Multiscale Community Blockmodel for Network Exploration

Qirong Ho
Carnegie Mellon University

Ankur P. Parikh
Carnegie Mellon University

Le Song
Carnegie Mellon University

Eric P. Xing
Carnegie Mellon University

Abstract

Real world networks exhibit a complex set of phenomena such as underlying hierarchical organization, multiscale interaction, and varying topologies of communities. Most existing methods do not adequately capture the intrinsic interplay among such phenomena. We propose a nonparametric Multiscale Community Blockmodel (MSCB) to model the generation of hierarchies in social communities, selective membership of actors to subsets of these communities, and the resultant networks due to within- and cross- community interactions. By using the nested Chinese Restaurant Process, our model automatically infers the hierarchy structure from the data. We develop a collapsed Gibbs sampling algorithm for posterior inference, conduct extensive validation using synthetic networks, and demonstrate the utility of our model in real-world datasets such as predator-prey networks and citation networks.

1 INTRODUCTION

How do complex networks and their self-organization arise from coordinated interactions and information sharing among the actors? One way to tap into this question is to understand the latent structures over actors which lead to the formation and organization of these networks. In particular, we are interested in uncovering the functional/sociological communities of network actors, and their influence on network connections. We consider a community to be a group of actors that share a common theme, like a clique of football fans in a social network, or an ecosystem of dependent organisms in a biological food web. Our objective is to gain a deeper understanding of the relationships within and among these communities, so as to shed insight into the network topology.

More specifically, we seek to address three critical aspects of network modeling and community discovery:

1. **Hierarchy** — not all communities are equal: a community can contain sub-communities, or be contained by super-communities. This is a natural way to structure the latent space of actors.

2. **Multiscale Granularity** — we must distinguish between coarse or generic associations that may occur in a large super-community, as opposed to fine grained interactions that occur within or among small, closely-interconnected sub-communities.

3. **Assortativity/Disassortativity** — some communities have strong within-community interactions and weak cross-community interactions (*assortativity*), yet others may exhibit the reverse (*disassortativity*).

These aspects are not independent, but are strongly interrelated. As an example, consider an oceanic food web (Figure 1), a directed network with species as actors and predator-prey relationships as edges. This network exhibits *hierarchy*: *cold-blooded animals* and *mammals* are large super-communities that can be sub-divided into smaller sub-communities, such as *sharks* and *squid*, or *toothed whales* and *pinnipeds*. These sub-communities can in turn be divided into even smaller communities (not shown). The ideas of hierarchy and network should *not* be confused with each other. The hierarchy is an organization of actors in some latent space learned from the observed network.

Next, the predator-prey relationships in the ocean are *multiscale*. Consider a sperm whale: it occasionally eats fish, which are common prey for many oceanic animals. Hence, this "sperm whale, fish" interaction is *generic*. Moreover, sperm whales usually eat giant squid, which are prey specific to them (making this interaction *fine-grained*). It is important to differentiate between such interactions of different scale.

Finally, the toothed whale sub-community demonstrates both *assortative* and *disassortative* behavior. Many toothed whales feed on small fish and seals, which are cross-community interactions. However, whales such as orcas feed on other whales, which are within-community interactions.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

We propose a nonparametric Multiscale Community Blockmodel (MSCB) that presents a unified approach to address these three concerns. Using the nested Chinese Restaurant Process (Blei, Griffiths, and Jordan 2010) as a nonparametric structural prior, our model learns the structure of the hierarchy from the data without requiring the branching factor at each node to be prespecified. Moreover, by exploiting latent space ideas from Blei *et al.* (2003) and Airoldi *et al.* (2008), we uncover the coarse/fine-grained interactions that underlie the network. Finally, our model builds upon the blockmodel concept (Wang and Wong 1987; Airoldi, Blei, Fienberg, and Xing 2008) to integrate assortativity and disassortativity into our hierarchy.

1.1 Comparison to Existing Work

Existing methods for graph clustering and inferring community structure do not adequately capture the three aspects we have described. Methods such as Girvan and Newman (2002), Hoff *et al.* (2002), Handcock *et al.* (2007), Krause *et al.* (2003) and Guimera and Amaral (2005) cannot discover *disassortative* communities characterized by weak within-community and strong cross-community interactions. Furthermore, they do not explicitly model organizational structure — and by extension, multiscale granularity of interactions. These methods do not meet any of our criteria, and are unsuited for our purposes.

