Stochastic Block Models with Multiple Continuous Attributes

Natalie Stanley, Thomas Bonacci, Roland Kwitt, Marc Niethammer, Peter J. Mucha

ABSTRACT

Stochastic block models (SBMs) are probabilistic models for community structure in networks in which nodes within a community are assumed to be connected to nodes within and between communities in a uniform, characteristic way. Typically, only the adjacency matrix is used to perform SBM parameter inference. In this paper, we consider circumstances in which nodes have an associated vector of continuous attributes. While this assumption is not realistic for every application, our model assumes that the attributes associated with the nodes in a network's community can be described by a particular multivariate Gaussian model. Moreover, in this augmented, attributed SBM, the objective is to learn the SBM connectivity probability parameters and the multivariate gaussian parameters describing each community. While there are recent examples in the literature that combine connectivity and attribute information, we are to our knowledge the first to consider the effect of multiple continuous attributes. We highlight the usefulness of our model for two two network prediction tasks: collaborative filtering and link prediction. As a result of fitting this attributed stochastic block model, one can predict the attribute vector or connectivity patterns for a new node in the event of the complementary source of information (connectivity or attributes, respectively). While our approach is the first stochastic block model (to our knowledge) to include multiple continuous attributes, we highlight the ability of our approach in two tasks: link prediction in collaborative filtering. In this work, we demonstrate the usefulness of our model in two different types of biological networks.

Keywords

Stochastic Block Model, Networks, Community Detection

1. INTRODUCTION

1.1 Network data and node attributes

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

Uncovering patterns in network data is a common pursuit across a range of fields, such as in biology [12], medicine [1, 8] and computational social science [7]. A powerful way to analyze mesoscale structural organization within a network is with community structure [16, 13, 17]. In this pursuit, the objective is to identify cohesive groups of nodes with a high density of within-group connections and few between-group connections. Numerous approaches exist to accomplish this task [16, 13], but typically only the adjacency matrix conveying the wiring of the graph is taken into account. In certain applications, each node in a network is equipped with additional information (or particular attributes) that was not implicitly taken into account in the construction of the network. For example, one could consider a collection of attributes measured for individuals in a social network (i.e., age, interest, level of education). The interplay between the connectivity-based (or structural) community organization of the network and attribute information has gained attention, as it is often unclear whether it is valid to assume that a *structural* community should necessarily correlate with an attribute-based *functional* community [10, 15, 18]. While these studies suggest that extreme caution should be taken in assuming a correlation between structural and functional communities, we limit our focus in this work to the assumption that a node's connectivity and attribute patterns can be jointly modeled based on its community assignment. In other words, we sought to develop an approach to assign node to communities based jointly on both sources of information, such that a community is defined as a group of nodes with similar connectivity and attribute patterns. Moreover, our objectives are two-fold. First, we seek to develop a probabilistic approach to jointly model connectivity and attributes. Second, we wish to ensure that our model can handle multiple, continuous attributes.

1.2 Related work

Recently, there have been numerous efforts to incorporate attribute information into the community detection problem [19, 14, 4, 10, 15]. First, we focus on an effective quality function based method, known as I-louvain [4]. This method approaches the problem as an extension to the Louvain algorithm, which is the state-of-the-art scalable quality function modularity-based community detection method [2]. The traditional modularity-based approach to community detection defines a null model for community structure under the assumption that there is not substantial structural organization in the network and seeks to identify a partition maximally different from this model through optimizing a quality function (modularity). The I-louvain method modi-

fies the standard modularity quality function as what they call 'inertia-based modularity'. This quality function incorporates the euclidean distance between nodes based on their attributes. This work demonstrates that incorporating connectivity and attributes allows for a partition of nodes to communities that aligns with ground truth better than the results obtained based on connectivity or attributes in isolation.

Alternatively, a variety of methods [14, 10, 15, 19] are probabilistic in nature and seek to learn a probabilistic model to describe how edges and attributes are generated using some aspect of community structure. Most similar to our work is [19], which seeks to learn a set of propensities or affiliations for each node across all possible communities, such that two nodes with similar propensities towards communities should have more in common in terms of connectivity and attributes. In this model, each node has a vector multiple binary attributes. The affiliation model is useful and flexible because it does not enforce a hard partitioning of nodes into communities, which is useful in social network applications. In this inference problem, the connectivity and attribute information are used to infer a node's affiliations to communities and then models the probability of an edge between two nodes as a function of the similarity in their community affiliation propensities.

In contrast to the affiliation model, the stochastic block model [9] (at least the more standard variants of it), seeks to determine a hard partition for each nodes across communities and models edges between a pair of nodes according to their community assignments. The partition of nodes to communities through a stochastic block model framework is accomplished through maximum likelihood optimization. The adaptation of the stochastic block model was first explored by Clauset *et al.*, [14], who adapts the stochastic block model to handle a single attribute with the assumption that metadata and communities are correlated. Next Hric *et al.*,, developed an attributed SBM [10] from a multilayer network perspective, with one layer modeling relational information between attributes and the other modeling connectivity. The authors then seek to assign nodes to communities maximizing the likelihood of the observed data in each layer. Finally, work of Peel *et al.* made important contributions in 1) establishing a statistical test to determine whether metadata actually correlates with community structure and 2) developing an SBM with flexibility in how strongly to couple metadata and community membership in the stochastic block model inference problem [15].

The incorporation of node attributes in network analysis tasks is not well explored, particularly in the context of probabilistic network models. In a modularity based approach, the authors of [4] modified the classical modularity quality function to incorporate node attribute information. Two probabilistic network and attribute models have also been proposed by [19, 14]. In [19], the authors developed a generative model for attributed network data based on node affiliation. However, in this approach only binary attribute data are considered. In [14], the authors developed a probabilistic framework for incorporating metadata or attributes in community detection, however, the model takes only a single metadata value for each node into account. We seek to extend these existing approaches to reflect continuous attribute data and particularly multidimensional attribute data drawn from a multivariate Gaussian distribution.

