Clustering Network Layers With the Strata Multilayer Stochastic Block Model

Natalie Stanley*†, Saray Shai†, Dane Taylor†, Peter J. Mucha†
*Curriculum in Bioinformatics and Computational Biology,
University of North Carolina, Chapel Hill
† Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics
University of North Carolina, Chapel Hill
stanleyn@email.unc.edu, {sshai, taylordr}@live.unc.edu, mucha@unc.edu

***Abstract*—Multilayer networks are a useful data structure for simultaneously capturing multiple types of relationships between a set of nodes. In such networks, each relational definition gives rise to a layer. While each layer provides its own set of information, community structure can be collectively utilized to discover and quantify underlying relational patterns. To most concisely extract information from a multilayer network, we propose to identify and combine sets of layers with meaningful similarities in community structure. In this paper, we describe the strata multilayer stochastic block model (sMLSBM), a probabilistic model for multilayer community structure. The assumption of the model is that there exist groups of layers, that we call "strata", with community structure described by a common stochastic block model (SBM). That is, layers in a stratum exhibit similar node-to-community assignments as well as SBM probability parameters. Fitting the sMLSBM to a multilayer network provides a joint clustering of nodes and layers with node-to-community and layer-to-strata assignments interactively aiding each other in inference. We describe an algorithm for separating layers into their appropriate strata and an inference technique for estimating the stochastic block model parameters describing each stratum. We demonstrate that our method works on synthetic networks and in a multilayer network inferred from human microbiome project data.**

***Keywords*—Stochastic Block Models, Clustering, Multilayer Networks, Strata, Probabilistic Models**

I. Introduction

Modeling relational information between a set of entities can often be successfully achieved through a network representation. Here, entities correspond to nodes and edges reflect a specific type of link between them. In many cases, there are multiple ways to define an edge that can be collectively analyzed for a more thorough understanding of the data. Multilayer networks provide a framework to do this, in that each relational definition leads to a layer in the network [1]. Such data and corresponding networks have shown to be useful in many contexts, such as, in the comparison of genetic and protein-protein interactions in a cell [2], in understanding underlying relationships and community structure across social networks [3], and in the analysis of temporal networks [4]. Thus, given the inherent mulitiplexity of network data across fields, there exists a need for the development of appropriate tools that can leverage information from all layers to elucidate structural patterns.

Each layer in the network provides its own information about interactions between nodes and it is useful to ask whether sets of layers are providing redundant information. Addressing this question requires the development of an approach to aggregate networks into a reduced-layer representation that still effectively conveys all of the information from the original multilayer network. Aggregating layers can potentially result in a loss of information, but can also successfully corroborate the existence of underlying structural patterns. This idea of reducibility in multilayer networks was previously explored in [5]; using an information theoretic notion of distance between pairs of networks, the authors performed hierarchical clustering of layers and chose the partition that maximized a quality function reflecting information loss due to aggregation of layers. While this approach reflects the validity and usefulness of combining layers, it does not result in a generative model describing the clusters of redundant layers. To further this intuition to a probabilistic framework, we have developed the strata multilayer stochastic block model (sMLSBM), which seeks to address this reducibility question by agglomerating sets of layers into structurally similar groups that we refer to as strata. Moreover, sMLSBM assumes that network layers in a given stratum have the same underlying generative model for community structure.

A. Similarities in Community Structure for Network Comparison

There are numerous ways to characterize and compare structure within and between networks, including motifs ([6],[7]), community structure [8], and network summary statistics ([9],[10]). In this work, we wish to analyze the layers in a multilayer network based on their community structure. Community detection in single-layer networks is an essential tool for understanding the organization and functional relatedness between nodes in a graph. Identifying communities in networks requires the identification of the best partitioning of nodes into groups to maximize number of within-community edges, which can be quantified by multiple approaches, including, modularity maximization [11], spectral methods [12], and through generative probabilistic models [13]. Because each of these approaches present computational challenges for efficiently detecting communities, numerous heuristics exist to accomplish these tasks ([14],[15],[16],[17]).

Here we consider the stochastic block model (SBM) [18], a popular generative model for community structure in networks.

The assumption of the SBM is that nodes in a particular community are related to nodes within and between communities in the same way. The inference procedure for fitting classical SBMs to an undirected network with n nodes and k communities involves learning the two parameters; $\boldsymbol{\pi}$ and $\mathbf{z}$: $\boldsymbol{\pi}$ is a $k \times k$ symmetric matrix, where π_{ij} gives the probability of an edge existing between nodes in communities i and j. Further, $\mathbf{z}$ is an n-length array indicating the community memberships for each node. These parameters are often inferred through a maximum likelihood approach and once learned, provide information about the within and between community relatedness. Given the usefulness of this model for the understanding of node organization in single-layer networks, it is natural to extend the intuition to multilayer networks. In this context, the assumption is that there are shared patterns in community structure across the layers of a multilayer network and the goal is to define a stochastic block model that captures this. We define the general notion of a probabilistic model characterizing a multilayer network as a multilayer stochastic block model (MLSBM).

B. Related Work in Multilayer SBMs

Providing an alternative to other methods for identifying communities in multilayer networks (e.g., maximizing multilayer modularity [4]), there have been many recent developments in related multilayer stochastic block models ([19],[20],[21],[22],[23]). Common to all of these approaches is that combining layers in the network in a principled way makes inference more accurate.

In [20], the authors define a version of MLSBM, where layers can be aggregated with different rules, such as through AND and OR conditions. They also provide an inference procedure for assessing whether or not a single-layer network is actually a projection of a multilayer network.

In [19], the authors explore asymptotic properties for inferring stochastic block model parameters in individual layers, by using information from all of the other layers, as the number of layers goes to infinity. As expected, as the number of layers increases, so does the quality of inference. Fitting their model to an n-node network with k communities requires learning an n-length vector $\mathbf{z}$ of community assignments across layers and a $k \times k$ matrix of block model probabilities, $\boldsymbol{\pi}^l$ for each individual layer, l. So, for a multilayer network with L layers, and k communities, there are $k(k+1)L/2$ total parameters to learn due to each $\boldsymbol{\pi}^{(l)}$, $l \in \{1, 2, \ldots, L\}$. Particularly, the authors extend the variational approximation for approximating the maximum likelihood estimates of SBM parameters introduced in single-layer SBMs introduced in [24] to the multilayer setting.