The Mixed Membership Stochastic Blockmodel (MMSB) (Airoldi, Blei, Fienberg, and Xing 2008) aims to discover the multiple latent roles played by each actor in the network, while employing a blockmodel to accommodate both disassortative and assortative types of interactions. The multi-role memberships discovered by MMSB are similar, but not identical, to our notion of coarse/fine-grained interactions. Furthermore, MMSB does not induce a hierarchical structure over the actors. These considerations prevent MMSB from modeling the organized network phenomena that our model is designed to explore. Another example of a latent space model is Miller *et al.*'s link prediction model (2009), which allows each actor to take on multiple binary features in an infinite-dimensional space. Like MMSB, this model does not learn a structure over its latent space, and therefore cannot replicate our model's ability to discover community hierarchies.

On the other hand, methods such as Clauset *et al.* (2004), Radicchi *et al.* (2004) and the infinite stochastic blockmodel (Kemp and Tenenbaum 2008) explicitly model some form of organizational structure. However, they do not permit actors to have multiple kinds of interactions, which precludes them from learning the kind of multiscale interactions we have described. Roy *et al.* (2007) generalize the infinite relational model (Kemp, Tenenbaum, Griffiths, Yamada, and Ueda 2006) for hierarchical group discovery, and extend their work to the nonparametric setting with Mondrian Processes (Teh and Roy 2009). However, their models are limited to *binary* hierarchies. Our model assumes no limit on the hierarchy's branching factor, which is more realistic for certain networks.

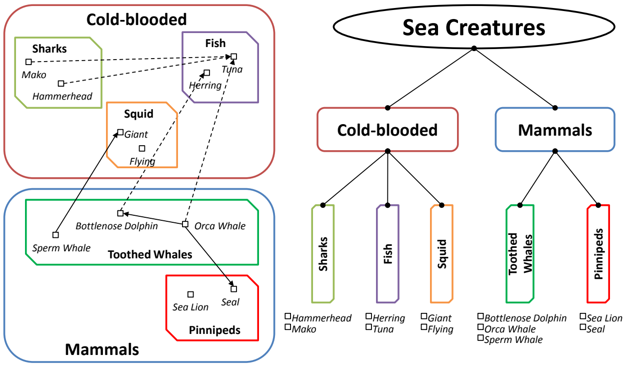

Figure 1: Illustration of an oceanic food web as a set of nested communities **(Left)**, and the corresponding hierarchy of communities **(Right)**. Vertices in the network represent individual species, and directed edges represent predator-prey relationships (not all shown). Solid edges are *fine-grained*, specific interactions, while dashed edges are *coarse-grained* and generic to a large community.

2 MULTISCALE COMMUNITY BLOCKMODEL (MSCB)

In the sequel, we describe the different aspects of the model, beginning with the hierarchy and then proceeding to network edge generation. We use the oceanic food web in Figure 1 as a running example.

2.1 The Community Hierarchy

In our model, the hierarchy is a tree where each node is a community. The root of the tree is designated as level 0. Nodes closer to the root represent large super-communities, (e.g. the "cold-blooded animals" and "mammals" in Figure 1), while those closer to the leaves represent finer-grained sub-communities (e.g. "toothed whales" or "sharks").

Each actor is associated with a single path of super-sub communities in the hierarchy (which we call its path c_i). This path delineates a sequence of communities from coarse to fine. For example, a *sperm whale* could have the path [mammal, toothed whale].

Selecting the number of branches at every tree node *a priori* can be daunting because of the huge number of possibilities. One might consider using heuristic methods that guess at the number of such children, but doing so would defeat the purpose of employing a probabilistic model.

We solve this problem by adopting a nonparametric Bayesian prior on paths through trees, the nested Chinese Restaurant Process (nCRP) (Blei, Griffiths, and Jordan 2010), which automatically selects the number of branches based on the data. The generative process for the nCRP works according to the following metaphor: Actor 1 chooses his tree path first, followed by actor 2, and so on. Consider actor i. He begins at the root, then with probability $n^{(1)}_{x,i-1}/(i-1+\gamma)$ he selects branch x of the tree, and with probability $\gamma/(i-1+\gamma)$, he picks a new branch. $n^{(1)}_{x,i-1}$ is the number of actors before i that chose branch x at level 1, and γ is a hyperparameter dictating the probability that an actor will start a new branch. Higher values of γ will increase the width of the hierarchy.