New stuff to add: peel, clauset controversy, tiago and dhric

1.3 Stochastic Block Models

To accomplish our objective of integrating node attributes in a probabilistic framework, we will model connectivity in a network with the widely-used stochastic block model (SBM) [9]. This model assumes that edges within a community are connected within and between communities in a characteristic or probabilistic way. To fit this model to network data, the objective is to partition the nodes into communities, such that these assignments maximize the likelihood of the model according to the observed edges. In this inference problem for a network with N nodes and K communities, a $K \times K$ probability matrix, **theta** and a N-length vector, $\mathbf{z}$ are learned. The matrx θ described the probability of connections within and between communities, while $\mathbf{z}$ gives the node-to-community assignments. While this modeling framework is well-studied, recent attention hs focused on the ability to integrate extra information (attributes or metadata), into the inference problem and whether it is appropriate to do so.

1.4 Contributions and Paper Outline

We summarize our contributions to the problem of finding communities in attributed network data as follows.

- We developed a probabilistic model for node-to-community structure that uses adjacency matrix connectivity and continuous node attributes to determine a community assignment for the nodes. Our model is the first to our knowledge to allow for the augmentation of a classic stochastic block model with multiple continuous attributes.
- We provide details for the inference technique that can be used to effectively estimate model parameters.
- We demonstrate that by fitting an attributed stochastic block model allows for good performance in link prediction and collaborative filtering results. In particular, we demonstrate these tasks on two different biological examples.

The paper is organized as follows. In section 2, we describe the model and inference methods for estimating parameters. In section 3, we perform experiments on synthetic networks. Section 4 shows how the method can be applied in image segmentation tasks. Finally in section 5 we discuss the method, its limitations and future work.

2. MODEL

2.1 Objective

We seek to incorporate both connectivity ($\mathbf{A}$) and attribute information ($\mathbf{X}$) to infer node-to-community assignments, $\mathbf{Z}$. Note that for a network with N nodes, K communities and p measured attributes, $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{Z}$ have dimensions $N \times N$, $N \times p$ and $N \times K$, respectively. In particular, $\mathbf{Z}$ is a binary indicator matrix, where entry z_{ic} is 1 if and only if node i belongs to community c. We also define $\mathbf{z}$ to be the N-dimensional vector of node-to-community assignments. Following the assumption of [19], we assume

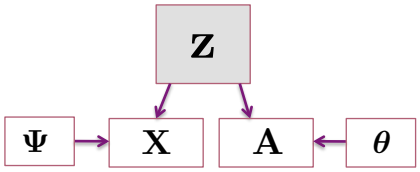

Figure 1: **Modeling community membership in terms of attributes and connectivity**. Node-to-community assignments specified by $\mathbf{z}$ are determined in terms of adjacency matrix information, $\mathbf{A}$ and attribute matrix information, $\mathbf{A}$. $\mathbf{A}$ and $\mathbf{X}$ are assumed by be generated from a stochastic block model and a mixture of multivariate gaussian distributions, parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$, respectively.

that connectivity and attributes are conditionally independent, given the community membership label. The graphical model for the relationship between node-to-community labels, connectivity and attribute information is shown in Fig. 1.

To infer the $\mathbf{Z}$ that best explains the data, we adopt a likelihood maximization approach. That is, we seek to find the partition of nodes to communities that best describes the observed connectivity and attribute information. Given the conditional independence assumption of $\mathbf{X}$ and $\mathbf{A}$, we can express the log likelihood of the data, $\mathcal{L}$ as the sum of connectivity and attribute log likelihoods, $\mathcal{L}_A$ and $\mathcal{L}_X$, respectively, as

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_X \ . \tag{1}$$

This likelihood reflects the joint distribution of the adjacency matrix, $\mathbf{A}$, the attribute matrix, $\mathbf{X}$, and the matrix of node-to-community indicators, $\mathbf{Z}$; formally, we have

$$\mathcal{L} = p(\mathbf{A}, \mathbf{X}, \mathbf{Z}) \ . \tag{2}$$

Given that $\mathbf{Z}$ is a latent variable that we are trying to infer, we can approach the problem using the expectation maximization (EM) algorithm [6]. By doing this, we will alternate between estimating the posterior probability that a node i has community label c, or

$$p(z_{ic} = 1 \mid \mathbf{X}, \mathbf{A}) \tag{3}$$

and estimates for $\boldsymbol{\theta}, \boldsymbol{\Psi}$, i.e., the model parameters specifying the adjacency and attribute matrices, respectively.

2.2 Attribute Likelihood

For a network with K communities, we assume that each particular community i has an associated p-dimensional mean $\boldsymbol{\mu}_i$ and $p \times p$ covariance matrix, $\boldsymbol{\Sigma}_i$. Note that these parameters uniquely identify a p-dimensional multivariate Gaussian distribution. To specify this model for all K communities, we define the parameter $\boldsymbol{\Psi} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots \boldsymbol{\Sigma}_K\}$. Then, the probability of the i-th row of $\mathbf{X}$, denoted as $\mathbf{x}_i$, giving the values of the p attributes for node i, can be modeled as

$$p(\mathbf{x}_i \mid \boldsymbol{\Psi}) = \sum_{c=1}^{K} \pi_c p(\mathbf{x}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \ . \tag{4}$$

Here, $p(\mathbf{x}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is the probability density function for the multivariate Gaussian and π_c is the probability that a node is assigned to community c.