The authors of [23] refer to the model in [19] as MLSBM (multilayer stochastic model) and point out the problems with this approach as the number of communities grows quickly or if layers are sparse overall. To address these problems, they proposed a modification to the model, known as restricted multilayer stochastic block model (RMLSBM). In this model instead of learning a set of L π_{ij} components for each i, j pair, each entry in $\boldsymbol{\pi}$ is fully layer-dependent. In other words, to determine the probability of an edge between a node from community i and a node from community j in layer l, they use a logistic link function and model the probability as $\text{logit}(\pi_{ij}^{(l)}) = \pi_{z_i z_j} + \beta_l$. The β_l is an offset parameter representing the particular layer or type of edge. In this model it is necessary in an L layer graph with k communities to learn $k(k+1)/2+L$ total parameters. Thus, the maximum likelihood estimate for RMLSBM is a regularized estimator.

The approach in [21] is similar to [19] and [23] in that the authors seek to leverage information in all layers by considering the joint distribution of layers. Using this, they can estimate quantities such as the marginal probabilities of node assignments to communities and the edge probabilities within and between groups. An interesting aspect of their approach is that they introduce a covariate capturing the coupling between pairs of nodes. For a network with k communities and L layers, this requires the estimation of $(2^L-1)k^2+(k-1)$ parameters. They demonstrate the usefulness of their model for analysis of a collaboration network of French cancer researchers.

Finally, Peixto [22] describes two possible generative processes for multilayer networks, named edge covariate and independent model, respectively. In the edge covariate model, he defines a collapsed graph, where certain edges appear in particular layers. Collapsing the multilayer network combines all of the edges from each of the layers. Thus, turning this into a generative model involves choosing a layer membership for each edge and sampling edges with a probability conditioned on adjacent nodes. In the independent case, layers are generated independently from each other and the only constraint is that group membership of the nodes are the same across all layers. The models were defined using a Bayesian framework and hence lend themselves to the opportunity for statistically rigorous model selection.

C. Contributions

While the literature regarding multilayer stochastic block models for multilayer networks has recently grown quickly, there is still a need for a probabilistic generative model where there are multiple stochastic block models underlying the observed network. In this paper, we develop a novel strata multilayer stochastic block model, sMLSBM, that assigns individual layers to strata, where a collection of layers in a stratum is assumed to be derived from the same underlying generative model. Our method can be viewed as a joint clustering procedure, where we seek to group layers into strata with similar node-to-community assignments and stochastic block model probability parameters. Thus, using community membership information of nodes in all of the layers within a stratum helps to more accurately estimate the underlying stochastic block model probability parameters and vice versa. Given the large combinatorial challenge of choosing a stratum assignment and community assignment for each layer and node within a layer, respectively, we describe an algorithm that effectively partitions layers into strata and an inference procedure to learn the stochastic block model parameters for each stratum's stochastic block model.

To describe the model, the algorithm for fitting the model and its performance, the rest of this paper is organized as follows. In section II, we define the model and an algorithm for fitting it. In section III, we perform numerical experiments on synthetic networks, and in section IV, we test the model on correlation networks constructed from data from the human microbiome project.

II. sMLSBM: Strata Multilayer Stochastic Block Model

A. Network Definition

Let $G(N, \mathcal{E})$ define a single network with N nodes and a set of undirected edges, $\mathcal{E} = \{(i, j)\}$. Further, we define a multiplex network, which is one kind of multilayer network, $G^l(N, \mathcal{E}^l)$, for a particular layer l, where $l \in \{1, 2, \cdots, L\}$. We denote the collection of all L layers as a set, $\mathcal{G}$, such that $\mathcal{G} = \{\mathbf{G}^1, \mathbf{G}^2, \cdots, \mathbf{G}^L\}$

B. Model Definition

For a network with n nodes and k communities, the objective in a traditional (single-layer) stochastic block model is to is to learn a $k \times k$ matrix, $\boldsymbol{\pi}$, and an $n \times k$ binary matrix $\mathbf{Z}$. Here, the parameters $\boldsymbol{\pi}$ and $\mathbf{Z}$ provide information about the distribution of edges within and between groups and the community memberships of each node, respectively. In particular, π_{qt} represents the probability of an edge between a node in community q and one in community t. Z_{im} is an indicator variable for whether or not node i belongs to community m and $\sum_m Z_{im} = 1$.

Under the sMLSBM, the network layers, $G^l(N, \mathcal{E}^l)$ are generated by a set of S different stochastic block models, where stratum s is parameterized by $\boldsymbol{\pi}^s$ and $\boldsymbol{Z}^s$. Note that here the parameters $\boldsymbol{\pi}^s$ and $\mathbf{Z}^s$ for a single stratum are analogous in meaning to their respective parameters in the single-layer SBM case. However, since a stratum is composed of multiple networks, the parameters represent a consensus for that group. Thus, our objective during inference is to identify the stratum assignment of each layer and to learn the collection of strata parameters, $\mathbf{\Pi} = \{\boldsymbol{\pi}^1, \boldsymbol{\pi}^2, \ldots, \boldsymbol{\pi}^S\}$ and $\mathcal{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \ldots \mathbf{Z}^S\}$. The learned SBM parameters for a stratum represent a consensus for the associated layers, and so in that sense can be interpreted as reducing the effective number of layers (cf. [5]). However, strata can also be interpreted as a way to simply identify layers with similarities in community structure. Figure 1 shows a toy example of a multilayer network with 3 strata, where each layer has 36 nodes and 3 communities. Each graph in this figure represents a layer in the network. The nodes in the layers belonging to each stratum are colored according to their stratum membership; moreover, it is easy to see that members of a stratum exhibit high similarities in community structure. Thus, during inference, we would like to be able to take all of the layers in the network and partition them into their appropriate strata.

