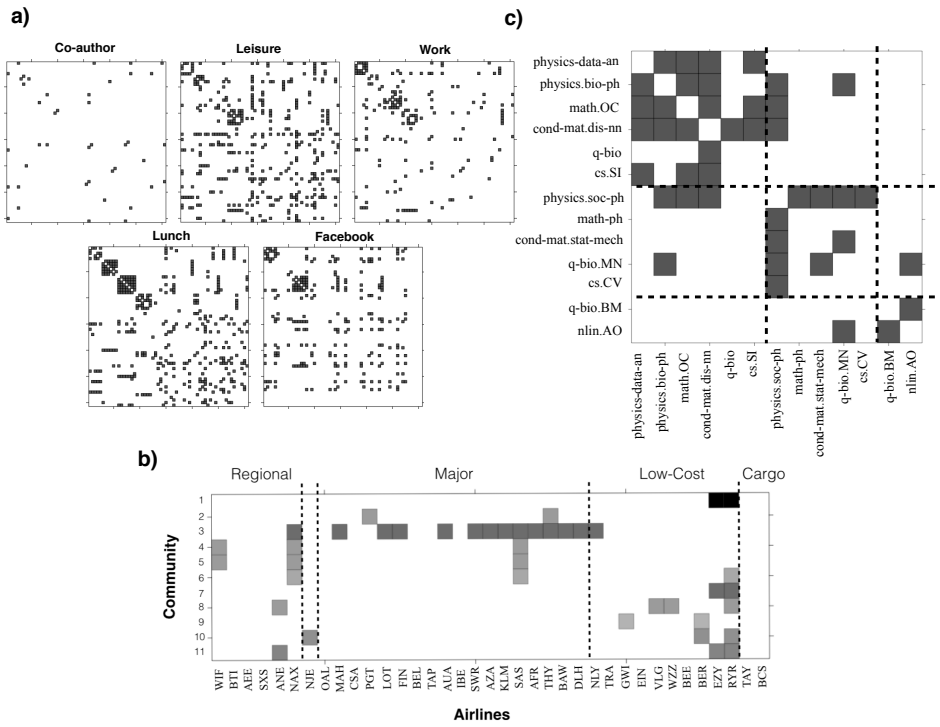

Figure 4: **a)** The AU-CS multilayer network. The vertices have been reordered based on the four communities identified by M-E. **b)** The layers of the eleven extracted significant communities identified in the EU transport network. The layers are ordered according to the type of airline. The darkness of the shaded in blocks represents the score of the identified community. **c)** Adjacency matrix of layers in the arXiv network, where edges are placed between layers that were contained in one or more of the communities identified by M-E. Dotted lines separate three communities of submission types that were identified using spectral clustering.

an average of 78 % of the layers in each community belonged to the same airline class, reflecting the fact that airlines of the same class tend to have direct flights between similar airports. Interestingly, the *regional* airline Norwegian Air Shuttle (NAX) and the *major* airline Scandanavian Airlines (SAS) appeared together in 4 unique communities. These airlines are in fact the top two air carriers in Scandanavia and fly primarily to airports in Norway, Sweden, Denmark, and Finland.

6.3 arXiv Network

The arXiv network of De Domenico et al. [15] represents the authors of all arXiv submissions that contained the word "networks" in its title or abstract between the years 2010 and 2012. The network has 14489 vertices representing authors, and 13 layers representing the arXiv category under which the submission was placed. An edge is placed between two authors in layer ℓ if they co-authored a paper placed in that category. The network is sparse, with each layer having edge density less than 1.5%.

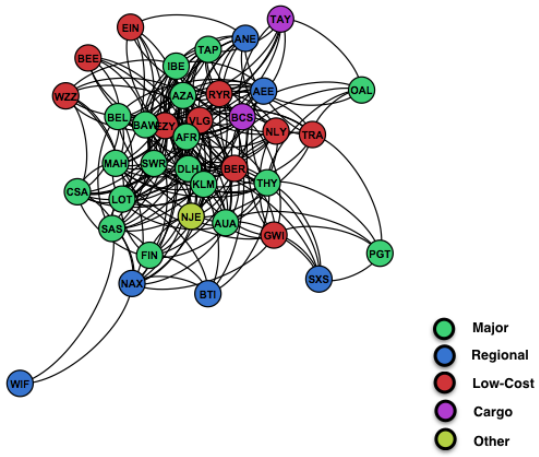

Figure 5: A one-dimensional visualization of the European Air Transportation Network. Edges are placed between airlines that share at least two routes between airports.

Results

M-E identified 272 multilayer communities, with an average of 2.39 layers per community. The communities were small in size, suggesting that network science collaboration groups are relatively tightly-knit. In Figure 4 **c**, we plot an adjacency matrix for layers whose (i, j) entry is 1 if and only if layers i and j were contained in at least one multilayer community. Using the adjacency matrix, the layers of the network were partitioned into communities using Spectral clustering. The results support the existence of three active interdisciplinary working groups among the selected researchers. Though directly identified by M-E, these interdisciplinary groups would be difficult to identify without using such a method.

Aggregate approaches identify on the order of 1000 small to moderately sized communities (mean number of vertices between 6.04 and 9.80) with approximately 400 background vertices. On the other hand, M-E and the single layer approaches identify a smaller number of communities (between 272 and 383), and classify about 12 thousand (roughly 86%) of the vertices as background. These findings suggest that the individual layers of the arXiv network have heterogeneous community structure, and that they contain many non-preferentially attached vertices. Figure 3 illustrates the discrepancy between the aggregate and single layer methods. In this application, Multilayer Extraction identifies community structure that is substantially different than any competing method, and the layers identified by the procedure highlight interesting and non-trivial features of these arXiv collaborations.

7 Simulation Study

As noted above, Multilayer Extraction has three key features: it allows community overlap; it identifies background; and it can identify communities that are present in a small subset of the available layers. Below we describe a simulation study that aims to evaluate the performance of M-E with respect to these features. The results of additional simulations

are described in the Appendix.

We compare the performance of M-E with layer-by-layer and aggregate approaches using the community detection methods described in Section 6. Define the *match* between two vertex families $\mathcal{B}$ and $\mathcal{C}$ by

$$M(\mathcal{B};\mathcal{C}) \;=\; \frac{1}{2}\,O(\mathcal{B};\mathcal{C}) \;+\; \frac{1}{2}\,O(\mathcal{C};\mathcal{B}), \tag{23}$$

where $O(\mathcal{B};\mathcal{C})$ is the coverage measure for vertex families from (22). The match $M(\mathcal{B};\mathcal{C})$ is symmetric and takes values in $[0,1]$. In particular, $M(\mathcal{B};\mathcal{C}) = 1$ if and only if $\mathcal{B} = \mathcal{C}$ and $M(\mathcal{B};\mathcal{C}) = 0$ if and only if $\mathcal{B}$ and $\mathcal{C}$ are disjoint.

In our simulations, we compute the match between the family of vertex sets identified by each method and the family of true simulated communities. For the layer-by-layer methods, we evaluate the average match of the communities identified in each layer. Suppose that $\mathcal{T}$ is the true family of vertex sets in a simulation and $\mathcal{D}$ is the family of vertex sets identified by a detection procedure of interest. Note that the value $O(\mathcal{D};\mathcal{T})$ quantifies the specificity of $\mathcal{D}$, while $O(\mathcal{T};\mathcal{D})$ quantifies its sensitivity; thus, $M(\mathcal{D};\mathcal{T})$ is a quantity between 0 and 1 that summarizes both the sensitivity and specificity of the identified vertex sets $\mathcal{D}$. The results of the simulation study are summarized in Figure 6, and discussed in more detail below.

7.1 Multilayer Stochastic Block Model

In the first part of the simulation study we generated multilayer stochastic block models with $m \in \{1, 5, 10, 15\}$ layers, $k \in \{2, 5\}$ blocks, and $n = 1000$ vertices such that each layer has the same community structure. In more detail, each vertex is first assigned a community label $\{1, \ldots, k\}$ according to a probability mass function $\pi = (0.4, 0.6)$ for $k = 2$ and $\pi = (0.2, 0.1, 0.2, 0.1, 0.4)$ for $k = 5$. In each layer, edges are assigned independently, based on vertex community membership, according to the probability matrix P with entries $P(i,i) = r + 0.05$ and $P(i,j) = 0.05$ for $i \neq j$. Here r is a parameter representing connectivity strength of vertices within the same community. The resulting multilayer network consists of m independent realizations of a stochastic k block model with the same communities. For each value of m and k we vary r from 0.00 to 0.10 in increments of 0.005. M-E and all other competing methods are run on ten replications of each simulation. The average match of each method to the true communities is given in the left panel of Figure 6.

Results

In the single-layer ($m = 1$) setting M-E is competitive with the existing single-layer methods for $r \geqslant 0.05$, and identifies the true communities without error for $r \geqslant 0.06$. For $m \geqslant 5$ M-E outperforms all competing single-layer methods for $r \geqslant 0.02$. As the number of layers increases, M-E exhibits improved performance across all values of r. As expected, aggregate approaches perform well in this simulation, outperforming or matching other methods when $m \leqslant 5$ (results not shown). These results suggest that in homogeneous multilayer networks M-E can outperform or match existing methods when the network contains a moderate to large number of layers.