2.3 Adjacency Matrix Likelihood

For the adjacency matrix, $\mathbf{A}$ and the $K \times K$ matrix of stochastic block model parameters, $\boldsymbol{\theta}$, the probability of observing the connectivity pattern of node i, or the i-th row in the adjacency matrix, denoted as $\mathbf{a}_i$, is given by

$$\begin{aligned} \log p(\mathbf{a}_i \mid \boldsymbol{Z}, \boldsymbol{\theta}) = \sum_{c_1=1}^{k} z_{i,c_1} [\sum_{c_2=1}^{k} \sum_{j \in \mathcal{N}(i)} z_{j,c_2} \log(\boldsymbol{\theta}_{c_1,c_2}) \\ - \sum_{j \notin \mathcal{N}(i)} z_{j,c_2} \log(1 - \boldsymbol{\theta}_{c_1,c_2})] \end{aligned} \tag{5}$$

2.4 Inference

To use EM to maximize the likelihood of the data, we break the process into the E-step and M-Step, and perform this step sequence iteratively until the estimates converge.

E-Step. During the E-step, we use the current value of learned model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$ to compute the posterior, given in Eq. (3). The posterior at each step, $\gamma(z_{ic})$, of node i belonging to community c, is given by

$$\begin{aligned} \gamma(z_{ic}) &= p(z_{ic} = 1 \mid \mathbf{x}_i, \mathbf{a}_i) \\ &= \frac{p(\mathbf{x}_i, \mathbf{a}_i, z_{ic})}{p(\mathbf{x_i}, \mathbf{a_i})} \\ &= \frac{p(\mathbf{x_i} \mid z_{ic}) p(\mathbf{a}_i \mid z_{ic}) \pi_c}{\sum_{c=1}^{K} p(\mathbf{x_i} \mid z_{ic}) p(\mathbf{a}_i \mid z_{ic}) \pi_c} \ . \end{aligned} \tag{6}$$

M-Step. In the M-step, we can compute updates for $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$ using this expectation.

Since, the attributes follow a Gaussian mixture model, it can be shown that the update for the mean vector describing community c, $\boldsymbol{\mu}_c$, can be computed as

$$\boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} \gamma(z_{ic}) \mathbf{x}_i}{\sum_{i=1}^{N} \gamma(z_{ic})} \ . \tag{7}$$

Similarly, the update for the covariance matrix describing a community, $\boldsymbol{\Sigma}_c$, is computed as

$$\boldsymbol{\Sigma}_c = \frac{\sum_{i=1}^{N} \gamma(z_{ic})(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T}{\sum_{i=1}^{N} \gamma(z_{ic})} \ . \tag{8}$$

To update the parameters of $\boldsymbol{\theta}$, we follow the method in [5] and update the probability of an edge existing between community q and l, given by $\boldsymbol{\theta}_{ql}$ as,

$$\boldsymbol{\theta}_{ql} = \frac{\sum_{i \neq j} \gamma(z_{iq}) \gamma(z_{jl}) x_{ij}}{\sum_{i \neq j} \gamma(z_{iq}) \gamma(z_{jl})} \tag{9}$$

We continue the process of iterating between the E-step and M-step until the change in the data log-likelihood, $\mathcal{L}$, is below a predefined tolerance threshold.

2.5 Initialization

Likelihood optimization approaches are often sensitive to initialization because it is easy to get stuck in a local optimum. As an initialization strategy for the nodes, we simply

cluster the nodes in the network using the Louvain algorithm (cite louvain). We chose this approach because this algorithm is efficient and stable.

3. SYNTHETIC DATA RESULTS

We first test the performance of our model and inference procedure on a synthetic example. Here, we generated a network for a stochastic block model with $N = 200$ nodes and $K = 4$ communities . Edges in the adjacency matrix were generated according to a stochastic block model, parameterized as follows:

$$p(A_{ij} = 1) \sim \begin{cases} \text{Bernoulli}(.10), & \text{if } z_i \neq z_j \\ \text{Bernoulli}(.25), & \text{if } z_i = z_j \end{cases} \quad (10)$$

Note that $\mathbf{z}$ is a 200-dimensional vector, where z_i identifies the community label for node i.

Fig. 2(a) shows an example network generated according to this parametrization. Nodes are colored by their community assignment. Looking at the network, it is apparent that there is not a strong notion of detectable community structure. That is, nodes are not separated into the desired homogeneous clumps that should be consistent with community assignment. To model attributes, for a community c, we randomly generated an 8-dimensional vector, $\boldsymbol{\mu}_c$, where each entry is from a Gaussian with 0-mean and unit variance. Associated with all c, $c \in \{1, 2, 3, 4\}$ is a 8×8 diagonal covariance, $\boldsymbol{\Sigma}_c = \text{diag}(1.25)$. Moreover, using the $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, a sample attribute vector can be generated. That is, the attribute vector for node i, $\mathbf{x}_i$ is generated as,

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \quad (11)$$

where $\mathcal{N}(\cdot, \cdot)$ denotes a multivariate Gaussian.

Fig. 2(b) shows a PCA plot of the attribute vectors associated with each node in an example synthetic experiment and hence, each point represents a node. Since, the true dimension of these feature vectors is 8, this plot provides a projection to 2 dimensions that allows for visualization of the relatedness between the nodes, according to the attributes. One can observe that members of community 2 are overall nicely separated in attribute space but members of communities 3 and 5 are especially hard to discern.