Actor i continues this process as he descends the tree. When picking a branch at level k, with probability $n^{(k)}_{y,i-1}/(n^{(k-1)}_{i-1}+\gamma)$ he selects branch y, and with probability $\gamma/(n^{(k-1)}_{i-1}+\gamma)$ he starts a new branch. Here, $n^{(k-1)}_{i-1}$ counts the number of actors $1,\ldots,i-1$ having the same path as actor i up to (and including) level $k-1$. Out of these actors, $n^{(k)}_{y,i-1}$ is the number that picked branch y (at level k). This sequence of branch choices defines the path of actor i, and the union of all these paths forms the hierarchy. A more formal treatment of the nCRP can be found in the Supplemental. In our model, we limit the hierarchy to a maximum depth of K.

2.2 Multiscale Membership

In order to enable multiscale granularity on the interactions, we associate each actor i with a Multiscale Membership (MM) vector θ_i. The MM vector is a K-dimensional multinomial that encodes an actor's tendencies to interact as a member of the different super- and sub- communities along his/her path of depth K. Consider two toothed whales: dolphins and sperm whales. Both have the same path in the tree, [mammal, toothed whale], yet both behave very differently. A dolphin's diet mainly consists of fish, which are common prey for many mammals. Thus it has a high probability of interacting as a member of the mammal super-community, although it occasionally chooses prey that are more specific to its species.

A sperm whale on the other hand barely eats fish, and thus rarely interacts as a member of its super-community. Instead, it eats giant squid, a more specific prey uncommon to most mammals. As a result, a sperm whale has a higher probability of participating in fine-grained interactions, instead of coarse ones like the dolphin does.

Like the mixed membership vector of the MMSB (Airoldi, Blei, Fienberg, and Xing 2008), which allows an actor to have a distribution over roles, our Multiscale Membership vector allows an actor to have a distribution over communities. However, there is a key difference: in MMSB, an actor may have a distribution over all possible latent roles, whereas in our model, an actor's Multiscale Membership vector is a distribution over only the set of super and sub-communities along the actor's path. This is because allowing an actor to have a distribution over all communities in the hierarchy can render the hierarchy virtually meaningless: a dolphin could simultaneously be in the shark and toothed whale communities, which is unrealistic.

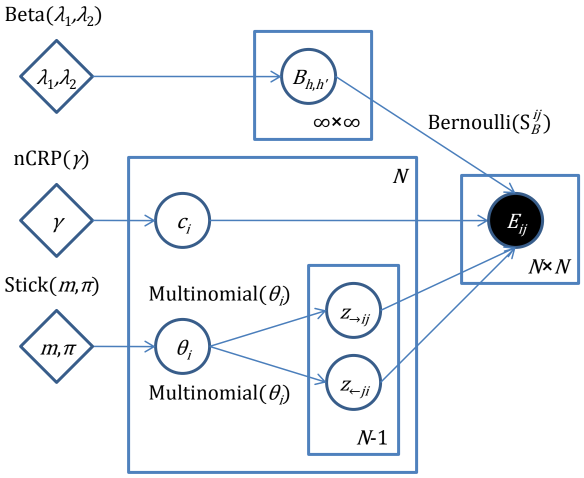

Figure 2: Graphical model for MSCB.

The Multiscale Membership vectors θ_i are drawn from a two-parameter stick breaking process $\mathrm{Stick}(m,\pi)$ (Blei, Griffiths, and Jordan 2010). The stick breaking prior makes it more intuitive to bias interactions toward coarser or finer levels compared to a Dirichlet prior with either a single parameter (which is not expressive enough), or $K-1$ parameters (which may be too expressive). The parameter $m>0$ influences the mean of θ_i, and $\pi>0$ influences its variance (details in the Supplemental). Because the hierarchy is only learnt up to depth K, we truncate the $\mathrm{Stick}(m,\pi)$.

2.3 Network Edge Generation

At this point, we shall introduce some notation. Let E be the $N\times N$ adjacency matrix of observed network edges, where E_{ij} corresponds to the *directed edge* or interaction/relationship from actor i to j. In the context of our food web, the actors are sea creatures like dolphins and sperm whales, and the edges represent predator-prey interactions. A value of $E_{ij}=1$ indicates that the interaction is present, while $E_{ij}=0$ indicates absence, and we ignore self-edges E_{ii}.