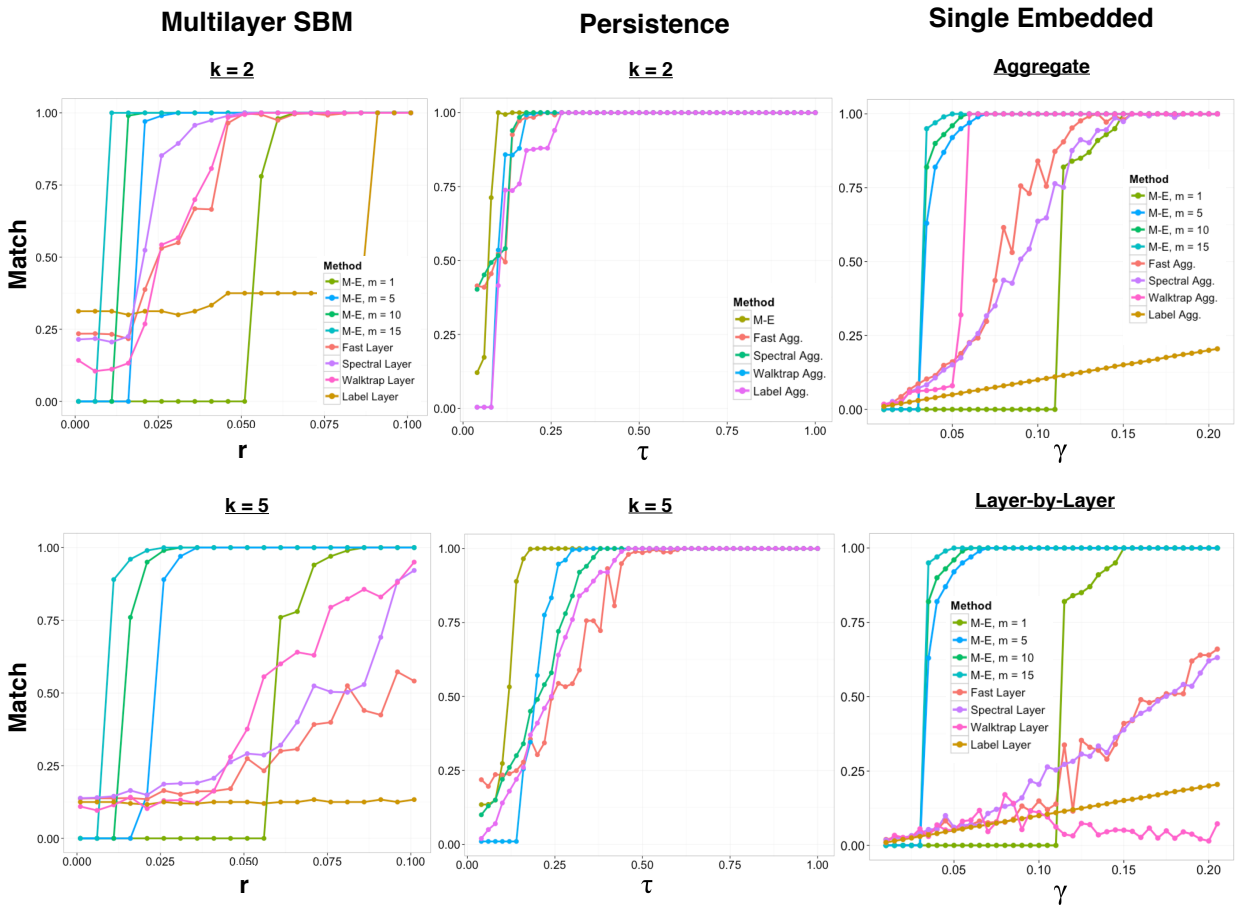

Figure 6: (Color) Simulation results for multilayer stochastic block model, persistence, and single embedded simulations. In each plot, we report the match of the identified communities with the true communities where the match is calculated using the match score in (23).

7.2 Persistence

In the second part of the simulation study we consider multilayer networks with heterogeneous community structure. We simulated networks with 50 layers and 1000 vertices. The first $\tau * 50$ layers follow the stochastic block model outlined in Section 7.1 with a fixed connection probability matrix P having entries $P(i,i) = 0.15$ and $P(i,j) = 0.05$ for $i \neq j$. The remaining $(1-\tau)*50$ layers are independent Erdős-Rényi random graphs with $p = 0.10$, so that in each layer every pair of vertices is connected independently with probability 0.10. For each $k \in \{2, 5\}$ we vary the persistence parameter τ from 0.02 to 1 in increments of 0.02, and for each value of τ, we run M-E as well as the competing methods on ten replications. The average match of each method is reported in the center panel of Figure 6.

Results

In both block model settings with $k = 2$ and 5 communities, M-E outperforms competing aggregate methods for small values of τ. At these values, aggregate methods perform poorly since the community structure in the layers with signal is hidden by the noisy Erdős-Rényi layers once the layers are aggregated. Though not shown in Figure 6, the layer-by-layer methods are able to correctly identify the community structure of the layers with signal. However, these methods identify on average of 4 or more non-trivial communities in each noisy layer where there is in fact no community structure present. Whereas the noisy Erdős-Rényi layers posed a challenge for all other competing methods, M-E never included any of these layers in an identified community. These results highlight M-E's ability to handle networks with noisy and heterogeneous layers.

7.3 Single Embedded Communities

We next evaluate the ability of M-E to detect a single embedded community in a multilayer network. We construct multilayer networks with $m \in \{1, 5, 10, 15\}$ layers and 1000 vertices according to the following procedure. Each layer of the network is generated by embedding a common community of size $\gamma * 1000$ in an Erdős-Rényi random graph with connection probability 0.05 in such a way that vertices within the community are connected independently with probability 0.15. The parameter γ is varied between 0.01 and 0.20 in increments of 0.005; ten independent replications of the embedded network are generated for each γ. For each method, we calculate the coverage $C(E, \mathcal{C})$ of the true embedded community E by the identified collection $\mathcal{C}$. We report the average coverage over the ten replications in the right panel of Figure 6.

Results

In the single layer setting, M-E is able to correctly identify the embedded community when the embedded vertex set takes up approximately 11 percent ($\gamma = 0.11$) of the layer. As before, the performance of M-E greatly improves as the number of layers in the observed multilayer network increases. For example at $m = 5$ and $m = 10$, the algorithm correctly identifies the embedded community (with at least 90% match) once the community has size taking as little as 6 percent ($\gamma = 0.055$) of each layer. At $m = 15$, M-E correctly extracts communities with size as small as three percent of the graph in each layer.

In the upper right plot of Figure 6, we illustrate the results for the aggregate methods applied to the simulated network with $m = 15$ layers. In the lower right plot of Figure 6, we show the results for the layer-by-layer methods in this simulation. For $m \geqslant 5$ M-E outperforms all of the aggregate methods. In addition, M-E outperforms every layer-by-layer method for all m. These results emphasize the extraction cabilities of M-E and show that the procedure, contrary to the competing methods, is able to detect small embedded communities in the presence of background vertices.

8 Discussion

Multilayer networks have been profitably applied to a number of complex systems, and community detection is a valuable exploratory technique to analyze and understand networks. In many applications, the community structure of a multilayer network will differ from layer to layer due to heterogeneity. In such networks, actors interact in tightly connected groups that persist across only a subset of layers in the network. In this paper we have introduced and evaluated the first community detection method to address multilayer networks with heterogeneous communities, Multilayer Extraction. The core of Multilayer Extraction is a significance score that quantifies the connection strength of a vertex-layer set by comparing connectivity in the observed network to that of a fixed degree random network.

Empirically, we showed that Multilayer Extraction is able to successfully identify communities in multilayer networks with overlapping, disjoint, and heterogeneous community structure. Our numerical applications revealed that Multilayer Extraction can identify relevant insights about complex relational systems beyond the capabilities of existing detection methods. We also established consistency of Multilayer Extraction under the multilayer stochastic block model under both regimes in which the number of layers was fixed or allowed to grow exponentially in n^{ϵ}.

The Multilayer Extraction method provides a first step in understanding and analyzing multilayer networks with heterogeneous community structure. This work encourages several interesting areas of future research. For instance, the techniques used in this paper could be applied, with suitable models, to networks having ordered layers (e.g. temporal networks), as well as to networks with weighted edges such as the recent work done in Palowitch et al. [38]. Furthermore, one could incorporate both node- and layer-based covariates in the null model to handle exogenous features of the multilayer network. Finally, it would be interesting to evaluate the consistency of Multilayer Extraction in multilayer networks in the high dimensional setting where the number of vertex-layer communities grows with the number of vertices.

A Proofs of Lemmas from Section 3.2

A.1 Proof of Lemma 6

It is easy to show that for any 2×2 symmetric matrix A and 2-vectors $\mathbf{x}$, $\mathbf{y}$,

$$(\mathbf{x}^T A\mathbf{x})(\mathbf{y}^T A\mathbf{y}) - (\mathbf{x}^T A\mathbf{y})^2 = (\mathbf{x}_1\mathbf{y}_2 - \mathbf{x}_2\mathbf{y}_1)^2 \det(A)$$

Fix $B \subseteq [n]$ and let s, ρ, and v correspond to B, as in Definition 4. Then for any $\ell \in [L]$, using the fact that $\kappa_\ell := \pi^T P_\ell \pi$ and the identity above, we have

$$v^t P_\ell v - \frac{(\pi^t P_\ell \pi)^2}{\kappa_\ell} = \frac{\kappa_\ell v^t P_\ell v}{\kappa_\ell} - \frac{(v^t P_\ell \pi)^2}{\kappa_\ell} = \frac{(\pi^t P_\ell \pi)(v^t P_\ell v) - (v^t P_\ell \pi)^2}{\kappa_\ell}$$

$$= (\pi_1(1-\rho) - \pi_2\rho)^2 \frac{\det P_\ell}{\kappa_\ell} = (\pi_1 - \rho)^2 \frac{\det P_\ell}{\kappa_\ell}$$