To assess how well the attribute SBM approach performed in successfully assigning nodes to communities, we compared the results obtained from our model to clustering results obtained either only clustering the nodes based on connectivity, and to results of clustering nodes based only on their attribute information. We quantify the correctness of the partition with normalized mutual information (NMI) (cite Danon). Letting $\mathbf{z}$ denote the true node-to-community assignments, then $\mathbf{z}^{\text{connectivity}}$, $\mathbf{z}^{\text{attributes}}$, and $\mathbf{z}^{\text{attribute sbm}}$ denote the partition of the nodes according to only the network connectivity only, attributes only, and with the attributes SBM. To cluster the network only according to connectivity, we fit a stochastic block model with 4 blocks. To cluster nodes with only attributes, we performed k-means clustering on only the attributes. Computing the NMI between $\mathbf{z}$ and each of these 3 cases, or $\text{NMI}(\mathbf{z}, \{\mathbf{z}^{\text{connectivity}}, \mathbf{z}^{\text{attributes}}, \mathbf{z}^{\text{attribute sbm}}\}) = \{0.65, 0.68, 0.83\}$. These results show that by combining both sources of information, there is an improvement in the ability to correctly identify communities. To further probe this idea, we sought to empirically look closer at the so-called 'detectability limit' (cite dane, cris moore). Generally, detectability refers to the difficulty of correctly identifying clusters in data. Multiple works (refs) have explored these limits in stochastic block models observing the sharp phase transition that occurs in accuracy as soon as within-community probability (p_{in}) is sufficiently larger than the between-community probability (p_{out}).

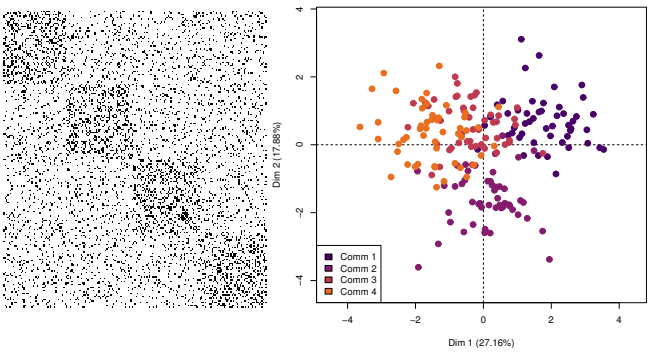

Figure 2: **Synthetic Example.** We generated a synthetic network with $N = 200$ nodes, $K = 4$ communities and an 8-dimensional multivariate gaussian for each community. **A.** A visualization of the adjacency matrix for this network where a black dot indicates an edge. We observe that there is an assortative block structure (blocks on the diagonal), but there are also many 'noisy' edges between communities making the true community structure with only a stochastic block model a bit harder to detect. **B.** We performed PCA on the $N \times p$ dimensional attribute array and plotted each of the N nodes in two dimensions. Points are colored by their true community assignments, $\mathbf{z}$. Clustering the nodes according to only connectivity, only attributes, and with the attributed SBM, we quantified the partition accuracy with normalized mutual information. This gave results $\text{NMI}(\mathbf{z}, \{\mathbf{z}^{\text{connectivity}}, \mathbf{z}^{\text{attributes}}, \mathbf{z}^{\text{attribute sbm}}\}) = \{0.65, 0.68, 0.83\}$.

Based on the results of the synthetic experiments in figure 2 where the attributes combined with connectivity lead to a more accurate partitioning of the nodes, we hypothesized that augmenting the network connectivity with attributes could somehow move this detectability limit. In figure 3, we explored how generating networks from a stochastic block model with varying ratios between p_{in} and p_{out} combined with the attributes used in figure 2 would affect the accuracy of the node-to-community partition. To do this, we considered values of p_{in} between 0.05 and 0.3 in increments of 0.05. For each of these p_{in} values, we found the corresponding value of p_{out} such that the mean degree was 20. For each of these p_{in} and p_{out} combinations, we generated 10 different networks using a stochastic block model. In figure 3 we plot the NMI between the true partition, $\mathbf{z}$ and the partitions using only the connectivity with the regular SBM $\mathbf{z}^{\text{connectivity}}$ and the attributed SBM $\mathbf{z}^{\text{attribute sbm}}$. These results are plotted in blue and pink, respectively. The shaded region around the points indicates standard deviation.

We see that while both inference approaches undergo a

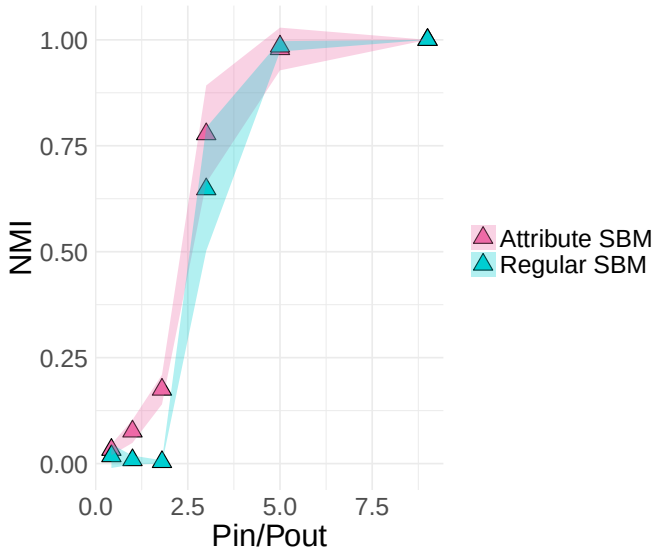

Figure 3: **Detectability Analysis in Synthetic Example.** To understand how attribute information can be combined with connectivity to assign nodes to communities accurately, we generated synthetic networks for within-probabilities of p_{in} between 0.05 and 0.3 with corresponding p_{out} or between-community probabilities such that the mean degree of the network was 20. For each of these synthetic networks, we used the attributes from the analysis in figure 2 to fit the attributed SBM. Here, we plot the correctness of the node-to-community assignment with normalized mutual information using the partition obtained from regular SBM (blue) and the partition under the attributed SBM model fit (pink). For each combination of p_{in} and p_{out}, we generated 10 networks and hence the bands around the points denote standard deviation. Incorporating attributes with the attributes stochastic block model shifts the detectability limit slightly to the left.

phase transition at a similar ratio of $p_{in}/p_{out} = 3$, we notice that the curve for the attribute SBM model is slightly shifted to the left suggesting that the extra attribute information does positively impact the ability to correctly identify communities. Future work could focus on understanding detectability questions in relation to the parameters for the underlying multivariate gaussian distributions parametrizing each community. For example, how does the difference in means between a pair of communities shift the detectability curve?