As part of our procedure, we specify another parameter that we refer to as the adjacency probability matrix, $\boldsymbol{\theta}^s$, which can be computed from $\boldsymbol{\pi}^s$ and $\mathbf{Z}^s$. Further, $\boldsymbol{\theta}^s$ is the $n \times n$ matrix, such that θ^s_{ij} gives the probability of an edge between the communities of nodes i and j in stratum s. That is, $\theta^s_{ij} = \pi^s_{z^s_i z^s_j}$, where z^s_i specifies the community number for node i in stratum s. Finally, we define the matrix $\mathbf{Y} = \{y_{ls}\}$ to be a binary matrix of indicators specifying whether or not layer l has been assigned to stratum s, such that $\sum_s y_{ls} = 1$.

C. Inference for sMLSBM

The procedure for fitting sMLSBM requires finding the layer-to-strata memberships and node-to-community memberships that best describe the multilayer network. In other words,

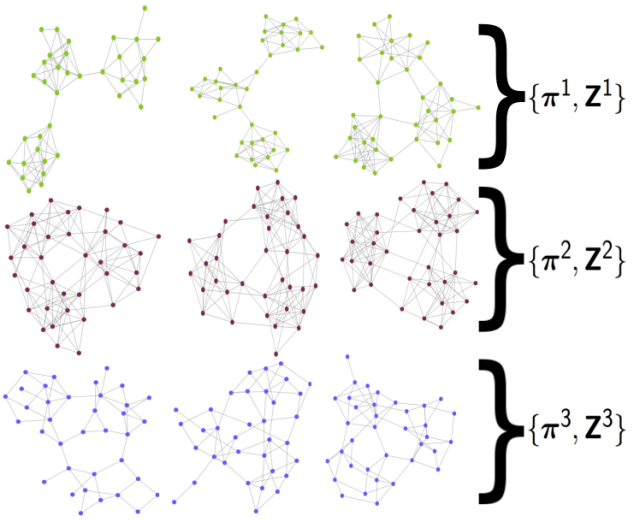

Fig. 1. **Objective of MLSBM.** Each of the 9 graphs here represents a layer in the network. Each graph has 36 nodes that are consistent across layers. The color of nodes in the network represents the stratum membership for that particular layer. Clearly, graphs within a stratum exhibit strong similarities in community structure. We would like to partition each layer into its appropriate stratum, and learn the associated parameters, $\boldsymbol{\pi}$ and $\mathbf{Z}$.

we can write down the marginal likelihood for the collection of graph layers, $\mathcal{G}$, as,

$$p(\mathcal{G} \mid \mathbf{\Pi}) = \sum_{\mathcal{Z}} \sum_{\mathbf{Y}} p(\mathcal{G}, \mathcal{Z}, \mathbf{Y} \mid \mathbf{\Pi}). \tag{1}$$

We assume the probability of an edge between two nodes in layer l belonging to stratum s can be modeled as a Bernoulli random variable, based on the community membership of the nodes. In particular, $p(G^l_{ij} = 1) \sim \text{Bernoulli}(\pi^s_{z_i z_j})$. Since the $\mathbf{Y}$ and $\mathcal{Z}$ are both latent quantities, summing over all of their possible values quickly becomes intractable. Thus, we have developed a two step approach to reduce the problem to only have the latent variable, $\mathcal{Z}$. In particular, we use clustering to come up with an estimate for $\mathbf{Y}$ that we can further use to infer $\mathcal{Z}$. We break this learning process in to two phases.

1) Phase I: Phase I for the fitting of sMLSBM to a network is comprised of two parts. First, a stochastic block model is fit to each individual layer, and then layers are clustered based on the similarities of their inferred block model parameters and node-to-community memberships, as specified by $\boldsymbol{\pi}$ and $\mathbf{Z}$, respectively. For the single-layer stochastic block model fits, we use the the inference method described in [24]. Here, the authors used a variational inference technique to approximate the maximum likelihood estimates for the stochastic block model parameters. For the set of L layers, this produces $\boldsymbol{\pi}$ and $\mathbf{Z}$ parameters for each layer, as denoted by $\mathbf{\Pi} = \{\boldsymbol{\pi}^1, \boldsymbol{\pi}^2, \ldots, \boldsymbol{\pi}^L\}$ and $\mathcal{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \ldots \mathbf{Z}^L\}$ (that is, at this stage of the procedure, each layer is temporarily treated as its own stratum). Using the fitted $\boldsymbol{\pi}^l$ and $\mathbf{Z}^l$ for a given layer, l, we can construct the corresponding adjacency probability matrix, $\boldsymbol{\theta}^l$. Doing this for each layer results in a collection of adjacency probability matrices, $\mathbf{\Theta} = \{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots, \boldsymbol{\theta}^L\}$. Now, we seek an initial partition of layers into strata, based on the adjacency probability matrices, where the total distance across strata between the stratum consensus adjacency probability matrix

and the adjacency probability matrices of stratum member layers is as small as possible. This is accomplished by treating each $\boldsymbol{\theta}$ as a feature vector and applying k-means clustering with S centers. S can be known *a priori*, or approximated with a measure such as the gap statistic [25]. This gives us an initial estimate for $\mathbf{Y}$. While this procedure initially treated each layer as a separate stratum, but provides a principled agglomeration of multiple layers into (ideally) fewer than L strata.

2) Phase II: After a first-pass approach for assigning layers to strata, we begin our iterative phase to more effectively estimate model parameters and the correct layer-to-strata assignments. Now, we would like to find the consensus parameters, $\boldsymbol{\pi}^s$, and $\mathbf{Z^s}$ that maximize the likelihood of the graphs in a particular stratum. We let $\mathcal{A}^s = \{\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^m\}$ denote the collection of networks corresponding to the m graphs in stratum s, where each $\mathbf{A}$ is the corresponding $n \times n$ adjacency matrix.