Recall that $q_\ell(B) := \frac{s}{\sqrt{2}}\left(v^t P_\ell v - (\pi^t P_\ell \pi)^2/\kappa_\ell\right)$ and $H_*(B, L) = |L|^{-1}\left(\sum_\ell q_\ell(B)\right)^2$. Part 1 follows by summation over L. For part 2, note that $\pi_1^2 P_\ell(1,1) + \pi_2^2 P(2,2) \geqslant 2\pi_1\pi_2\sqrt{P_\ell(1,1)P_\ell(2,2)}$. Therefore,

$$\begin{aligned}
\kappa_\ell &= \pi_1^2 P_\ell(1,1) + 2\pi_1\pi_2 P_\ell(1,2) + \pi_2^2 P(2,2) \\
&\geqslant 2\pi_1\pi_2\left(\sqrt{P_\ell(1,1)P_\ell(2,2)} + P_\ell(1,2)\right) \\
&\geqslant 2\pi_1\pi_2\left(\sqrt{P_\ell(1,1)P_\ell(2,2)} + P_\ell(1,2)\right)\left(\sqrt{P_\ell(1,1)P_\ell(2,2)} - P_\ell(1,2)\right) \\
&= 2\pi_1\pi_2\delta \geqslant \pi_1\delta
\end{aligned}$$

Thus $\delta \leqslant \frac{\det P_\ell}{\kappa_\ell} \leqslant \frac{1}{\pi_1\delta}$. Part 2 follows. ■

A.2 Proof of Lemma 7

Define $g : 2^{[n]} \mapsto \mathbb{R}$ by $g(B) := \frac{s(B)}{2}(\rho(B) - \pi_1)^4$. Recall the function $\phi(L)$ defined in Lemma 6. Note that part 1 of Lemma 6 implies $H_*(B, L) = |L|\phi(L)g(B)$. It is therefore sufficient to show that there exists a constant $a > 0$ such that for sufficiently small t, $B \in \mathcal{R}(t)^c$ implies $g(B) < g(C_{1,n}) - at$. We will show this separately for the $\pi_1 < \pi_2$ and $\pi_1 = \pi_2$ cases.

Part 1 ($\pi_1 < \pi_2$): Define the intervals $I_1 := [0, \pi_1]$, $I_2 := (\pi_1, \pi_2]$, and $I_3 := (\pi_2, 1]$. We trisect $2^{[n]}$, the domain of g, with the collections $\mathcal{D}_{i,n} := \{B \subseteq [n] : s(B) \in I_i\}$, for $i = 1, 2, 3$. We will prove that the inequality $g(B) < g(C_{1,n}) - at$ holds for all $B \in \mathcal{R}(t)$ on each of those collections. We will continually rely on the fact that $B \in \mathcal{R}(t)$ implies at least one of the inequalities (I) $|s(B) - \pi_1| > t$ or (II) $1 - \rho(B) < t$ is true.

Suppose $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{1,n}$ and inequality (I) is true. Then $s(B) < \pi_1 - t$, and

$$\begin{aligned}
g(B) := \frac{s(B)^2}{2}(\rho(B) - \pi_1)^4 &\leqslant \frac{s(B)^2}{2}(1 - \pi_1)^4 \text{ (since } \pi_1 \leqslant 1/2) \\
&< \frac{(\pi_1 - t)^2}{2}(1 - \pi_1)^4 = \frac{\pi_1^2}{2}(1 - \pi_1)^4 - 2t(1 - \pi_1)^4 + o(t) \\
&< \frac{\pi_1^2}{2}(1 - \pi_1)^4 - t(1 - \pi_1)^4 = g(C_{1,n}) - t(1 - \pi_1)^4 \qquad (24)
\end{aligned}$$

for sufficiently small t. If inequality (II) is true, then

$$(\rho(B) - \pi_1)^4 \leqslant \max\{(1 - t - \pi_1)^4, \pi_1^4\} = \max\{(\pi_2 - t)^4, \pi_1^4\} = (\pi_2 - t)^4$$

for sufficiently small t, as $\pi_1 < \pi_2$. Therefore,

$$g(B) \leqslant \frac{\pi_1^4}{2}(\pi_2 - t)^4 = \frac{\pi_1^4}{2}\pi_2^4 - 4\pi_2^3 t + o(t) < g(C_{1,n}) - 2\pi_2^3 t \tag{25}$$

for sufficiently small t. Thus for all $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{1,n}$, $g(B) < g(C_{1,n}) - a_1 t$ with $a_1 = \min\{(1 - \pi_1)^4, 2\pi_2^3\}$.

Suppose $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{2,n}$ and inequality (I) is true. Then $s(B) > \pi_1 + t$. Note that $0 \leqslant \rho(B)|B| \leqslant |C_{1,n}|$, yielding the useful inequality

$$0 \leqslant \rho(B) \leqslant \pi_1 / s(B). \tag{26}$$

Subtracting through by π_1 gives

$$(\rho(B) - \pi_1)^4 \leqslant \max\{\pi_1^4, \pi_1^4(1/s(B) - 1)^4\} = \pi_1^4(1/s(B) - 1)^4.$$

Therefore,

$$g(B) \leqslant \frac{s(B)^2}{2}\pi_1^4(1/s(B) - 1)^4 = \frac{\pi_1^4}{2}(1/\sqrt{s(B)} - \sqrt{s(B)})^4 < \frac{\pi_1^4}{2}(1/\sqrt{\pi_1 + t} - \sqrt{\pi_1 + t})^4, \tag{27}$$

since $F(x) := (1/\sqrt{x} - \sqrt{x})^4$ is decreasing on $(0, 1]$, and $s(B) > \pi_1 + t$. Note that

$$\frac{d}{dt}\left(\frac{1}{\sqrt{\pi_1 + t}} - \sqrt{\pi_1 + t}\right)^4 = -3\left(\frac{1}{\sqrt{\pi_1 + t}} - \sqrt{\pi_1 + t}\right)^3 \left[\frac{1}{2(\pi_1 + t)^{3/2}} + \frac{1}{2\sqrt{\pi_1 + t}}\right]. \tag{28}$$

By Taylor's theorem, this implies that

$$(1/\sqrt{\pi_1 + t} - \sqrt{\pi_1 + t})^4 = (1/\sqrt{\pi_1} - \sqrt{\pi_1})^4 - a_2 t + o(t) < (1/\sqrt{\pi_1} - \sqrt{\pi_1})^4 - a_2 t/2$$

for sufficiently small t, where a_2 is the right-hand-side of (28) at $t = 0$. Note further that $(1/\sqrt{\pi_1} - \sqrt{\pi_1})^4 = (\pi_2/\sqrt{\pi_1})^4 = \pi_2^4/\pi_1^2$. Putting these facts together with inequality (27), we obtain

$$g(B) < \frac{\pi_1^4}{2}\frac{\pi_2^4}{\pi_1^2} - a_2 t/2 = \frac{\pi_1}{2}\pi_2^4 - a_2 t/2 = g(C_{1,n}) - a_2 t/2 \tag{29}$$

If inequality (II) is true, $\rho(B) < 1 - t$. If $\rho(B) \leqslant \pi_1$, $(\rho(B) - \pi_1)^4$ is maximized when $\rho(B) = 0$, so that, since $s(B) \leqslant \pi_2$,

$$g(B) \leqslant \frac{\pi_2^2}{2}\pi_1^4 = g(C_{1,n}) + \frac{\pi_2^2}{2}\pi_1^4 - \frac{\pi_1^2}{2}\pi_2^4 = g(C_{1,n}) + \frac{\pi_1^2\pi_2^2}{2}(\pi_2^2 - \pi_1^2) < g(C_{1,n}) - t \tag{30}$$

for sufficiently small t, since π_1 is fixed. If $\rho(B) > \pi_1$, note that inequality (26) implies

$s(B) \leqslant \pi_1/s(B)$. Therefore,

$$g(B) \leqslant \frac{\pi_1^2}{2\rho(B)^2}(\rho(B)-\pi_1)^4 = \frac{\pi_1^2}{2}(\sqrt{\rho(B)}-\pi_1/\sqrt{\rho(B)})^4 < \frac{\pi_1^2}{2}(\sqrt{1-t}-\pi_1/\sqrt{1-t})^4 \quad (31)$$

since $G(x) := (\sqrt{x}-\pi_1/\sqrt{x})^4$ is increasing on $(\pi_1, 1]$. A similar Taylor expansion argument to that yielding inequality (29) yields, for a constant a_3 depending only on π_1,

$$g(B) < \frac{\pi_1^2}{2}(1-\pi_1)^4 - a_3 t/2 = g(C_{1,n}) - a_3 t/2, \quad (32)$$

for sufficiently small t. Pulling together inequalities (29), (30), and (32), we have that for all $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{1,n}$, $g(B) < g(C_{1,n}) - a_4$ with $a_4 := \min\{a_2/2, 1, a_3/2\}$.

Suppose $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{3,n}$. Note that $|B| - |C_{2,n}| \leqslant |B \cap C_{1,n}| \leqslant |C_{1,n}|$. Dividing through by $|B|$ yields the useful inequality

$$1 - \pi_2/s(B) \leqslant \rho(B) \leqslant \pi_1/s(B). \quad (33)$$

Subtracting inequality (33) by π_1 gives

$$\pi_2(1 - 1/s(B)) \leqslant \rho(B) - \pi_1^4 \leqslant \pi_1(1/s(B) - 1).$$

Since $\pi_1 < \pi_2$, this implies that $(\rho(B)-\pi_1)^4 \leqslant \pi_2^4(1-1/s(B))^4$. Therefore,

$$g(B) \leqslant \frac{s(B)^2}{2}\pi_2^4(1/s(B)-1)^4 = \frac{\pi_2^4}{2}(1/\sqrt{s(B)} - \sqrt{s(B)})^4 < \frac{\pi_2^4}{2}(1/\sqrt{\pi_2} - \sqrt{\pi_2})^4, \quad (34)$$

since $F(x) := (1\sqrt{x} - \sqrt{x})^4$ is decreasing on $I_3 := (\pi_2, 1]$ and $s(B) \in I_3$. Note that $\sqrt{\pi_2} - 1/\sqrt{\pi_2} = -\pi_1/\sqrt{\pi_2}$. Therefore,

$$g(B) < \frac{\pi_2^4}{2}\frac{\pi_1^4}{\pi_2^2} = \frac{\pi_2^2}{2}(0-\pi_1)^4 = g(C_{2,n}) < g(C_{1,n}) - t$$

for t sufficiently small. Thus, for $a := \min\{a_1, a_4, 1\}$, for sufficiently small t we have $g(B) < g(C_{1,n}) - at$ whenever $B \in \mathcal{R}(t)$. This completes the proof in the case $\pi_1 < \pi_2$.