4. USING THE FITTED ATTRIBUTED SBM FOR LINK PREDICTION AND COLLABORATIVE FILTERING

One of the benefits of a generative network model is that it can be applied to prediction tasks. Most notably, in the absence of one source of information about a node (connectivity or attributes), the model can be used to predict the complementary information source (attributes or connectivity, respectively). By fitting an attributed SBM, we found that obtain successful performance in two fundamental network prediction tasks, link prediction and collaborative filtering.

In the link prediction problem, when given two node stubs, the objective is to determine whether a link exists between them. Since we are modeling connectivity with a stochastic block model, we can predict links using the learned parameters. In particular, we highlight how this task can be performed using just the attribute information of the node stubs of interest. In the experiments to follow, we compare to 3 commonly-used link prediction methods. In all of these methods, a score is computed for all pairs of edge candidates and ultimately the top x set of edges with highest weights are kept (where x is some user-defined parameter). Let m and n be a pair of nodes and $\Gamma(m)$ denote the set of neighbors for a node m. Then, under the following 3 link prediction methods, we can calculate the score of the potential link as, Score(m, n)

Jaccard: $\text{Score}(m,n) = \frac{\Gamma(m)\cap\Gamma(n)}{\Gamma(m)\cup\Gamma(n)}$

Adamic Adar: $\text{Score}(m,n) = \sum_{c\in(m)\cap\Gamma(n)} \frac{1}{\log|\Gamma(c)|}$

Preferential Attachment: $\text{Score}(m, n = |\Gamma(m)|\times|\Gamma(n)|$

Conversely, the collaborative filtering problem seeks to predict a node's attributes based on its similarity to its neighbors. For some node of interest, we can use our fitted attributed SBM model to predict a node's attributes, given only the information about its connectivity. Formally, for node i, we seek to predict $\mathbf{x}_i$. In the following experiments, we compare our results to two common collaborative filtering approaches. Let $\mathcal{N}^k(m)$ be the set of k-nearest neighbors in the network for node m. Let $\hat{\mathbf{x}}_i$ be the predicted attribute vector for node i and s_{ij} be a similarity measure between nodes i and j.

Neighborhood Avg: $\hat{\mathbf{x}}_i = \frac{1}{|\mathcal{N}^k(i)|}\sum_{j\in\mathcal{N}^k(i)}\mathbf{x}_j$

Weighted Neighborhood Avg: $\hat{\mathbf{x}}_i = \frac{1}{\sum_{jin\mathcal{N}^k(i)} s_{ij}}\sum_{j\in\mathcal{N}^k(i)} s_{ij}\mathbf{x}_j$

We show results for these two tasks in two different biological network examples in section X. In particular, the experiments were designed in the following ways.

4.1 Link Prediction Experiments

For our link prediction (figures x and y), we performed a link prediction task by sampling pairs of nodes and utilizing the complementary source of attribute information. We sampled 10 different sets of 50 pairs of nodes. In each sample, 25 of the node pairs were those having an edge in the original network and 25 were pairs with no edge. For each of the 50 edges in each sample, we sought to predict whether an edge existed between the corresponding node pair in a leave one out manner. To do this, for each edge we fit the attributed SBM to the network with the pair of nodes (stubs) associated with the edge removed. We then use the nearest neighbor in attribute space of each stub as the the input to each of the 3 baseline community detection methods (Jaccard, Adamic Adar, and Preferential Attatchment). To use our attributed SBM in this link prediction task, we also consider the most commonly observed community among the 3 nearest neighbors for the stubs of the edge of interest. Again, using the nearest neighbors, which we denote by n and m of the stubs, then we define the link prediction score for the edge as θ_{z_n,z_m}, or the probability that an edge exists between nodes n and m according to the fitted model. After generating 10 samples of 50 edge pairs, this results in 500 total edge scores. Since we know the ground truth of

whether or not these edges actually exist from the original network, we can construct an ROC curve for each method. From these curves we can plot area under the curve (AUC) to quantify the quality of the link prediction result. Two link prediction results are reported in figures 5 and 9 for the microbiome and protein network examples.

4.2 Collaborative Filtering Experiments

In collaborative filtering experiments, the objective is to predict the vector of attributes for each node. In our experiments, we used leave-one-out validation to predict the attribute vector for each node. That is, for each node in the network, we created a single node test set. The training set, was then the rest of the network with the node to predict removed. For this single test set node, we identified neighbors it connects to in only connectivity space within the training set. For standard collaborative filtering approaches (Neighborhood average and weighted neighborhood average), the predicted attribute for the test set node is then the specified averaging of the neighbors. To use our model for this task, we first fit the attributed SBM model to the training set. Similar to the standard link prediction approaches, we identify the nearest neighbors for our test node in connectivity space within the training set. We then predict the community membership of our test node to be the most-frequently observed community among its neighbors. Using this community assignment, c, we then predict the attribute vector for our test node to be μ_c, or the mean vector that was learned to describe community c. The results of collaborative filtering experiments for the microbiome and protein network examples are shown in figures 6 and 10. For node a node i and its associated vector of attributes, $\mathbf{x}_i$ we quantify the accuracy of the predicted attribute vector, $\hat{\mathbf{x}}_i$ with a a relative error measure, $\mathcal{E}$, such that

$$\mathcal{E} = \frac{||\hat{\mathbf{x}}_i - \mathbf{x}_i||_2}{||\mathbf{x}_i||_2}. \tag{12}$$

5. APPLICATIONS IN BIOLOGICAL NETWORKS

While the motivation for the development of this method was not motivated by problems in biological data, we evaluate the potential to combine similarity or relational information between a set of entities, whether that be proteins, or biological samples, and experimental data and metadata. Our application of this model to biological problems provides a framework to predict attribute or connectivity information about a new observation. Note that we do not intend to suggest any new biological insights, but rather that we can combine two sources of information for prediction tasks and alternative definitions of what constitutes a community in the data. Applying the attributed stochastic block model to integrate connectivity and attribute data provides a way to find a partition that takes into account two different sources of information, or a method to predict one source of information (connectivity, attributes) in the absence of the other (attributes, connectivity).