We now proceed to maximize the likelihood in each stratum, extending the framework of [24] to a multilayer context. Note that this is similar to [19], except that we are not aiming to infer an SBM probability matrix for each layer, individually. In particular, the complete data-log-likelihood for stratum s can be written as,

$$p(\mathcal{A}^s, \mathbf{Z}^s) = p(\mathcal{A}^s \mid \mathbf{Z}^s) p(\mathbf{Z}^s). \tag{2}$$

We now introduce a parameter α_q^s for stratum s and community q, representing the probability that a node in a layer in stratum s belongs to community q. So, α_q^s is $p(Z_{iq}^s) = 1$, with the constraint that $\sum_q \alpha_q^s = 1$. Further, we let $\mathcal{L}^s$ be the set of layers belonging to stratum s. Then,

$$p(\mathbf{Z}^s) = \prod_i \prod_q \alpha_q^{s(Z_{iq}^s)}. \tag{3}$$

Also,

$$p(\mathcal{A}^s \mid \mathbf{Z}^s) = \prod_{l \in \mathcal{L}^s} \prod_{i<j} \prod_{qt} {\pi_{qt}^s}^{A_{ij}^l} (1 - \pi_{qt}^s)^{(1 - A_{ij}^l)}. \tag{4}$$

Then, the complete-data log-likelihood for the graphs in stratum s can be expressed as,

$$\begin{aligned} \log P(\mathcal{A}^s, \mathbf{Z}^s) &= \log(P(\mathbf{Z}^s)) + \log(P(\mathcal{A}^s \mid \mathbf{Z}^s)) \\ &= \sum_i \sum_q Z_{iq}^s \log(\alpha_q^s) \\ &+ \sum_{l \in \mathcal{L}^s} \sum_{i<j} \sum_{qt} A_{ij}^l \log(\pi_{qt}^s) \\ &+ \sum_{l \in \mathcal{L}^s} \sum_{i<j} \sum_{qt} (1 - A_{ij}^l) \log(1 - \pi_{qt}^s). \end{aligned} \tag{5}$$

Problems of this variety involving the need to compute maximum likelihood estimates with incomplete data are typically addressed with the expectation maximization (EM) framework [26]. Doing so requires the ability to compute $P(\mathbf{Z}^s \mid \mathcal{A}^s)$; however, [24] showed calculating the conditional distribution is intractable on the single-layer network case. To address this challenge, we can use a variational approximation, as shown in ([24],[19],[21]). In general, the variational approximation seeks to optimize a lower bound on the log-likelihood. To do this, we first approximate the conditional distribution, $P(\mathbf{Z}^s \mid \mathcal{A}^s)$, with $R_{\mathcal{A}^s}$, where,

$$R_{\mathcal{A}^s}(\mathbf{Z}^s) = \prod_i h(\mathbf{Z}_{i\cdot}^s; \boldsymbol{\tau}_{i\cdot}). \tag{6}$$

Here, $\boldsymbol{\tau}^s = \{\tau_{iq}^s\}$ represents an approximation of the probability that node i belongs to community q in stratum s. Further, $h(\cdot)$ represents the multinomial distribution, with parameters, $\boldsymbol{\tau}_{i\cdot}^s$. Using this, we define the variational approximation as ,

$$\mathcal{J}(R_{\mathcal{A}^s}) = \ell\ell(\mathcal{A}^s) - \mathrm{KL}(R_{\mathcal{A}^s}(\mathbf{Z}^s), P(\mathbf{Z}_s \mid \mathcal{A}^s)). \tag{7}$$

Here, $\ell\ell$ represents log likelihood and KL is the Kullback-Leibler divergence.

Through maximizing $\mathcal{J}(R_{\mathcal{A}^s})$, we minimize the KL divergence between the true conditional distribution, $P(\mathbf{Z}_s \mid \mathcal{A}^s)$, and its approximation, $R_{\mathcal{A}^s}(\mathbf{Z}^s)$. Moreover, we follow the derivation in [24] and rewrite $\mathcal{J}(R_{\mathcal{A}^s})$ as,

$$\begin{aligned} \mathcal{J}(R_{\mathcal{A}^s}) &= \sum_i \sum_q \tau_{iq}^s \log(\alpha_q^s) \\ &+ \sum_{l \in \mathcal{L}^s} \sum_{i<j} \sum_{qt} \tau_{iq}^s \tau_{jt}^s [A_{ij}^l \log(\pi_{qt}^s)] \\ &+ \sum_{l \in \mathcal{L}^s} \sum_{i<j} \sum_{qt} \tau_{iq}^s \tau_{jt}^s [(1 - A_{ij}^l) \log(1 - \pi_{qt}^s)] \\ &- \sum_i \sum_q \tau_{iq}^s \log(\tau_{iq}^s). \end{aligned} \tag{8}$$

We can now differentiate $\mathcal{J}(R_{\mathcal{A}^s})$ and use Lagrange multipliers to enforce constraints (i.e. probabilities summing to 1), with respect to each parameter to compute the updates. Doing so yields the following, where the hat notation symbolizes the current best estimate for the given parameter:

$$\hat{\alpha_q^s} = \sum_i \hat{\tau}_{iq}^s / n\,, \tag{9}$$

$$\hat{\pi}_{qt}^s = \sum_{l \in \mathcal{L}^s} \frac{\sum_{i<j} \hat{\tau}_{iq}^s \hat{\tau}_{jt}^s A_{ij}^l}{\sum_{i<j} \hat{\tau}_{iq}^s \hat{\tau}_{jt}^s}\,, \tag{10}$$

$$\hat{\tau}_{iq}^s \propto \hat{\alpha_q^s} \prod_{l \in \mathcal{L}^s} \prod_{i<j} \prod_t [\hat{\pi}_{qt}^s{}^{A_{ij}^l} (1 - \hat{\pi}_{qt}^s)^{1 - A_{ij}^l}]^{\hat{\tau}_{jt}^s}\,. \tag{11}$$

We alternate between updating $\hat{\boldsymbol{\tau}^s}$ and $\hat{\boldsymbol{\pi}^s}$ until convergence. When convergence has occurred, we refer to the resulting estimates as the consensus $\boldsymbol{\tau}^s$ and $\boldsymbol{\pi}^s$ for stratum s. Since $\boldsymbol{\tau}^s$ and $\boldsymbol{\pi}^s$ are computed in terms of each other, we can use one of the consensus parameters to compute the other parameter in individual layers. This allows us to determine whether or not the stratum consensus estimates affect the estimation of analogous parameters in the single-layer case. Particularly, it indicates whether the stratum consensus parameters are proper

descriptions of the node-to-community assignments and the stochastic block model parameters in single layers of the stratum. We are now able to represent each layer by the adjacency probability matrix computed in two different ways, letting $\boldsymbol{\theta}(\boldsymbol{\tau}, \boldsymbol{\pi})$ represent the $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$ being used to compute the adjacency probability matrix for layer l. Specifically, we compute,

$$\boldsymbol{\theta}^l_{(1)} = \boldsymbol{\theta}^l(\boldsymbol{\tau}^s, \boldsymbol{\pi}^l) \tag{12}$$

with the $\boldsymbol{\pi}$ that provides the best match to layer l using information about node-to-community assignments given by $\boldsymbol{\tau}^s$.