Part 2 ($\pi_1 = \pi_2$): Recall that when $\pi_1 = \pi_2$ we define $\mathcal{R}(t)$ by

$$\mathcal{R}(t) := \{B \subseteq [n] : |s(B) - \pi_1| \vee \rho(B) \vee [1-\rho(B)] \leqslant t\}$$

Hence we will use the fact that $B \in \mathcal{R}(t)$ implies at least one of the inequalities (I) $|s(B) - \pi_1| > t$ or (II) $t < \rho(B) < 1 - t$ is true. Define the intervals $I_1 := [0, \pi_1]$, $I_2 := (\pi_1, 1]$. We bisect $2^{[n]}$, the domain of g, with the collections $\mathcal{D}_{i,n} := \{B \subseteq [n] : s(B) \in I_i\}$, for $i = 1, 2$. We will prove that the inequality $g(B) < g(C_{1,n}) - at$ holds for all $B \in \mathcal{R}(t)$ on each of those collections.

Suppose $B \in \mathcal{R}(t)^c \cap \mathcal{D}_{1,n}$ and inequality (I) is true. Then the same derivation yielding inequality (24) gives $g(B) < g(C_{1,n}) - t(1-\pi_1)^4$ for sufficiently small t. If inequality (II) is

true, then

$$(\rho(B) - \pi_1)^4 \leqslant \max\{(1 - t - \pi_1)^4, (\pi_1 - t)^4\} = \max\{(\pi_2 - t)^4, (\pi_1 - t)^4\} = (\pi_2 - t)^4,$$

since $\pi_1 = \pi_2$. Therefore, inequality (25) remains intact. Both inequalities hold on I_2 as well, for the roles of π_1 and π_2 may be interchanged, and the derivations treated symmetrically. This completes the proof in the case $\pi_1 = \pi_2$. ■

A.3 Proof of Lemma 9

Recall the definitions of set modularity and population set modularity from Definitions 2 and 9. Define $W := \sum_{\ell \in [L]} \widehat{Q}_\ell(B)$ and $w := \sum_{\ell \in [L]} q_\ell(B)$. Note that by Part 1 of Lemma 6, $q_\ell(B) \geqslant 0$ regardless of B, a fact which will allow the application of Lemma 16 in what follows. We have $\widehat{H}(B, L) = |L|^{-1} W_+^2$, $H_*(B, L) = |L|^{-1} w^2$, and for any B such that $|B| \geqslant n\varepsilon$,

$$\mathbb{P}_n\left(\left|\widehat{H}(B, L) - H_*(B, L)\right| > \frac{4|L|t}{n^2} + \frac{52|L|}{\kappa n}\right) = \mathbb{P}_n\left(\left|W_+^2 - w^2\right| > \frac{4|L|^2 t}{n^2} + \frac{52|L|^2}{\kappa n}\right)$$

$$\leqslant \mathbb{P}_n\left(\max_{\ell \in [L]} \left|\widehat{Q}_\ell(B) - q_\ell(B)\right| > \frac{t}{n^2} + \frac{13}{\kappa n}\right) \leqslant 4|L| \exp\left(-\kappa^2 \frac{\varepsilon t^2}{16 n^2}\right)$$

for large enough $t > 0$, where the first inequality follows from Lemma 16 for large enough n, and the second inequality follows from Lemma 19 and a union bound. Applying a union bound over sets $B \in \mathcal{B}_n$ yields the result. ■

A.4 Proof of Lemma 12

Assume first that $k > 1$. By definition, $B \in \widetilde{N}_{n,k}(A)$ implies that at least one of $d_h(B, C_1) \leqslant A \cdot n \cdot b_{n,k-1}$ or $d_h(B, C_2) \leqslant A \cdot n \cdot b_{n,k-1}$ is true. Suppose the first inequality holds. Since $d_h(B, C_1) = |B \setminus C_1| + |C_1 \setminus B|$, we have the inequality

$$\big||B| - n\pi_1\big| = \big||B| - |C_1|\big| \leqslant \big||B| - |B \cap C_1| - |C_1| + |B \cap C_1|\big| \leqslant \big||B| - |B \cap C_1|\big|$$

$$+ \big||C_1| - |B \cap C_1|\big| = |B \setminus C_1| + |C_1 \setminus B| \leqslant A \cdot n \cdot b_{n,k-1}$$

Alternatively, if $d_h(B, C_2) \leqslant A \cdot n \cdot b_{n,k-1}$, we have the same bound for $\big||B| - n\pi_2\big|$. Therefore, since $\pi_1 \leqslant \pi_2$, $B \in \widetilde{N}_{n,k}(A)$ implies that $|B| \geqslant n\pi_1 - A \cdot n \cdot b_{n,k-1}$. Since $b_{n,k-1} = o(1)$ as $n \to \infty$ and $\varepsilon < \pi_1$, this implies that for large enough n, $\widetilde{N}_{n,k}(A) \subseteq \mathcal{B}_n(\varepsilon)$. By Lemma 9, therefore, for large enough n, we have

$$\mathbb{P}_n\left(\sup_{\widetilde{N}_{n,k}(A)} \left|\widehat{H}(B, L) - H_*(B, L)\right| > \frac{4|L|t}{n^2} + \frac{52|L|}{\kappa n}\right) \leqslant 4|L||\widetilde{N}_{n,k}(A)| \exp\left(-\kappa^2 \frac{\varepsilon t^2}{16 n^2}\right) \quad (35)$$

for all $t > 0$. We now bound the right-hand side of inequality (35) with t replaced by $t_n := n^{1+\frac{1}{2^k}}(\log n)^{1-\frac{1}{2^k}}$. Note that

$$\frac{t_n^2}{n^2} = \frac{1}{n^2} n^{2+\frac{1}{2^{k-1}}}(\log n)^{2-\frac{1}{2^{k-1}}} = n \cdot n^{\frac{1}{2^{k-1}}-1}(\log n)^{1-\frac{1}{2^{k-1}}} \log n = n \cdot b_{n,k-1} \log n.$$

Furthermore, by Corollary 15 (see Appendix B) we have $|\widetilde{N}_{n,k}(A)| \leqslant 2\exp\left[3A \cdot n \cdot b_{n,k-1} \log\left(1/b_{n,k-1}\right)\right]$. These facts yield the bound

$$\begin{aligned}
|\widetilde{N}_{n,k}(A)| \exp\left(-\kappa^2 \frac{\varepsilon t_n^2}{16n^2}\right) &\leqslant 2\exp\left\{-\kappa^2 \frac{\varepsilon}{16} n \cdot b_{n,k-1}\left[\log n - \frac{16}{\kappa^2 \varepsilon} 3A \log\left(1/b_{n,k-1}\right)\right]\right\} \\
&\leqslant 2\exp\left(-\kappa^2 \frac{\varepsilon}{32} n \cdot b_{n,k-1} \log n\right) \text{ (for large } n\text{, since } 1/b_{n,k-1} = o(n)) \\
&< 2\exp\left(-\kappa^2 \frac{\varepsilon}{32} n \gamma_n^{1-\varepsilon} \log n\right)
\end{aligned}$$

where the final inequality follows from the choice of k satisfying $\frac{1}{2^{k-1}} < \varepsilon$. Therefore,

$$4|L||\widetilde{N}_{n,k}(A)| \exp\left(-\kappa^2 \frac{\varepsilon t_n^2}{16n^2}\right) \leqslant 2\exp\left\{-\frac{\kappa^2 \varepsilon}{32} n \gamma_n^{1-\varepsilon} \log n + O(\log |L|)\right\} \tag{36}$$

for large enough n. Notice now that $t_n/n^2 = b_{n,k}$ vanishes slower than $1/n$, and is therefore the leading order term in the expression $\frac{4|L|t_n}{n^2} + \frac{52|L|}{\kappa n}$ (see equation 35). Hence for large enough n we have $\frac{4|L|t_n}{n^2} + \frac{52|L|}{\kappa n} \leqslant 5|L|b_{n,k}$. Combining this observation with lines (35) and (36) proves the result in the case $k > 1$.