5.1 Microbiome Subject Similarity Results

Motivation

In the analysis of biological data, it is often useful to cluster subjects based on a set of their measure biological features and to then determine what makes each of the subgroups different. One type of biological data gaining much attention in recent years is metagenomic sequencing data, used to profile the composition of a microbiome. We refer to this as the 'metagenomic profile' and each feature is a count for each bacterial species, also known as operational taxonomic unit (OTU). Lahti et al. conducted a study among subjects across a variety of ethnicities, body mass (BMI) classifications, and age groups to understand differences in the intestinal microbiota [11]. Using metagenomic sequencing, the counts for 130 OTUs were provided for each subject. We created an experiment to test our model by seeing if we could overlay a similarity network between subjects with the individual OTU count vectors for each subject.

Pre-Processing The data were downloaded from http://datadryad.org/resource/doi:10.5061/dryad.pk75d.We extracted a subset of the subjects from Eastern Europe,Southern Europe,Scandinavia, and the United States. Using only these subjects, a between-subject similarity network was constructed between the 121 individuals who had a BMI measurement. This resulted in a network of 121 nodes, where each edge is the pearson correlation between their microbial compositions. We then removed all edges in the network with weight (correlation) <0.7. Note that our attributed SBM does not allow for edge weights, so as input of this network into the model, we simply ignored the edge weights.

Constructing Node Attributes Since each node had a 130-dimensional vector of attributes (counts), we used this information to create a lower-dimensional attribute vector for each node by performing PCA and then representing each node with the first 5 principal components. Each dimension of this new attribute vector was then centered and scaled and has an approximately gaussian distribution.

We first visualized the differences in partitions obtained according to the classic and attributed stochastic block models in figure 4A-B, respectively. In both networks, nodes are colored by their community assignment. Using the classic stochastic block model and the model selection criteron described in [5], 7 blocks were identified. With the attributed stochastic block model, 6 blocks were identified. While we do not have ground truth labels on the nodes, it is visually apparent that adding the attributes to the inference problem helps to 'clean up' the partition. For example, in figure 4A there is mixing between the dark and lighter purple communities in the upper left of the network. In figure 4B, this mixing was reduced by assigning all of the nodes in the general region to the lighter purple community.

Microbiome Link Prediction In figure 4a-b we show the results for the link prediction in the microbiome sample similarity network. These experiments were performed in the manner described in section 4.1. To perform link prediction, we selected 25 pairs of nodes (i.e. edge stubs) from our 121-node subject network. In our prediction task , we sought to model the ability to use only the attribute information and our fitted model to predict whether there was a link. Repeating this experiment 25 times, we obtained a distribution of area under the curve (AUC) values in figure 4A. and plotted the ROC curves corresponding to the experiment closest to the median for all link prediction methods.

Microbiome Collaborative Filtering We performed the collaborative filtering experiments on the microbiome subject similarity network in the manner described in sec-

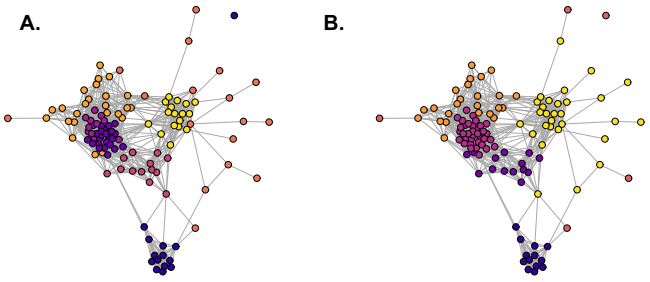

Figure 4: **Microbiome subject similarity network:** A visualization of the 121 node microbiome subject similarity network with nodes colored by the partition using the classic (**A.**) and attributed (**B.**) stochastic block model. **A.** Fitting the classic stochastic block model to the network, 7 communities were identified. **B.** Fitting the attributed stochastic block model to the network with the attributes being the first 5 principle components of each subject's OTU count vector (metagenomic profile), 6 communities were identified. Incorporating attributes in inferring this partition removed some of the noise in the partition on the network, specifically in the mixed purple community in the left of **A.**

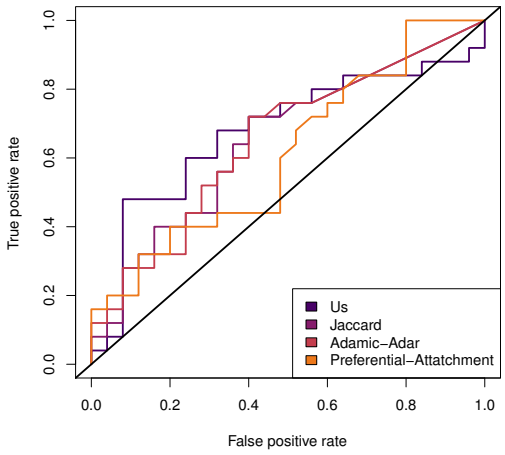

Figure 5: **Link Prediction on the microbiome subject similarity matrix:**

tion 4.2 to predict the 5-dimensional attribute vector for each node. The box plots indicate the distribution of relative errors over the 121 nodes for the attribute SBM (blue), neighbor average (pink) and weighted neighbor average (orange). While these distributions look similar, While the attributed SBM plotted has a similar distribution of relative errors with the standard collaborative filtering methods, the mean is slightly lower, at 0.21, compared to 0.26 and 0.27 in the neighbor average and weighted neighbor average, respectively.