We introduce our *generative process* for network edges:

- For each actor $i\in\{1,\ldots,N\}$
    - Sample i's path $c_i\sim\mathrm{nCRP}(\gamma)$.
    - Sample i's MM $\theta_i\sim\mathrm{Stick}(m,\pi)$.
- To generate the network, for each directed edge E_{ij}:
    - Sample donor level $z_{\rightarrow ij}\sim\mathrm{Multinomial}(\theta_i)$.
    - Let[1] $h=c_i[z_{\rightarrow ij}]$.
    - Sample receiver level $z_{\leftarrow ij}\sim\mathrm{Multinomial}(\theta_j)$.
    - Let $h'=c_j[z_{\leftarrow ij}]$.
    - Sample the edge $E_{ij}\sim\mathrm{Bernoulli}(\mathrm{S}_B(h,h'))$. We shall explain the meaning of S_B later.

The basic idea is as follows: for every directed edge E_{ij}, both actor i (the donor) and actor j (the receiver) pick

[1] Formally, h is the community at level $z_{\rightarrow ij}$ on path c_i.

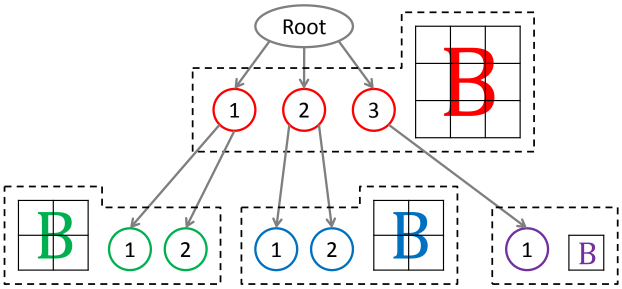

Figure 3: Sibling groups in the hierarchy, and their associated community compatibility matrices $\mathbf{B}$.

communities h and h' from their respective paths c_i, c_j, according to their MM vectors θ_i, θ_j. The communities h, h' are then used to select a *community compatibility* parameter $\mathrm{S}_B(h, h')$, which in turn generates $E_{ij} \sim \mathrm{Bernoulli}(\mathrm{S}_B(h, h'))$. Note that the arrow in $z_{\rightarrow ij}$ or $z_{\leftarrow ij}$ denotes donor or receiver respectively, *not* edge direction between i and j.

2.4 Community Compatibility Matrices B

We now discuss the $\mathrm{S}_B()$ function. Intuitively, the *compatibility* from h to h' is high if actors from h often interact with actors from h'. Conversely, a low compatibility indicates that actors from h rarely interact with actors from h'. Thus, it is natural to define compatibility to be a Bernoulli parameter in $[0, 1]$, where 1 indicates perfect compatibility. This notion of compatibility is what allows our model to account for both assortative and disassortative behavior. (For example, strong assortative interactions correspond to high compatibility parameters when $h = h'$).

There are many ways to associate compatibility parameters with pairs of communities h, h'. Our goal is to meaningfully integrate compatibility with the hierarchy and multiscale interactions over communities. A first attempt might be to ignore the hierarchy, and place a full $H \times H$ compatibility matrix over all community pairs h, h', which is analogous to the blockmatrix of MMSB (Airoldi, Blei, Fienberg, and Xing 2008). However, this formulation does not capture the multiscale nature of interactions, because there is no connection between the compatibility parameter for h, h' and those communities' levels in the hierarchy.

Instead, we *restrict* the compatibility parameters by defining a compatibility matrix for each *sibling group* (a set of children under the same parent) in the hierarchy. Each sibling group's compatibility matrix defines the interaction probability between every pair of siblings in that group — refer to Figure 3 for an illustration. Since the number of hierarchy nodes is not pre-specified, the number and size of the sibling group compatibility matrices must be determined automatically from the data. We shall address this issue when we describe our inference procedure.

When the interacting communities h, h' share the same parent, we simply choose the appropriate entry from their sibling group matrix. However, if h, h' do not share the same parent, then we invoke the following *coarsening* procedure:

1. Recall that $h = c_i[z_{\rightarrow ij}]$ and $h' = c_j[z_{\leftarrow ij}]$.
2. Find $z_{min} = \min(z_{\rightarrow ij}, z_{\leftarrow ij})$.
3. If $h_{coarse} = c_i[z_{min}]$ and $h'_{coarse} = c_j[z_{min}]$ are in the same sibling group, then we look up its compatibility matrix entry $\mathbf{B}_{h_{coarse}, h'_{coarse}}$. We then generate $E_{ij} \sim \mathrm{Bernoulli}(\mathbf{B}_{h_{coarse}, h'_{coarse}})$.
4. Otherwise, h_{coarse}, h'_{coarse} have zero compatibility, and we generate $E_{ij} = 0$.