If $k = 1$, assume $A = \varepsilon$. By definition, then (see Definition 11), $N_{n,k}(A) = \mathcal{B}_n(\varepsilon)$. Returning to inequality (35), we note that $\log|\mathcal{B}_n(\varepsilon)| = O(n)$, and thus we can derive the bound (36) with the same choice of $t_n := n^{1+\frac{1}{2^k}}(\log n)^{1-\frac{1}{2^k}} = n\sqrt{n \log n}$. The rest of the argument goes through unaltered. ■

B Technical Results

Lemma 14. *Fix $\pi_1 \in [0,1]$. For each n, let $C_1 \subseteq [n]$ be an index set of size $\lfloor n\pi_1 \rfloor$. Let $C_2 := [n] \setminus C_1$. Let $\gamma_n \in [0,1]$ be a sequence such that $\gamma_n \to 0$ and $n\gamma_n \to \infty$. Then for large enough n,*

$$|N(C_1, \gamma_n)| \leqslant \exp\{3n\gamma_n \log(1/\gamma_n)\}$$

Proof. Define the boundary of a neighborhood of $C \subseteq [n]$ by

$$\partial N(C, r) := \{B \subseteq [n] : d_h(B, C) = \lfloor nr \rfloor\}.$$

Note that any $B \subseteq [n]$ may be written as the disjoint union $B = \{C_2 \cap B\} \cup \{C_1 \cap B\}$. Since $C_1 \cap B = C_1 \setminus \{C_1 \setminus B\}$, for fixed $k \in [n]$ it follows that each set $B \in \partial N(C, k/n)$ is uniquely

identified with choices of $|C_2 \cap B|$ indices from C_2 and $|C_1 \setminus B|$ indices from C_1 such that

$$|B \cap C_2| + |C_1 \setminus B| = |B \setminus C_1| + |C_1 \setminus B| = d_h(B, C_1) = k$$

Therefore, we have the equality

$$|\partial N(C_1, k)| \;=\; \sum_{m=0}^{k} \left[\binom{|C_2|}{m} + \binom{|C_1|}{k-m} \right] \tag{37}$$

Note that for positive integers K, N with $K < N/2$, properties of the geometric series yield the following bound:

$$\begin{aligned} \binom{N}{K}^{-1} \sum_{m=0}^{K} \binom{N}{m} &= \sum_{m=0}^{K} \frac{(N-K)!K!}{(N-m)!m!} = \sum_{m=0}^{K} \prod_{j=m+1}^{K} \frac{j}{N-j+1} \\ &< \sum_{m=0}^{K} \left(\frac{K}{N-K+1} \right)^m < \frac{N-(K-1)}{N-(2K-1)} \end{aligned} \tag{38}$$

For sufficiently small K/N, the right-hand side of inequality (38) is less than 2, and thus $\sum_{m=0}^{K} \binom{N}{m} < 2\binom{N}{K}$ if $K \ll N$. We apply this inequality to equation (37). Choose n large enough so that $n\gamma_n < \frac{1}{2}\min\{|C_1|, |C_2|\}$, which is possible since $\gamma_n \to 0$. Then for fixed $k \leqslant n\gamma_n$, we have that $|\partial N(C_1, k)| < 2\left[\binom{|C_2|}{k} + \binom{|C_1|}{k}\right]$ for large enough n. By another application of the inequality derived from (38), using the fact that $n\gamma_n = o(n)$, we therefore obtain

$$\begin{aligned} |N(C_1, \gamma_n)| = \sum_{k=0}^{\lfloor n\gamma_n \rfloor} |\partial N(C_1, k)| &< \sum_{k=0}^{\lfloor n\gamma_n \rfloor} 2\left[\binom{|C_2|}{k} + \binom{|C_1|}{k} \right] \\ &< 4\left[\binom{|C_2|}{\lfloor n\gamma_n \rfloor} + \binom{|C_1|}{\lfloor n\gamma_n \rfloor} \right] \leqslant 8\binom{n}{\lfloor n\gamma_n \rfloor} \end{aligned}$$

As $\binom{N}{K} \leqslant \left(\frac{N \cdot e}{K}\right)^K$, we have

$$|N(C_1, \gamma_n)| \leqslant \exp\left\{\log(8) + n\gamma_n \left[\log(e) + \log(1/\gamma_n)\right]\right\} \leqslant \exp\{3n\gamma_n \log(1/\gamma_n)\}$$

for large enough n, since $1/\gamma_n \to \infty$. ■

Here we give a short Corollary to Lemma 14 which directly serves the proof of Lemma 12. Recall $\widetilde{N}_{n,k}(A)$ from Definition 11 in Section 4.1.2.

Corollary 15. *Fix an integer $k > 1$. For large enough n,*

$$|\widetilde{N}_{n,k}(A)| \leqslant 2\exp\left[3A \cdot n \cdot b_{n,k-1} \log\left(1/b_{n,k-1}\right)\right]$$

Proof. The corollary follows from a direct application of Lemma 14 to $N(C_1, A \cdot b_{n,k-1})$ and

$N(C_2, A \cdot b_{n,k-1})$. ■

Lemma 16. *Let $x_1, \ldots, x_k \in (0,1)$ be fixed and let $X_1, \ldots, X_k$ be arbitrary random variables. Define $W := \sum_i X_i$ and $w := \sum_i x_i$. Then for t sufficiently small, $\mathbb{P}(|W_+^2 - w^2| > 4k^2t) \leqslant \mathbb{P}(\max_i |X_i - x_i| > t)$.*

Proof. Define $D_i := |X_i - x_i|$ and fix $t < \min_i x_i$. Then if $\max_i D_i \leqslant t$, all X_i's will be positive, and thus $W_+ = W$ and $|W - w| \leqslant kt$, by the triangle inequality. Therefore $\max_i D_i \leqslant t$ implies that

$$|W_+^2 - w^2| = |(W-w)^2 + 2w(W-w)| \leqslant k^2t^2 + 2wkt \leqslant k^2t^2 + 2k^2t \tag{39}$$

Thus by the law of total probability, we have

$$\mathbb{P}(|W_+^2 - w^2| > 4k^2t) \leqslant \mathbb{P}(\{|W_+^2 - w^2| > 4k^2t\} \cap \{\max_i D_i \leqslant t\}) + \mathbb{P}(\max_i D_i > t)$$

Inequality (39) implies that for sufficiently small t, the first probability on the right-hand side above is equal to 0. The result follows. ■

In what follows we state and prove Lemma 19, a concentration inequality for the modularity of a node set (see Definition 2) from a single-layer SBM with n nodes and two communities. We first give a few short facts about the 2-community SBM. For for all results that follow, let s, ρ, and v (see Definition 4) correspond to the fixed set $B \subseteq [n]$ in each result (though sometimes we will make explicit the dependence on B). Define a matrix V by $V(i,j) := P(i,j)(1 - P(i,j))$ for $i = 1, 2$, where P is the probability matrix associated with the 2-block SBM.

Lemma 17. *Consider a single-layer SBM with $n > 1$ nodes, two communities, and parameters P and π_1. Fix a node set $B \subseteq [n]$ with $|B| \geqslant \alpha n$ for some $\alpha \in (0,1)$. Then*

1. $\left|\mathbb{E}[Y(B)] - \binom{|B|}{2} v^t P v\right| \leqslant 3|B|/2$
2. $\left|\mathbb{E}\left[\sum_{u \in B} \widehat{d}(u)\right] - |B| n v^t P \pi\right| \leqslant |B|$
3. $Var\left[\sum_{u \in B} \widehat{d}(u)\right] \leqslant 9|B|n$

Proof. For part 1, note that by definition,

$$\mathbb{E}\left[Y(B)\right] = \sum_{u,v \in B : u < v} \mathbb{P}\left((u,v) \in \widehat{E}\right) = \frac{1}{2} \sum_{u \neq v : u,v \in B} \mathbb{P}\left((u,v) \in \widehat{E}\right)$$

The right-hand sum can be expressed the sum of the entries of a 2×2 symmetric block matrix with zeroes on the diagonal. In this matrix, the upper diagonal block is of size $|B \cap C_1|$ with off-diagonal entries equal to $P(1,1)$. The lower diagonal block is of size $|B \cap C_2|$ with off-diagonal entries equal to $P(2,2)$. The off-diagonal blocks have entries equal to $P(1,2)$.

Therefore, summing over blocks and accounting for the zero diagonal, we have

$$\mathbb{E}\left[Y(B)\right] = \frac{1}{2}\Big[|B\cap C_1|^2 P(1,1) + |B\cap C_1||B\cap C_2|P(1,2) + |B\cap C_2|^2 P(2,2)\Big]$$

$$- \frac{1}{2}\Big[|B\cap C_1|P(1,1) + |B\cap C_2|P(2,2)\Big]$$

By dividing and multiplying by $|B|^2$ and collapsing cross-products, we get

$$\begin{aligned}
\mathbb{E}\left[Y(B)\right] &= \frac{|B|^2}{2}\left[v^t P v - \frac{\rho P(1,1) + (1-\rho)P(2,2)}{|B|}\right] \\
&= \binom{|B|}{2}\left[1 + \frac{1}{|B|-1}\right]\left[v^t P v - \frac{\rho P(1,1) + (1-\rho)P(2,2)}{|B|}\right] \\
&= \binom{|B|}{2}\left[v^t P v - \frac{\rho P(1,1) + (1-\rho)P(1,2)}{|B|} + \frac{v^t P v}{|B|-1} - \frac{\rho P(1,1) + (1-\rho)P(2,2)}{|B|(|B|-1)}\right]
\end{aligned}$$

Part 1 follows by carrying out the multiplication by $\binom{|B|}{2}$ in the last expression.

For part 2, let $P(\cdot, i)$ denote the i-th column of P. Note that $\mathbb{E}\left[\widehat{d}(u)\right] = n\pi^T P(\cdot, c_u) - P(c_u, c_u)$, with $c_u \in \{1,2\}$ denoting the community index of u. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{u\in B}\widehat{d}(u)\right] &= \sum_{u\in B}\mathbb{E}\left[\widehat{d}(u)\right] = \sum_{u\in B\cap C_1}\mathbb{E}\left[\widehat{d}(u)\right] + \sum_{u\in B\cap C_2}\mathbb{E}\left[\widehat{d}(u)\right] \\
&= |B\cap C_1|\left[n\pi^T P(\cdot,1) - P(1,1)\right] + |B\cap C_2|\left[n\pi^T P(\cdot,2) - P(2,2)\right] \\
&= |B|\left[n\rho\pi^T P(\cdot,1) + n(1-\rho)\pi^T P(\cdot,2) - \rho P(1,1) - (1-\rho)P(2,2)\right] \\
&= |B| n v^t P\pi - |B|\big[\rho P(1,1) + (1-\rho)P(2,2)\big]
\end{aligned}$$

which completes part 2.