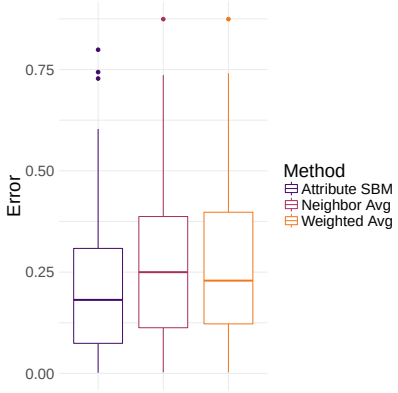

Figure 6: **Collaborative Filtering Accuracy in Microbiome Subject Similarity Network**: For each of the 121 nodes, we fit a model to the remaining 120 node network and given the node's closest neighbors (based on network connectivity) sought to predict its 5-dimensional attribute vector. The reported error is the relative error $\mathcal{E}$ between the difference between the true attribute vector ($\mathbf{x}_i$) and its predicted attribute vector ($\hat{\mathbf{x}}_i$). The mean error in $\mathbf{x}_i$ is 0.21, as opposed to the neighbor average and weighted neighbor averages, having errors of 0.26 and 0.27, respectively.

5.2 Protein Interaction Network Results

We also apply our attributed SBM approach to the protein interaction network presented in [3]. This network represents interactions between proteins, predicted from the literature. Associated with each each node (protein), is a classification of one of 6 experimental modifications observed from the exposure of cancer cells to a chemotherapeutic drug. While communities in this network should reflect functional relatedness among proteins (i.e. similar biological functions, in general), we also expect that members of a community should share similarities in the observed modification type. Also associated with each of the 6 modification types is whether that particular type of modification became either more or less prominent after treatment with the drug. Since we have two types of labels associated with these nodes, we also sought to explore how these two labeling schemes (6 class vs. 2 class) aligned with the communities returned by the algorithm.

Data Pre-Processing: We downloaded the unweighted protein interaction network data and the modification information from the supplement of [3]. We removed 13 nodes that were not connected to the giant component of the network and considered only the 82 node giant connected component.

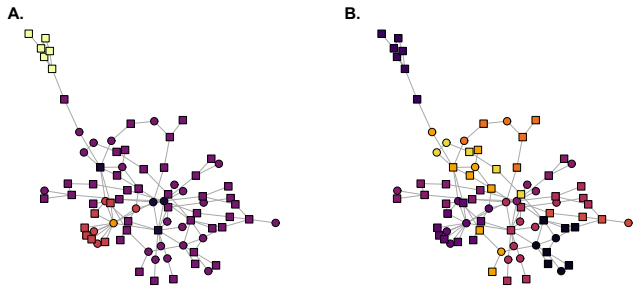

Figure 7: **Protein interaction network.** We visualize the 82 node protein interaction network under the classic stochastic block model **A.** and the attributed stochastic block model **B.** In both networks, nodes are colored by their community assignment and the node shape indicates whether the modification status increased (square) or decreased. **A.** Nodes colored according to the community partition under the stochastic block model. Nodes are assigned to one of five communities. **B.** Nodes are colored to the community partition under one of nine communities.

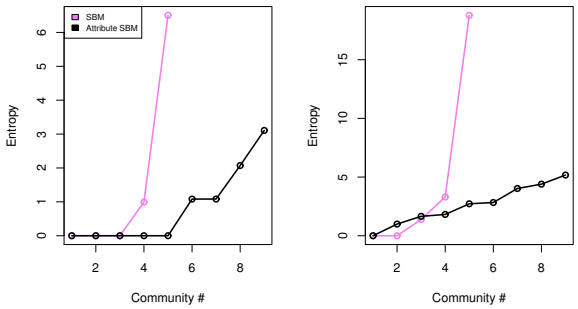

Figure 8: **Community entropies in the protein interaction network.** We studied the entropy of the 2 class and 6 class classifications of the nodes in **A.** and **B.**, respectively under the classic SBM (black) and attributed SBM (purple) partitions. For **A** − **B** the horizontal axis denotes the community index for the particular partition. Nodes belonged to 1 of 5 communities under the classic SBM and belong to 1 of 9 communities with the attributed SBM. Incorporating attributes under both classifications succeeds in breaking up a high entropy community (5) from the classic SBM partition to lower entropy communities in the attributed SBM partition.

Constructing Node Attributes: As previously described, each node is classified with 1 of 6 possible modification type. For each node, we created an attribute vector that captured the modification types of its neighbors. To do this, we considered the 4th order neighborhood of each node. That is, for each node, we collected its neighbors who were four hops or less away in the network. Then to define the value for attribute c of node i, or x_{ic}, we counted the number of fourth order neighbors of node i with label c. After defining these attributes across all nodes, for each of the 6 classes, we centered and scaled each attribute across all of the nodes.

Figure 7A-B show the results of fitting a classic SBM and attributed SBM, respectively. The 6 possible modifications exist for 3 biological processes that can either increase or decrease. The node shape reflects whether the experimental modification for a node increased (square) or decreased (circle) after treatment with the chemotherapeutic agent. Again by fitting an SBM with the model selection criterion in [5], 5 communities were identified. With our attributed SBM, 9 communities were identified.

Using the partition of the nodes under the classic and attributed stochastic block models, we sought to use the two different classifications of the nodes (6 class modification type and 2 class increase/decrease) to compute entropy of labels within communities. The expectation is that by incorporating attribute information that is related to the functional protein information into the community detection problem, we should see a decrease in the entropy over the classification labels in communities. In figure 8A-B, we plot the entropy for the 2 class and 6 class node classifications, respectively. We define $\mathbf{E}_c$, the entropy for community c as,

$$\mathbf{E}_c = -\sum_k p_k \log(p_k). \tag{13}$$

.