$$\boldsymbol{\theta}^l_{(2)} = \boldsymbol{\theta}^l(\boldsymbol{\tau}^l, \boldsymbol{\pi}^s) \tag{13}$$

with the $\boldsymbol{\tau}$ that provides the best match to layer l using information about the stochastic block model probabilities given by $\boldsymbol{\pi}^s$.

That is, for each layer in stratum s, $\theta^l_{(1)}$ uses the consensus $\boldsymbol{\tau}$ computed for stratum s and the $\boldsymbol{\pi}$ computed to best fit a particular layer. Conversely, $\boldsymbol{\theta}^l_{(2)}$ uses the consensus $\boldsymbol{\pi}$ from the stratum paired with the single-layer estimates for $\boldsymbol{\tau}$ to compute the adjacency probability matrix for each layer in the stratum.

During the phase I, we took the adjacency probability matrix for each of the L layers and clustered these matrices using k-means clustering. We employ a similar procedure here, but instead of clustering L matrices, we now cluster $2L$ matrices, since each layer is represented in two different ways. Moreover, clustering these $2L$ matrices yields two cluster assignments for each layer. The total number of partition combinations induced by the two representations of each layer determines the number of strata in the next iteration. Ideally, both representations of a layer will receive identical cluster assignments for an individual layer. However, an interesting case arises when the two representations induce different stratum assignments on the same layer because this implies that the consensus $\boldsymbol{\pi}$ and single-layer $\boldsymbol{\tau}$ (and vice versa) do not have sufficient agreement. We iterate phase II until the assignment of layers into strata does not change. Theoretically, new strata could arise in every iteration, so one could specify a maximum number of iterations to terminate the process. However, we did not observe this problem in any of our synthetic or real data experiments.

III. Numerical Experiments

A. Comparison of sMLSBM to other SBM Approaches

To demonstrate a situation where sMLSBM would be useful, we designed a synthetic experiment and compared the results of using different SBM approaches: i) fitting a stochastic block model to each layer individually (denoted single-layer SBM), and ii) fitting a single SBM to all of the layers (denoted single SBM). We generated a multilayer network, where each layer has $n = 200$ nodes, $k = 5$ communities and a mean degree, c, of $c = 25$. We specified an sMLSBM with 3 strata and 10 layers per strata. Note that this results in 30 total layers. We define the $\boldsymbol{\pi}^s$ for each stratum in terms of two parameters, p_{in} and p_{out}, giving the within-community edge probabilities and between-community edge probabilities, respectively. That is, the diagonal elements of each $\boldsymbol{\pi}$ are p_{in} and the off-diagonal values are p_{out}. The p_{in} values for strata 1,2 and 3 are assigned to be .6,.45 and .35, respectively. Given the mean degree for networks belonging to each stratum is 25, this gives corresponding values of 0.00625, 0.04375, and 0.06875 for p_{out}. In figure 2 A, we show a sample of a graph plotted from strata 1, 2, and 3 in panels i, ii and iii, respectively. Nodes are colored by their community assignments in stratum 1. Further, we can see that the node-to-community assignments are different in each stratum and that the extent of block structure decreases from stratum 1 to stratum 3.

We attempt to learn parameters and community assignments for each layer with 3 methods: the single-layer SBM involves fitting an SBM to each layer separately, the single SBM fits one SBM across all of the layers, and strata SBM fits an SBM to each stratum. First, for each layer, we quantified the error ($\ell 2$ norm) between its true $\boldsymbol{\pi}$ parameter, and the $\boldsymbol{\pi}$ learned to describe it under each of the 3 models. The mean errors across layers under each model are shown in figure 2 B. Second, we computed the normalized mutual information (NMI) [27] between the true $\mathbf{Z}^s$, (node-to-community assignments) and the inferred $\mathbf{Z}^s$ under each model. Figure 2 C shows the mean NMI for community assignments across layers.

B. Two-Strata Synthetic Experiment

To test how sMLSBM performs in comparison to a baseline k-means clustering of adjacency matrices and for different p_{in} and p_{out} parameters, we created a synthetic experiment with 2 strata. In each simulation, layers in the network have 200 nodes, 5 communities and a mean degree of 50. Additionally, there are 100 total layers and 50 were assigned to each stratum. Thus, in experiments in figure 3 we hold p_{in} and consequently p_{out} constant at 0.5 and 0.19, respectively in stratum 1. Then in each separate experiment (horizontal axis), we choose a p_{in} value for stratum 2 and again choose the corresponding p_{out} value such that the mean degree is 50.

In figure 3 A we quantify the quality of layer-to-strata assignment with normalized mutual information. Each numerical experiment consists of running 50 simulations with a p_{in} value specified by the horizontal axis. Fitting sMLSBM to the networks in a particular simulation results in a vector of strata assignments, $\hat{\mathbf{y}}$. Thus, we compute the NMI between $\hat{\mathbf{y}}$ and the true $\mathbf{y}$ for the strata (green curve). As a baseline we also perform k-means clustering of the adjacency matrices (purple curve). We can see that as the p_{in} of model 2 approaches the p_{in} of model 1 inference becomes more difficult, as expected. We can also see that in the experiments having a p_{in} parameter close to 0.5 in model 2, sMLSBM dramatically outperforms k-means.

Since phase II of fitting sMLSBM is an iterative process, we recorded the mean number of iterations required for the strata assignments to converge. Figure 3 B plots the mean number of iterations (vertical axis) against the corresponding experimental parameter. Again, for p_{in} values close to 0.5, more iterations are required for sMLSBM to converge.