Finally, for part 3, we have

$$\mathrm{Var}\left[\sum_{u\in B}\widehat{d}(u)\right] = \mathrm{Var}\left[2Y(B)\right] + \sum_{u,v:u\in B, v\in B^C}\mathrm{Var}\left[\widehat{X}(u,v)\right]. \tag{40}$$

We address these two terms separately. For the first term, a calculation analogous to that from part 1 yields that $|\mathrm{Var}\left[Y(B)\right] - \binom{|B|}{2} v^t V v| \leqslant 3|B|/2$. Defining $\bar{v} := (\rho(B^C), 1-\rho(B^C))^t$, it is easy to show that $\sum_{u,v:u\in B, v\in B^C}\mathrm{Var}\left[\widehat{X}(u,v)\right] = |B||B^C| v^t V \bar{v}$, which is simply the sum of variances of all edge indicators for edges from B to B^C. Applying these observations to

equation (40), we have

$$\operatorname{Var}\left[\sum_{u\in B}\widehat{d}(u)\right] \leqslant 4\binom{|B|}{2}v^tVv + 12|B|/2 + |B||B^C|v^TV\bar{v}$$

$$\leqslant |B|\big[2(|B|-1)v^tVv + 6 + |B^C|v^tV\bar{v}\big] \leqslant 9|B|n$$

■

Lemma 18. *Under a single-layer SBM with $n > 1$ nodes, two communities, and parameters P and π_1, define $\kappa := \pi^T P\pi$. Then for large enough n, $\mathbb{P}\left(\left|2|\widehat{E}| - n^2\kappa\right| > t + 4n\right) \leqslant 2\exp\left\{-\frac{t^2}{n^2}\right\}$ for any $t > 0$.*

Proof. Note that $|\widehat{E}| = Y([n])$. Thus part 1 of Lemma 17 with $B = [n]$ yields $\left|\mathbb{E}[|\widehat{E}|] - \binom{n}{2}\kappa\right| \leqslant 3n/2$ for large enough n. As $n^2/2 = \binom{n}{2} + n/2$, by the triangle inequality,

$$\left|\mathbb{E}[|\widehat{E}|] - \frac{n^2}{2}\kappa\right| \leqslant \left|\mathbb{E}[|\widehat{E}|] - \binom{n}{2}\kappa\right| + \frac{n}{2} \leqslant 2n$$

Thus for any $t > 0$, Hoeffding's inequality gives

$$\mathbb{P}\left(\left|2\widehat{E} - n^2\kappa\right| > t + 4n\right) \leqslant \mathbb{P}\left(\left|2\widehat{E} - n^2\kappa\right| > t + 2\left|\mathbb{E}[|\widehat{E}|] - \frac{n^2}{2}\kappa\right|\right)$$

$$\leqslant \mathbb{P}\left(\left|2\widehat{E} - 2\mathbb{E}[|\widehat{E}|]\right| > t\right) \leqslant 2\exp\left\{-2\frac{t^2}{4\binom{n}{2}}\right\} \leqslant 2\exp\{-t^2/n^2\}$$

■

Lemma 19. *Consider a single-layer 2-block SBM having $n > 1$ nodes and parameters P and π. Fix $\alpha \in (0,1)$ and $B \subseteq [n]$ such that $|B| \geqslant \alpha n$. Then for large enough n we have*

$$\mathbb{P}_n\left(\left|\widehat{Q}(B) - q(B)\right| > \frac{t}{n^2} + \frac{8}{\kappa n}\right) \leqslant 4\exp\left(-\frac{\kappa^2\alpha t^2}{16n^2}\right) \tag{41}$$

for any $t > 0$.

Proof. With notation laid out in Section 2.2, define

$$\widetilde{Q}(B) := n^{-1}\binom{|B|}{2}^{-1/2}(Y(B) - \widetilde{\mu}(B)) \tag{42}$$

where

$$\widetilde{\mu}(B) := \frac{\sum_{u,v\in B:u<v}\widehat{d}(u)\widehat{d}(v)}{n^2\kappa} \tag{43}$$

We will prove the inequality in three steps: *Step 1*: bounding $\left|\widehat{Q}(B) - \widetilde{Q}(B)\right|$ in probability; *Step 2*: deriving a concentration inequality for $\widetilde{Q}(B)$; and *Step 3*: showing that $\left|\mathbb{E}\left[\widetilde{Q}(B)\right] - q(B)\right|$ is eventually bounded by a constant.

Step 1. As $\sum_{u,v\in B;u<v} \widehat{d}(u)\widehat{d}(v) \leqslant \sum_{u\in B} \widehat{d}(u)^2$, we have

$$\sum_{u,v\in B;u<v} \widehat{d}(u)\widehat{d}(v) \leqslant \sqrt{\sum_{u,v\in B;u<v} \widehat{d}(u)\widehat{d}(v)}\sqrt{\sum_{u\in B} \widehat{d}(u)^2} \leqslant n\binom{|B|}{2}^{1/2}|\widehat{E}|$$

Therefore,

$$\left|\widehat{Q}(B) - \widetilde{Q}(B)\right| = n^{-1}\binom{|B|}{2}^{-1/2}\left|\frac{(2|\widehat{E}| - n^2\kappa)\sum_{u,v\in B;u<v} \widehat{d}(u)\widehat{d}(v)}{2|\widehat{E}|n^2\kappa}\right| \leqslant \frac{\left|2|\widehat{E}| - n^2\kappa\right|}{2n^2\kappa} \quad (44)$$

Combining the inequality in (44) with Lemma 18, for any $t > 0$,

$$\mathbb{P}\left(\left|\widehat{Q}(B) - \widetilde{Q}(B)\right| > \frac{t}{2n^2} + \frac{2}{\kappa n}\right) \leqslant \mathbb{P}\left(\left|2|\widehat{E}| - n^2\kappa\right| > \kappa t + 4n\right) \leqslant 2\exp\left(-\frac{\kappa^2 t^2}{n^2}\right). \quad (45)$$

Step 2. This step relies on McDiarmid's concentration inequality. Recall from Section 2.1 that $\widehat{X}(u,v)$ denotes the indicator of edge presence between nodes u and v. Note that node pairs have a natural, unique ordering along the upper-diagonal of the adjacency matrix. Define $ord\{u,v\} = 2(u-1) + (v-1)$, for $\{u,v\} \in [n]^2$ with $u < v$ (e.g. $ord\{1,2\} = 1$, $ord\{1,3\} = 2$, etc.). For all $n > 1$ and $i \leqslant n(n-1)/2$, define $\widehat{Z}(i) := \widehat{X}(u,v)$ such that $ord\{u,v\} = i$. If $ord\{u,v\} = i$, we call $\{u,v\}$ the "i-th ordered node pair". Define the set

$$\mathcal{I}(B) := \{i : \text{the } i\text{-th ordered node pair has at least one node in } B\}$$

and let $\widehat{\mathcal{Z}}(B) := \{\widehat{Z}(i) : i \in I(B)\}$. Note that the proxy score $\widetilde{Q}(B)$ is a function $f(z_1, z_2, \ldots)$ of the indicators $\widehat{\mathcal{Z}}(B)$.

Consider a *fixed* indicator set $\mathcal{Z}(B)$. For each $j \in I(B)$, define $\mathcal{Z}^j(B) := \{Z^j(i) : i \in I(B)\}$ with

$$Z^j(B) := \begin{cases} Z^j(i) = 1 - Z(i), & i = j \\ Z^j(i) = Z(i), i \neq j \end{cases} \quad (46)$$

To apply McDiarmid's inequality, we must bound $\Delta(j) := |f(\mathcal{Z}(B)) - f(\mathcal{Z}^j(B))|$ uniformly over $j \in \mathcal{I}(B)$. Fix $j \in \mathcal{I}(B)$ and let $\{u', v'\}$ be the j-th ordered edge. Without loss of generality, we assume $Z(j) = 1$. Since $f(\mathcal{Z}(B)) = Q(B)$, $f(\mathcal{Z}(B))$ has a representation in terms of $Y(B)$ and $\widetilde{\mu}(B)$. We let $Y^j(B)$ and $\widetilde{\mu}^j(B)$ correspond to $f(\mathcal{Z}(B)^j)$. Notice that

$$n\binom{|B|}{2}^{1/2}\Delta(j) = \left|Y(B) - Y^j(B) - \left[\widetilde{\mu}(B) - \widetilde{\mu}^j(B)\right]\right| \quad (47)$$

We bound the right hand side of equation (47) in two cases: (i) $u', v' \in B$, and (ii) $u' \notin$

$B, v' \in B$. In case (i), $Y(B) - Y^j(B) = 1$, and

$$\widetilde{\mu}(B) - \widetilde{\mu}^i(B) = \frac{\sum_{u,v \in B; u \neq v} d(u)d(v) - d^j(u)d^j(v)}{n^2\kappa} = \frac{d(u')d(v') - d^j(u')d^j(v')}{n^2\kappa}$$
$$= \frac{d(u')d(v') - (d(u')-1)(d(v')-1)}{n^2\kappa} = \frac{d(u') + d(v') - 1}{n^2\kappa},$$

which is bounded in the interval $(0,1)$ for large enough n. Thus in case (i), $\Delta(j) \leqslant 2\binom{|B|}{2}^{-1/2}$ by the triangle inequality, for large enough n. In case (ii), $Y(B) - Y'(B) = 0$, and

$$\widetilde{\mu}(B) - \widetilde{\mu}^j(B) = \frac{\sum_{u,v \in B; u \neq v} d(u)d(v) - d^j(u)d^j(v)}{n^2\kappa} = \frac{\sum_{u \in B; u \neq v'} d(u)\left[d(v') - d^j(v')\right]}{n^2\kappa}$$
$$= \frac{\sum_{u \in B; u \neq v'} d(u)}{n^2\kappa} \leqslant \frac{n|B|}{n^2\kappa} \leqslant \kappa^{-1}$$