Here k is indexing the unique classifications found in community c and p_k is the probability that a node in community c belonged to classification k. In these plots the black and purple curves correspond to the fits of the classic and attribute SBM fits, respectively. Using both types of node classifications to compute these entropy quantities, we see that the attribute SBM succeeds in breaking up one high entropy community (5) from the classic SBM partition into lower entropy communities.

Link Prediction in the Protein interaction network We performed link prediction on the protein interaction network using the procedure described in section 4.1. Given that this protein network is sparse, none of the link prediction methods performed particularly well. The AUC values for the attributed SBM, jaccard, adamic adar and preferential attachment were 0.61, 0.58, 0.58, and 0.54, respectively. The associated ROC curves are shown in figure 9.

Collaborative filtering in the protein interaction network Collaborative filtering was performed using the method described in section 4.2. Note that unlike the microbiome sample similarity network, the edges in this network are unweighted and hence the neighbor average and weighted neighbor average methods produce the same result. We note that performing collaborative filtering with the attributed stochastic block model results in a lower mean error of 0.21 compared to that of 0.48 when using the neighbor average. Similar to figure 5, the box plots represent the distribution of errors across each of the 82 nodes.

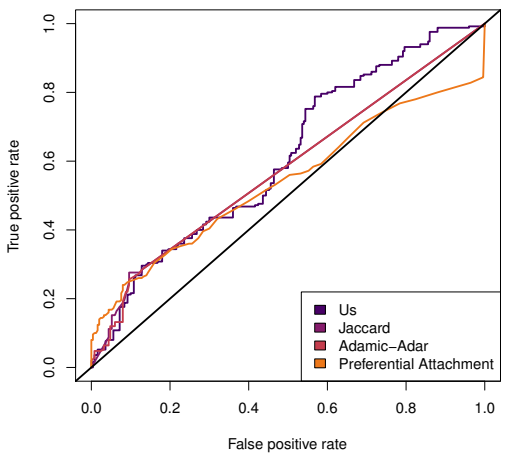

Figure 9: **Link Prediction in the protein interaction network**. Performing link prediction using the attributed SBM, jaccard, agamic adar, and preferential attachment. The corresponding AUC curves for these methods were 0.61, 0.58, 0.58, and 0.51, respectively.

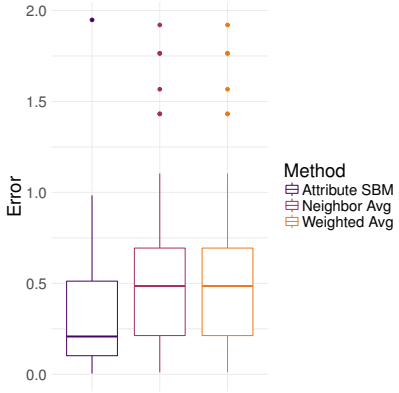

Figure 10: **Collaborative filtering in the protein interaction network**. For each of the 82 nodes, we fit a model to the remaining 81 node network and given the node's closest neighbors (based on network connectivity) sought to predict its 6-dimensional attribute vector. The reported error is the relative error $\mathcal{E}$ between the difference between the true attribute vector ($\mathbf{x}_i$) and its predicted attribute vector ($\hat{\mathbf{x}}_i$). The mean error in $\mathbf{x}_i$ using the attributed SBM is 0.21, as opposed to the neighbor average error where it is 0.48.

6. DISCUSSION

Detectability problems

7. DISCUSSION

To be filled in.

8. REFERENCES

[1] N. Aghaeepour, E. A. Ganio, D. Mcilwain, A. S. Tsai, M. Tingle, S. Van Gassen, D. K. Gaudilliere, Q. Baca, L. McNeil, R. Okada, et al. An immune clock of human pregnancy. *Science immunology*, 2(15):eaan2946, 2017.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[3] T. Bonacci, S. Audebert, L. Camoin, E. Baudelet, G. Bidaut, M. Garcia, I.-I. Witzel, N. D. Perkins, J.-P. Borg, J.-L. Iovanna, et al. Identification of new mechanisms of cellular response to chemotherapy by tracking changes in post-translational modifications by ubiquitin and ubiquitin-like proteins. *Journal of proteome research*, 13(5):2478–2494, 2014.

[4] D. Combe, C. Largeron, M. Géry, and E. Egyed-Zsigmond. I-louvain: An attributed graph clustering method. In *Advances in Intelligent Data Analysis XIV*, pages 181–192. Springer, 2015.

[5] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

[6] 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)*, pages 1–38, 1977.

[7] 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*, pages 118–121. ACM, 2013.

[8] J. Guinney, R. Dienstmann, X. Wang, A. de Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 2015.

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

[10] D. Hric, T. P. Peixoto, and S. Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, 6(3):031038, 2016.

[11] L. Lahti, J. Salojärvi, A. Salonen, M. Scheffer, and W. M. De Vos. Tipping elements in the human intestinal ecosystem. *Nature communications*, 5, 2014.

[12] D. B. Larremore, A. Clauset, and C. O. Buckee. A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Comput Biol*, 9(10):e1003268, 2013.

[13] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

[14] M. Newman and A. Clauset. Structure and inference in annotated networks. *arXiv preprint arXiv:1507.04001*, 2015.

[15] L. Peel, D. B. Larremore, and A. Clauset. The ground truth about metadata and community detection in

networks. *Science Advances*, 3(5):e1602548, 2017.

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

[17] S. Shai, N. Stanley, C. Granell, D. Taylor, and P. J. Mucha. Case studies in network community detection. *arXiv preprint arXiv:1705.02305*, 2017.

[18] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[19] J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *Data mining (ICDM), 2013 ieee 13th international conference on*, pages 1151–1156. IEEE, 2013.