Essentially, if actor i picks a community at a deeper level than actor j, then i coarsens up along his path to the same level as j. We can now formally define the $\mathrm{S}_B()$ function from the previous section:

$$\mathrm{S}_B(h, h') = \begin{cases} \mathbf{B}_{h_{coarse}, h'_{coarse}} \\ \quad h_{coarse} \text{ and } h'_{coarse} \text{ have same parent} \\ 0 \quad \text{otherwise} \end{cases}$$

$$\begin{aligned} h &= c_i[z_{\rightarrow ij}] \quad h' = c_j[z_{\leftarrow ij}] \\ h_{coarse} &= c_i[z_{min}] \quad h'_{coarse} = c_j[z_{min}] \\ z_{min} &= \min(z_{\rightarrow ij}, z_{\leftarrow ij}). \end{aligned}$$

For brevity, we define the shorthand $\mathrm{S}_B^{ij} := \mathrm{S}_B(h, h')$.

Finally, a $\mathrm{Beta}(\lambda_1, \lambda_2)$ prior is placed over every community compatibility parameter $\mathbf{B}_{h_{coarse}, h'_{coarse}}$. This adds the following step to our generative process:

- For each h_{coarse}, h'_{coarse}:
  - Sample $\mathbf{B}_{h_{coarse}, h'_{coarse}} \sim \mathrm{Beta}(\lambda_1, \lambda_2)$.

Increasing λ_1 will bias the model towards having more edges, whereas increasing λ_2 biases the model towards having fewer edges. Intuitively, λ_1 is the number of fake edges we are willing to assume, while λ_2 is our assumed number of fake non-edges.

A graphical model representation of our generative process can be found in Figure 2.

3 COLLAPSED GIBBS SAMPLER INFERENCE

Exact inference on our model is intractable, so we derive a collapsed Gibbs sampling scheme for posterior inference. The compatibility matrices $\mathbf{B}$ present a challenge since they can change in number/size as nodes are added and deleted during the sampling process (because the hierarchy structure is not prespecified). To finesse this issue, we analytically integrate them out using Beta-Bernoulli conjugacy, which adds dependence among interactions that implicitly share a compatibility parameter.

For faster mixing, the θ_i's are also integrated out. Thus, the only variables that need to be explicitly sampled are the levels $\mathbf{z}$ and the paths $\mathbf{c}$. The sampling equations are provided below.

Sampling Levels The distribution of $z_{\rightarrow ij}$ conditioned on all other variables is

$$\begin{aligned}&\mathbb{P}(z_{\rightarrow ij} \mid \mathbf{c}, \mathbf{z}_{-(\rightarrow ij)}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \propto \qquad (1)\\ &\qquad \mathbb{P}(E_{ij} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(ij)}, \lambda_1, \lambda_2)\mathbb{P}(z_{\rightarrow ij} \mid \mathbf{z}_{i,(-j)}, m, \pi)\end{aligned}$$

where $\mathbf{E}_{-(ij)}$ is the set of all edges except E_{ij}, and $\mathbf{z}_{i,(-j)} = \{z_{\rightarrow i\cdot}, z_{\leftarrow \cdot i}\} \setminus z_{\rightarrow ij}$. The first term, for a particular value of $z_{\rightarrow ij}$, is equal to

$$\begin{cases} \frac{\Gamma(a+b+\lambda_1+\lambda_2)}{\Gamma(a+\lambda_1)\Gamma(b+\lambda_2)} \cdot \frac{\Gamma(a+E_{ij}+\lambda_1)\Gamma(b+(1-E_{ij})+\lambda_2)}{\Gamma(a+b+1+\lambda_1+\lambda_2)} & \mathrm{S}_B^{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} a &= \left|\left\{(x,y) \mid (x,y) \neq (i,j), \mathrm{S}_B^{xy} = \mathrm{S}_B^{ij}, E_{xy} = 1\right\}\right| \\ b &= \left|\left\{(x,y) \mid (x,y) \neq (i,j), \mathrm{S}_B^{xy} = \mathrm{S}_B^{ij}, E_{xy} = 0\right\}\right| \qquad (2) \end{aligned}$$

In Eq. 2, the compatibility matrices $\mathbf{B}$ have been integrated out via Beta-Bernoulli conjugacy. As a result, $z_{\rightarrow ij}$ depends on the interactions E_{xy} that share E_{ij}'s compatibility parameter *at this point in the sampling process*.