Hence due to equation (47), we have for sufficiently large n that

$$\Delta(j) \leqslant n^{-1}\binom{|B|}{2}^{-1/2} \cdot \max\{2, \kappa^{-1}\} \leqslant n^{-1}\binom{|B|}{2}^{-1/2} \cdot 2 \cdot \kappa^{-1} \tag{48}$$

for all $j \in \mathcal{I}(B)$, as $\kappa \leqslant 1$. Since $|\mathcal{I}(B)| = \binom{|B|}{2} + |B||B^C| \leqslant n|B|$, McDiarmid's bounded-difference inequality implies that for sufficiently large n,

$$\mathbb{P}\left(\left|\widetilde{Q}(B) - \mathbb{E}\left[\widetilde{Q}(B)\right]\right| > \frac{t}{n}\right) = 2\exp\left(\frac{-t^2}{n|B|\Delta(j)}\right) \leqslant 2\exp\left(-\kappa^2\frac{n^2\binom{|B|}{2}t^2}{4n^3|B|}\right)$$
$$\leqslant 2\exp\left(-\kappa^2\frac{(|B|-1)t^2}{8n}\right) \leqslant 2\exp\left(-\kappa^2\frac{\alpha t^2}{16}\right)$$

for any $t > 0$. Replacing t by t/n gives

$$\mathbb{P}\left(\left|\widetilde{Q}(B) - \mathbb{E}\left[\widetilde{Q}(B)\right]\right| > \frac{t}{n^2}\right) \leqslant 2\exp\left(-\kappa^2\frac{\alpha t^2}{16n^2}\right) \tag{49}$$

Step 3. Turning our attention to $\mathbb{E}[\widetilde{Q}(B)]$, recall that $n\binom{|B|}{2}^{1/2}\widetilde{Q}(B) = Y(B) - \widetilde{\mu}(B)$ and that $\widetilde{\mu}(B) := \sum_{u,v \in B; u<v} \widehat{d}(u)\widehat{d}(v)/(n^2\kappa)$. As in previous lemmas, we will shorthand the quantities

$s(B), \rho(B)$, and $v(B)$, by s, ρ, and v (respectively). Note that

$$\mathbb{E}\left[2 \cdot \sum_{u,v \in B; u<v} \widehat{d}(u)\widehat{d}(v)\right] = \mathbb{E}\left\{\left[\sum_{u \in B} \widehat{d}(u)\right]^2 - \sum_{u \in B} \widehat{d}^2(u)\right\}$$

$$= \mathrm{Var}\left[\sum_{u \in B} \widehat{d}(u)\right] + \mathbb{E}\left[\sum_{u \in B} \widehat{d}(u)\right]^2 - \sum_{u \in B} \mathbb{E}\left[\widehat{d}^2(u)\right] \tag{50}$$

Part 3 of Lemma 17 gives $\mathrm{Var}\left[\sum_{u \in B} \widehat{d}(u)\right] \leqslant 9sn^2$. Furthermore, for $u \in C_i$ we have

$$\mathbb{E}\left[\widehat{d}^2(u)\right] = \mathrm{Var}\left[\widehat{d}(u)\right] + \mathbb{E}\left[\widehat{d}(u)\right]^2$$

$$= n\pi^T V(\cdot, i) - V(i,i) + n^2\left[\pi^T P(\cdot, i) - P(i,i)\right]^2,$$

and therefore $\sum_{u \in B} \mathbb{E}\left[\widehat{d}^2(u)\right] \leqslant 2sn^3$. Finally, Part 2 of Lemma 17 gives $\left|\mathbb{E}\left[\sum_{u \in B} \widehat{d}(u)\right] - |B| n v^T P \pi\right| \leqslant |B|$. By expansion, this implies there exists a constant a with $|a| < 3$ such that for large enough n, $\mathbb{E}\left[\sum_{u \in B} \widehat{d}(u)\right]^2 = s^2n^4(v^tP\pi)^2 + as^2n^3$. Therefore overall, line (50) implies there exists a constant b with $|b| < 6$ such that for large enough n, $\mathbb{E}\left[2 \cdot \sum_{u,v \in B; u<v} \widehat{d}(u)\widehat{d}(v)\right] = s^2n^4(v^T P\pi)^2 + bsn^3$. Therefore, using the definition of $\widetilde{\mu}(B)$,

$$\mathbb{E}\left[\widetilde{\mu}(B)\right] = s^2n^4\frac{(v^tP\pi)^2 + b(sn)^{-1}}{2n^2\kappa} = \binom{sn}{2}\left[1 + \frac{1}{sn-1}\right]\left[\frac{(v^tP\pi)^2}{\kappa} + \frac{b}{\kappa sn}\right]$$

$$= \binom{sn}{2}\left[\frac{(v^tP\pi)^2}{\kappa} + \frac{b}{\kappa sn} + \frac{(v^tP\pi)^2}{\kappa(sn-1)} + \frac{b}{\kappa sn(sn-1)}\right]$$

$$= \binom{sn}{2}\left[\frac{(v^tP\pi)^2}{\kappa} + \frac{1}{\kappa sn}\left(b + \frac{sn(v^tP\pi)^2 + b}{sn-1}\right)\right] = \binom{sn}{2}\left[\frac{(v^tP\pi)^2}{\kappa} + \frac{c_1}{\kappa sn}\right] \tag{51}$$

for a constant c_1 with $|c_1| < 8$, for large enough n. Now, part 1 of Lemma 17 gives that $\left|\mathbb{E}[Y(B)] - \binom{|B|}{2} v^tPv\right| \leqslant 3|B|/2$ for large enough n. Thus there exists a constant c_2 with $|c_2| < 3$ such that for large enough n, $\mathbb{E}[Y(B)] = \binom{|B|}{2}\left[v^tPv + \frac{c_2}{sn}\right]$. Thus

$$n\mathbb{E}\left[\widetilde{Q}(B)\right] = \binom{|B|}{2}^{-1/2}(\mathbb{E}[Y(B)] - \mathbb{E}[\tilde{\mu}(B)]) = \frac{sn}{\sqrt{2}}\left[v^tPv - \frac{(v^tP\pi)^2}{\kappa} + \frac{1}{sn}(c_1/\kappa + c_2)\right]$$

$$* \left(\sqrt{1 - \frac{1}{sn}}\right) = \frac{sn}{\sqrt{2}}\left[v^tPv - \frac{(v^tP\pi)^2}{\kappa}\right] + \frac{c_1/\kappa + c_2}{\sqrt{2}}\left(\sqrt{1 - \frac{1}{sn}}\right)$$

Thus there exists a constant c with $|c| \leqslant |c_1|/\kappa + |c_2| < 8/\kappa + 3$ such that for large enough n, $\mathbb{E}[\widetilde{Q}(B)] = q(B) + c/n$. This completes Step 3.

Completion of the proof: We now recall the results of the three steps:

(i) For large enough n, we have $\mathbb{P}\left(\left|\widehat{Q}(B) - \widetilde{Q}(B)\right| > \frac{t}{2n^2} + \frac{2}{\kappa n}\right) \leqslant 2\exp\left(-\frac{\kappa^2 t^2}{n^2}\right)$

(ii) $\mathbb{P}\left(\left|\widetilde{Q}(B) - \mathbb{E}\left[\widetilde{Q}(B)\right]\right| > \frac{t}{n^2}\right) \leqslant 2\exp\left(-\kappa^2 \frac{\alpha t^2}{16n^2}\right)$

(iii) There exists c with $|c| < 8/\kappa + 3$ such that for large enough n, $\mathbb{E}\left[\widehat{Q}(B)\right] = q(B) + c/n$

Noting that $\alpha/16 < 1$, we apply a union bound to the results of steps (i) and (ii):

$$\mathbb{P}\left(\left|\widehat{Q}(B) - \mathbb{E}\left[\widetilde{Q}(B)\right]\right| > \frac{t}{n^2} + \frac{2}{\kappa n}\right) \leqslant 4\exp\left(-\frac{\kappa^2 \alpha t^2}{16n^2}\right) \tag{52}$$

Applying the inequality $|x - a| \geqslant |x| - |a|$ with (iii) and some algebra gives the result. ■

C Competing Methods

In Sections 6 and 7, we compare and contrast the performance of Multilayer Extraction with the following methods:

Spectral clustering [35]: an iterative algorithm based on the spectral properties of the modularity matrix of an observed network. In the first step, the modularity matrix of the observed network is calculated and its leading eigenvector is identified. The graph is divided into two disjoint communities so that each vertex is assigned according to its sign in the leading eigenvector. Next, the modularity matrix is calculated for both of the subgraphs corresponding to the previous division. If the modularity of the partition increases, these communities are once again divided into two disjoint communities, and the procedure is repeated in this fashion until the modularity no longer increases.

Label Propagation [44]: an iterative algorithm based on propagation through the network. At the first step, all vertices are randomly assigned a community label. Sequentially, the algorithm chooses a single vertex and updates the labels of its neighborhood to be the majority label of the neighborhood. The algorithm continues updating labels in this way until no updates are available.

Fast and greedy [14]: an iterative and greedy algorithm that seeks a partition of vertices with maximum modularity. The algorithm is an agglomerative approach that is a modification of the Kernighan-Lin algorithm commonly used in the identification of community structure in network.

Walktrap [42]: an agglomerative algorithm that seeks a partition of vertices that minimizes the total length of a random walk within each community. At the first stage, each vertex of the network is placed in its own community. At each subsequent stage, the two closest communities (according to walk distance) are merged. This process is continued until all vertices have been merged into one large community, and a community dendrogram is formed. The partition with the smallest random walk distance is chosen as the final partition.

When used, we apply each of the above methods to the aggregate (weighted) network computed from the average of the layers in the analyzed multilayer network as well as to each layer separately. For each method, we use the default settings from the *igraph* package version 0.7.1 set in **R**.