Finally, we analyzed the mean quality of node to community assignments across all layers. That is, after fitting sMLSBM, each layer was assigned to a particular stratum

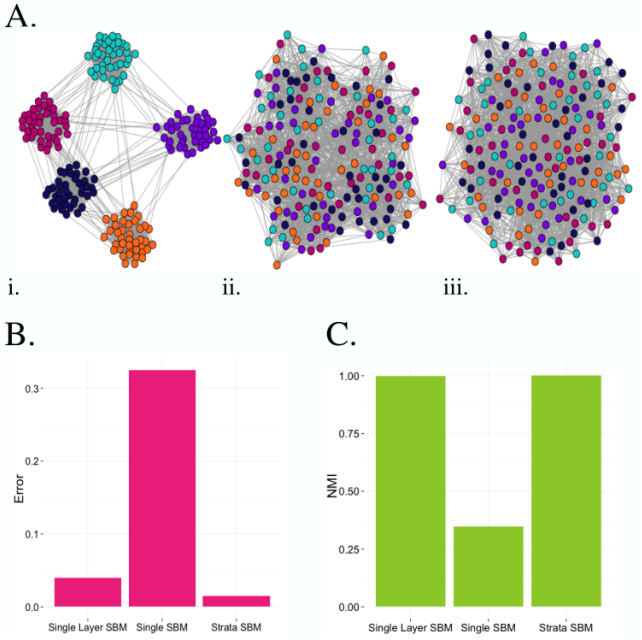

Fig. 2. **Mixture of SBMs Synthetic Experiment. A**. We specified a model with 3 strata and 10 layers per stratum. Panels i, ii and iii represent sample networks from strata 1, 2 and 3, respectively. Note that nodes in all networks are colored according to their community membership in graph i. Each network has $n = 200$ nodes, mean degree, $c = 25$. The p_{in} parameters for strata 1,2 and 3 are .6, .45 and .35, respectively. Corresponding values of p_{out} were selected to maintain the desired mean degree. **B**. We fit 3 types of models to the 30 networks such that each model yields a representation, $\boldsymbol{\pi}^l$ for layer l. The three models fit are 1) Single Layer SBM: fitting an individual SBM to each layer, 2) Single SBM: fitting a single SBM to all of the layers, and 3) Strata SBM: fitting SBMs based on strata memberships. On the vertical axis we plot the mean $\ell 2$ norm error between each layer's true underlying $\boldsymbol{\pi}^l$ and that inferred under the given model. **C**. For the community assignments inferred for each layer, given by $\mathbf{Z}^l$, under each of the 3 models, single-layer SBM, single SBM and strata SBM, we computed the normalized mutual information (NMI) between the true $\mathbf{Z}^s$ and $\mathbf{Z}^l$.

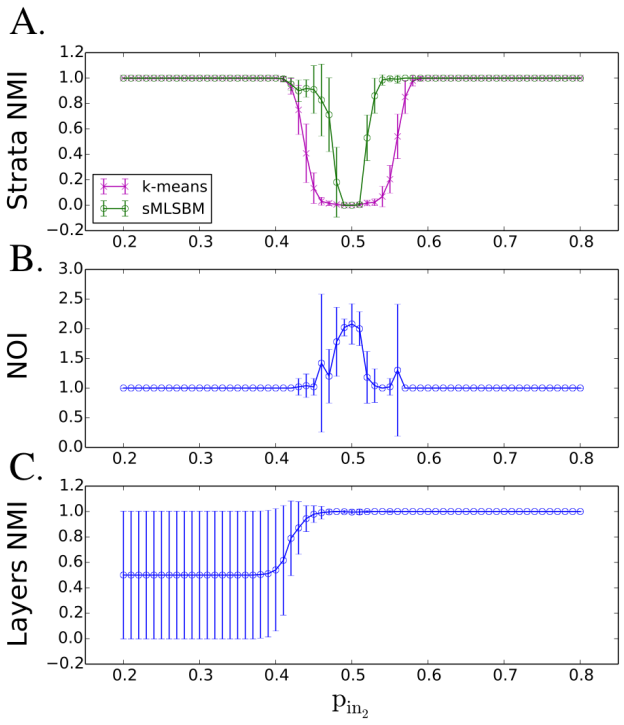

Fig. 3. **2 Stratum Synthetic Experiment.** We considered numerical experiments consisting of multilayer networks with 200 nodes, 2 strata and 50 layers per stratum. Within-community edge probability, $p_{in} = .5$ for stratum 1 and a corresponding p_{out} was chosen such that the mean degree, $c = 50$. Numerical experiments consisted of varying the within-community edge probability (p_{in}) for stratum 2, and measuring 3 quantities. Results shown correspond to mean and standard deviation obtained from 50 random networks. **A.** As a baseline, we compared the performance of sMLSBM to k-means clustering of the adjacency matrices. Curves show the mean NMI across 50 simulations and error bars show standard deviation. **B.** The mean number of iterations (NOI) required for sMLSBM to converge. **C.** Each numerical experiment resulted in a node-to-community membership vector for each layer, depending on its strata assignment. We used NMI to compare this vector to the true node-to-community assignments in each layer. Here we have plotted the mean NMI across layers as a function of experimental parameter.

with the corresponding community membership parameter, $\mathbf{Z}^s$. Thus, we let the community memberships for all layers within the stratum be represented by this parameter and computed the NMI between the inferred $\mathbf{Z}^s$ representation and the true $\mathbf{Z}^s$. Figure 3 C shows the mean NMI (vertical axis) against the p_{in} experiment parameters for model 2. As reflected by the low NMI, detectability of communities is difficult in networks where the p_{in} for model 2 is less than 0.5.