D Extraction Simulations

D.1 Simulation

We now investigate several intrinsic properties of Multilayer Extraction by applying the method to multilayer networks with several types of community structure, including I) disjoint, II) overlapping, III) persistent, IV) non-persistent, and V) hierarchical structure. Figure 7 illustrates six multilayer networks that we analyze for this purpose. Each simulated network contains 1000 nodes and 90 layers. Embedded communities have inner connection probability 0.15; whereas, the remaining vertices independently connected to all other vertices with probability 0.05.

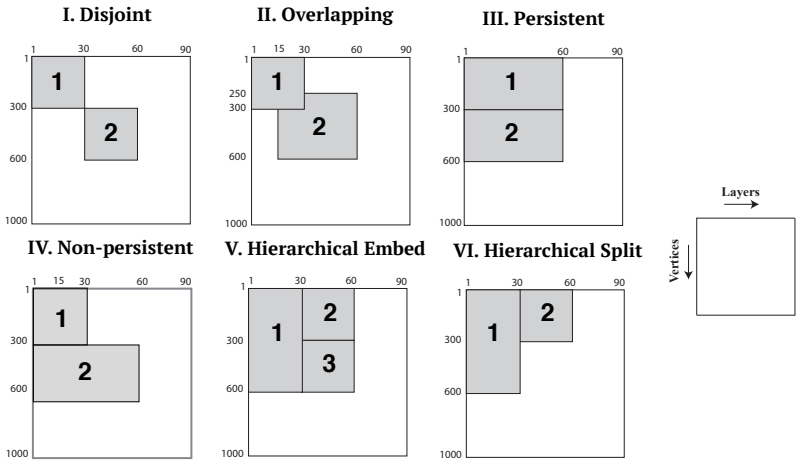

Figure 7: Simulation test bed for extraction procedures. Each graphic displays a multilayer network on 1000 nodes and 90 layers. In each plot, shaded rectangles are placed over the nodes (rows) and layers (columns) that are included in a multilayer community. Communities are labeled by number. Vertices within the same community are randomly connected with probability 0.15 while all other vertices have connection probability 0.05 to vertices in their respective layer.

D.2 Results

In the disjoint, overlapping, persistent, and non-persistent networks (I, II, III, and IV, respectively), Multilayer Extraction identifies communities that perfectly match the true embedded communities. On the other hand, in the hierarchical community setting, Multilayer Extraction is unable to identify the full set of communities. In example V, Multilayer Extraction does not identify community 1, and in example VI Extraction identifies a community with vertices 1 - 300 across layers 1 - 60, which combines community 1 and community 2.

Together, these results suggest two properties of the Multilayer Extraction procedure. First, the method can efficiently identify disjoint and overlapping community structure in multilayer networks with heterogeneous community structure. Second, Multilayer Extraction tends to disregard communities with a large number of vertices (e.g. communities that include over half of the vertices in a network). The inverse relationship between the score and the number of vertices in a community may provide some justification as to why this is the case. In networks with large communities, one can in principle modify the score by introducing a reward for large collections. We plan to pursue this further in future research.

References

[1] Bader, G. D. and C. W. Hogue (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics 4*(1), 2.

[2] Barigozzi, M., G. Fagiolo, and G. Mangioni (2011). Identifying the community structure of the international-trade multi-network. *Physica A: statistical mechanics and its applications 390*(11), 2051–2066.

[3] Bassett, D. S., N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton (2011, May). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences 108*(18), 7641 7646.

[4] Bender, E. A. and E. R. Canfield (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A 24*(3), 296–307.

[5] Berlingerio, M., M. Coscia, and F. Giannotti (2011). Finding redundant and complementary communities in multidimensional networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2181–2184. ACM.

[6] Berlingerio, M., F. Pinelli, and F. Calabrese (2013). Abacus: frequent pattern mining-based community discovery in multidimensional networks. *Data Mining and Knowledge Discovery 27*(3), 294–320.

[7] Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences 106*(50), 21068–21073.

[8] Boccaletti, S., G. Bianconi, R. Criado, C. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin (2014). The structure and dynamics of multilayer networks. *Physics Reports 544*(1), 1–122.

[9] Bollobás, B. and A. Universitet (1979). *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*. Aarhus Universitet.

[10] Cardillo, A., J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. Del Pozo, and S. Boccaletti (2013). Emergence of network features from multiplexity. *Scientific reports 3*.

[11] Celisse, A., J.-J. Daudin, L. Pierre, et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics 6*, 1847–1899.

[12] Chung, F. R. (1997). *Spectral graph theory*, Volume 92. American Mathematical Soc.

[13] Clauset, A. (2005). Finding local community structure in networks. *Physical review E 72*(2), 026132.

[14] Clauset, A., M. Newman, and C. Moore (2004). Finding community structure in very large networks. *Physical review E 70*(6), 066111.

[15] De Domenico, M., A. Lancichinetti, A. Arenas, and M. Rosvall (2014). Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems. *arXiv preprint arXiv:1408.2925*.

[16] De Domenico, M., A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas (2013). Mathematical formulation of multilayer networks. *Physical Review X 3*(4), 041022.

[17] Decelle, A., F. Krzakala, C. Moore, and L. Zdeborová (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E 84*(6), 066106.

[18] Fasino, D. and F. Tudisco (2016). Modularity bounds for clusters located by leading eigenvectors of the normalized modularity matrix. *arXiv preprint arXiv:1602.05457*.

[19] Ferriani, S., F. Fonti, and R. Corrado (2013). The social and economic bases of network multiplexity: Exploring the emergence of multiplex ties. *Strategic Organization 11*(1), 7–34.

[20] Fienberg, S. E., M. M. Meyer, and S. S. Wasserman (1980). Analyzing data from multivariate directed graphs: An application to social networks. Technical report, DTIC Document.

[21] Fienberg, S. E., M. M. Meyer, and S. S. Wasserman (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association 80*(389), 51–67.

[22] Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3), 75–174.

[23] Gleich, D. F. and C. Seshadhri (2012). Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 597–605. ACM.

[24] Greene, D., D. Doyle, and P. Cunningham (2010). Tracking the evolution of communities in dynamic social networks. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 176183.

[25] Han, Q., K. S. Xu, and E. M. Airoldi (2014). Consistent estimation of dynamic and multi-layer networks. *arXiv preprint arXiv:1410.8597*.

[26] Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

[27] Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter (2014). Multilayer networks. *Journal of Complex Networks 2*(3), 203–271.

[28] Lancichinetti, A., F. Radicchi, J. J. Ramasco, and S. Fortunato (2011). Finding statistically significant communities in networks. *PloS one 6*(4), e18961.

[29] Leskovec, J., K. J. Lang, A. Dasgupta, and M. W. Mahoney (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pp. 695–704. ACM.

[30] Lewis, A. C., N. S. Jones, M. A. Porter, and C. M. Deane (2010, Dec). The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology 4*(1), 114.

[31] Mossel, E., J. Neeman, and A. Sly (2012). Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*.

[32] Mucha, P. J., T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science 328*(5980), 876–878.

[33] Nadakuditi, R. R. and M. E. Newman (2012). Graph spectra and the detectability of community structure in networks. *Physical review letters 108*(18), 188701.

[34] Newman, M. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems 38*(2), 321–330.

[35] Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E 74*(3), 036104.

[36] Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*(23), 8577–8582.

[37] Onnela, J.-P., S. Arbesman, M. C. Gonzlez, A.-L. Barabsi, and N. A. Christakis (2011, Apr). Geographic constraints on social network groups. *PLoS ONE 6*(4), e16939.

[38] Palowitch, J., S. Bhamidi, and A. B. Nobel (2016). The continuous configuration model: A null for community detection on weighted networks. *arXiv preprint arXiv:1601.05630*.

[39] Parker, K. S., J. D. Wilson, J. Marschall, P. J. Mucha, and J. P. Henderson (2015). Network analysis reveals sex-and antibiotic resistance-associated antivirulence targets in clinical uropathogens. *ACS Infectious Diseases 1*(11), 523–532.

[40] Paul, S. and Y. Chen (2015). Community detection in multi-relational data with restricted multi-layer stochastic blockmodel. *arXiv preprint arXiv:1506.02699*.

[41] Paul, S. and Y. Chen (2016). Null models and modularity based community detection in multi-layer networks. *arXiv preprint arXiv:1608.00623*.

[42] Pons, P. and M. Latapy (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pp. 284–293. Springer.

[43] Porter, M. A., J.-P. Onnela, and P. J. Mucha (2009). Communities in networks. *Notices of the AMS 56*(9), 1082–1097.

[44] Raghavan, U. N., R. Albert, and S. Kumara (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E 76*(3), 036106.

[45] Rocklin, M. and A. Pinar (2013). On clustering on graphs with multiple edge types. *Internet Mathematics 9*(1), 82–112.

[46] Rohe, K., S. Chatterjee, B. Yu, et al. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*(4), 1878–1915.

[47] Snijders, T. A. and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification 14*(1), 75–100.

[48] Sporns, O. (2011). *Networks of the Brain*. MIT press.

[49] Stanley, N., S. Shai, D. Taylor, and P. Mucha (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE*.

[50] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

[51] Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*(397), 8–19.

[52] Wasserman, S. and J. Galaskiewicz (1994). *Advances in social network analysis: Research in the social and behavioral sciences*, Volume 171. Sage Publications.

[53] Wilson, J., S. Bhamidi, and A. Nobel (2013). Measuring the statistical significance of local connections in directed networks. In *NIPS Workshop on Frontiers of Network Analysis: Methods, Models, and Applications.*

[54] Wilson, J. D., S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel (2014). A testing based extraction algorithm for identifying significant communities in networks. *The Annals of Applied Statistics 8*(3), 1853–1891.

[55] Zhao, Y., E. Levina, J. Zhu, et al. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics 40*(4), 2266–2292.