IV. Correlation Networks from the Human Microbiome

As a real-world motivation for sMLSBM, we consider correlation networks constructed from data from the human microbiome project [28]. For various sites on the body, the human microbiome project has successfully collected multiple human samples in order to better understand interactions between bacterial species. In this context, network inference is particularly interesting, as such methods aim to capture the signed relationships between various organisms. Microorganisms exhibit intricate ecologies within the gut of their human host and particular body sites have been shown to possess characteristic interactions. Further, certain interactions between microbes can often be associated with particular health and disease states [29]. Microbiome data is typically collected through metagenomic sequencing and reads are further binned into groups, known as operational taxonomic units (OTUs), to represent particular organisms. The nature of this count-based sequencing data makes network inference challenging, and is thus an interesting field in itself. To demonstrate the potential use for sMLSBM in the context of the human microbiome, we applied MLSBM to networks constructed from the SparCC [30] network inference method.

SparCC is a correlation network inference method that aims to approximate the linear Pearson correlation between components in a system. Their method accounts for the extent of diversity in the microbial community, which plays a significant role in detecting valid interactions. Furthermore, networks are constructed with the assumptions that the number of components in the system (e.g. OTUs) is large and that the correlation network should be sparse. As supplemental data in their paper, the authors provided their inferred microbial interaction networks for 18 sites in the human body. The edges in these networks have positive and negative real-valued weights, based on the results of SparCC inference. In

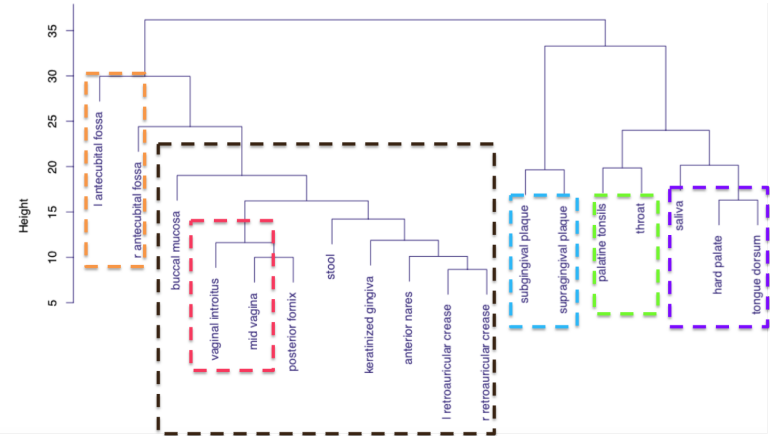

Fig. 4. **Hierarchical Clustering of SparCC Networks** Hierarchical clustering was performed on the thresholded binary adjacency matrices corresponding to the interactions between OTUs at each body site. Leaves of the tree correspond to body sites and colored boxes around the leaves indicate the strata assignment, according to sMLSBM. The variability in hierarchical clustering results, depending on where the tree is cut, highlights the usefulness of fitting sMLSBM.

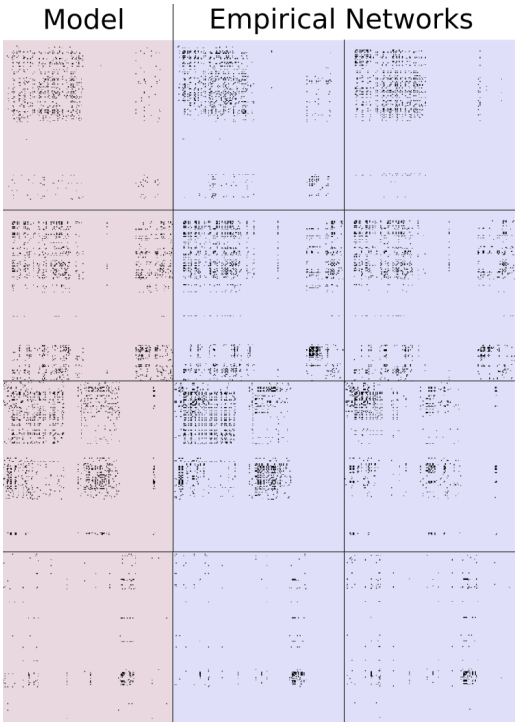

Fig. 5. **Visualization of Strata in SparCC Networks.** We visualize the adjacency matrices corresponding to the SparCC networks corresponding to body sites in each stratum (empirical networks) in blue, as well as a sample network generated from the inferred stratum model parameters, $\boldsymbol{\pi}^s$ and $\mathbf{Z}^s$, in pink. Black dots indicate an edge in the adjacency matrix. The following gives a legend for the networks in the figure in terms of the row and column (c) indices (rows go top to bottom). **row 1:** c2-saliva, c3-tongue dorsum. **row 2:** c2-subgingival plaque, c3-supra gingival plaque. **row 3:** c2-l antecubital fossa, c3-r antecubital fossa. **row 4:** c2-mid vagina, c3-vaginal introits.

this analysis, we converted the SparCC networks into binary adjacency matrices by allowing a link only if the SparCC edge-weight between two OTUs was at least 0.2. To convert the 18 single-layer networks corresponding to species interactions in 18 body sites, we found the collection of nodes (OTUs) that occurred in at least 2 of the layers. This resulted in 213 unique OTUs (nodes) for our multilayer network analysis.

We ran sMLSBM on the multilayer network and found 6 strata. The partitions of layers (body sites) into strata was interesting because similar body sites tended to group together. We compared the sMLSBM results to those obtained performing hierarchical clustering of the networks. Figure 4 shows the dendrogram, depicting the hierarchical clustering result. Also captured in this figure at the sMLSBM results. Leaves correspond to body sites and the colored boxes indicate strata assignments assigned with sMLSBM. We note that the orange, red, blue, green, and purple strata (as colored in the figure) are appropriate in terms of their location in the body. For example, it makes sense that the saliva, hard palate and tongue dorsum layers have very similar microbe species interaction networks. However, the brown stratum seems to be a miscellaneous cluster. Using the dendrogram to compare the sMLSBM results with hierarchical clustering, we see that the quality of the clustering partition is highly dependent on where the tree is cut. It is difficult to find a cut of the tree that partitions the body sites in a way that is as meaningful as the result of fitting sMLSBM. Moreover, fitting sMLSBM also provides a generative model for each stratum.

To further visualize the quality of sMLSBM, we can highlight 4 of the 6 strata. Each row of figure 5 provides information about the networks and their fitted sMLSBM models in a particular stratum. Each grid in the figure represents the binary adjacency matrix encoding interactions between OTUs. Edges, or a 1 in the matrix, are colored black. The first column of each row (pink) is a sample network generated with the learned model parameters of that stratum, $\boldsymbol{\pi}^s$ and $\mathbf{Z}^s$. Columns 2 and 3 (blue) show adjacency matrices for two representative layers within the stratum. Note that while some strata have more than two members, we only show two example members in this figure, for illustrative purposes. It is easy to see the block structure that naturally arises in all of the networks, corroborating the usefulness of the stochastic block model. The caption of figure 5 provides an indication of what each grid represents. Finally, model sample networks closely mimic the members of its stratum in terms of community structure.

V. Conclusion and Future Work

We developed a novel model for multilayer stochastic block models and an associated algorithm to jointly partition layers to strata and nodes to communities. Our model assumes that layers belonging to a stratum have the same underlying stochastic block model for community structure. To fit sMLSBM to a multilayer network, we iteratively alternate between rearranging layer-to-strata assignments and updating the model parameters for each stratum. Having multiple networks within a stratum and hence multiple realizations from some underlying model helps to make inference more accurate. Particularly, more accurate assignments of nodes-to-communities within a stratum leads to improved estimation of SBM probability parameters, and vice versa. If layers from different models were all considered to have arisen from the same SBM, both the community memberships and SBM parameters used to represent each layer would be noisier and inaccurate. Our model allows for an understanding of the similarities between

layers in a network, in terms of their community structure. The ability to identify strata within collections of networks holds promise in numerous applications.

There are several extensions to sMLSBM that could make the approach more accurate and applicable to a wider range of applications. First, as always in the context of stochastic block models, it is useful to consider the degree-corrected [31] and overlapping community [31] varieties. Next, sMLSBM as implemented here, is only appropriate in unweighted, undirected networks. Extensions analogous to weighted and directed networks as shown in [32] and [33] could be quite useful. Next, we could consider the case where there exist layers that should not belong to any stratum and should be assigned to singleton clusters.

Finally, the microbiome example shown in the paper reveals some interesting computational biology questions that could facilitate the development of more advanced network tools. To construct the multilayer network, negative edges were thresholded away. However, an understanding of antagonistic relationships between microbes is interesting. Thus, it would be useful to develop a signed version of sMLSBM, where edges could be modeled as either positive, or negative.

The rise of a greater number of multilayer network datasets is providing the need for additional tools for the construction and analysis of such networks. The sMLSBM provides a new method to find signal in inherently noisy and complex network data.

ACKNOWLEDGMENTS

We thank James Wilson for helpful discussions about related work in multilayer networks and in multilayer stochastic block models. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD075712, the James S. McDonnell Foundation 21st Century Science Initiative Complex Systems Scholar Award grant # 220020315 and training grant T32 GM 067553 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

[1] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[2] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi *et al.*, "The genetic landscape of a cell," *science*, vol. 327, no. 5964, pp. 425–431, 2010.

[3] D. Greene and P. Cunningham, "Producing a unified graph representation from multiple social network views," in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 118–121.

[4] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.

[5] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Layer aggregation and reducibility of multilayer interconnected networks," *arXiv preprint arXiv:1405.0425*, 2014.

[6] M. M. Hasan, Y. Kavurucu, and T. Kahveci, "A scalable method for discovering significant subnetworks," *BMC systems biology*, vol. 7, no. Suppl 4, p. S3, 2013.

[7] K. Tsuda and T. Kudo, "Clustering graphs by weighted substructure mining," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 953–960.

[8] J.-P. Onnela, D. J. Fenn, S. Reid, M. A. Porter, P. J. Mucha, M. D. Fricker, and N. S. Jones, "Taxonomies of networks from community structure," *Physical Review E*, vol. 86, no. 3, p. 036104, 2012.

[9] U. Brandes, J. Lerner, and U. Nagel, "Network ensemble clustering using latent roles," *Advances in Data Analysis and Classification*, vol. 5, no. 2, pp. 81–94, 2011.

[10] U. Brandes, J. Lerner, U. Nagel, and B. Nick, "Structural trends in network ensembles," in *Complex networks*. Springer, 2009, pp. 83–97.

[11] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[12] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi, "A local spectral method for graphs: With applications to improving graph partitions and exploring data graphs locally," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2339–2365, 2012.

[13] A. Z. Jacobs and A. Clauset, "A unified view of generative models for networks: models, methods, opportunities, and challenges," *arXiv preprint arXiv:1411.4070*, 2014.

[14] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[15] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[16] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[17] A. Clauset, C. Moore, and M. E. Newman, "Structural inference of hierarchies in networks," in *Statistical network analysis: models, issues, and new directions*. Springer, 2007, pp. 1–13.

[18] T. A. Snijders and K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of classification*, vol. 14, no. 1, pp. 75–100, 1997.

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

[20] T. Valles-Catala, F. A. Massucci, R. Guimera, and M. Sales-Pardo, "stochastic block models reveal the multilayer structure of complex networks," *arXiv preprint arXiv:1411.1098*, 2014.

[21] P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen, "Stochastic block models for multiplex networks: an application to networks of researchers," *arXiv preprint arXiv:1501.06444*, 2015.

[22] T. P. Peixoto, "Inferring the mesoscale structure of layered, edge-valued and time-varying networks," *arXiv preprint arXiv:1504.02381*, 2015.

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

[24] J.-J. Daudin, F. Picard, and S. Robin, "A mixture model for random graphs," *Statistics and computing*, vol. 18, no. 2, pp. 173–183, 2008.

[25] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[27] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[28] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[29] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial co-occurrence relationships in the human microbiome," *PLoS computational biology*, vol. 8, no. 7, p. e1002606, 2012.

[30] J. Friedman and E. J. Alm, "Inferring correlation networks from

genomic survey data," *PLoS computational biology*, vol. 8, no. 9, p. e1002687, 2012.

[31] B. Karrer and M. E. , "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016107, 2011.

[32] C. Aicher, A. Z. Jacobs, and A. Clauset, "Learning latent block structure in weighted networks," *Journal of Complex Networks*, vol. 3, no. 2, pp. 221–248, 2015.

[33] Y. J. Wang and G. Y. Wong, "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 8–19, 1987.