The second term is computed by conditioning on the stick-breaking lengths $V_1, ..., V_K$ associated with $z_{\rightarrow ij}$:

$$\begin{aligned}&\mathbb{P}(z_{\rightarrow ij} = k \mid \mathbf{z}_{i,(-j)}, m, \pi) = \qquad (3)\\ &\qquad \frac{m\pi + \#[\mathbf{z}_{i,(-j)} = k]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq k]} \prod_{u=1}^{k-1} \frac{(1-m)\pi + \#[\mathbf{z}_{i,(-j)} > u]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq u]}\end{aligned}$$

Since we have limited the maximum depth to K, we simply ignore the event $z_{\rightarrow ij} > K$, and renormalize the distribution of $z_{\rightarrow ij}$ over the domain $\{1, \ldots, K\}$. The distribution of $z_{\leftarrow ij}$ is derived in similar fashion.

The runtime complexity of sampling a single z_{ij} is $\mathrm{O}(K)$, where K is the (fixed) depth of our hierarchy. Hence the total runtime for all z is $\mathrm{O}(N^2K)$.

Sampling Paths The distribution of c_i conditioned on all other variables is

$$\begin{aligned}&\mathbb{P}(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \propto \qquad (4)\\ &\qquad \mathbb{P}(\mathbf{E}_{(i\cdot),(\cdot i)} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i\cdot),-(\cdot i)}, \lambda_1, \lambda_2)\mathbb{P}(c_i \mid \mathbf{c}_{-i}, \gamma)\end{aligned}$$

where $\mathbf{E}_{(i\cdot),(\cdot i)} = \{E_{xy} \mid x = i \text{ or } y = i\}$ is the set of edges E_{xy} that depend on c_i, and $\mathbf{E}_{-(i\cdot),-(\cdot i)}$ is its complement. The second term can be computed using the nCRP definition (refer to the Supplemental). The first term, for a particular value of c_i, is

$$\begin{cases} \displaystyle\prod_{B \in \mathbb{B}_{(i\cdot),(\cdot i)}} \frac{\Gamma(g_B+h_B+\lambda_1+\lambda_2)}{\Gamma(g_B+\lambda_1)\Gamma(h_B+\lambda_2)} \cdot \frac{\Gamma(g_B+r_B+\lambda_1)\Gamma(h_B+s_B+\lambda_2)}{\Gamma(g_B+h_B+r_B+s_B+\lambda_1+\lambda_2)} \\ \qquad \forall E_{xy} \in \mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy} \neq 0 \\ 0 \qquad \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbb{B}_{(i\cdot),(\cdot i)} &= \{\mathbf{B}_{h,h'} \mid \exists(i,j)[E_{ij} \in \mathbf{E}_{(i\cdot),(\cdot i)}, \mathrm{S}_B^{ij} = \mathbf{B}_{h,h'}]\} \\ g_B &= \left|\{(x,y) \mid E_{xy} \in \mathbf{E}_{-(i\cdot),-(\cdot i)},\ \mathrm{S}_B^{xy} = B, E_{xy} = 1\}\right| \\ h_B &= \left|\{(x,y) \mid E_{xy} \in \mathbf{E}_{-(i\cdot),-(\cdot i)},\ \mathrm{S}_B^{xy} = B, E_{xy} = 0\}\right| \\ r_B &= \left|\{(x,y) \mid E_{xy} \in \mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy} = B, E_{xy} = 1\}\right| \\ s_B &= \left|\{(x,y) \mid E_{xy} \in \mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy} = B, E_{xy} = 0\}\right| \qquad (5) \end{aligned}$$

where $\mathbb{B}_{(i\cdot),(\cdot i)}$ is the set of community compatibility parameters $\mathbf{B}_{h,h'}$ associated with some edge in $\mathbf{E}_{(i\cdot),(\cdot i)}$. Similar to Eq. (2), Eq. (5) is a consequence of integrating out $\mathbf{B}$ for all interactions E associated with actor i.

The runtime for a single c_i is $\mathrm{O}(NH)$, where H is the number of hierarchy nodes. Hence the time to sample all c is $\mathrm{O}(N^2H)$. Note that $H = \mathrm{O}(NK)$, so the complexity of sampling all c is $\mathrm{O}(N^3K)$.

4 SIMULATION

We first evaluate our model's ability to recover hierarchies on simulated data. Our focus is to examine how our model's ability to model both assortative (within-community) interactions and disassortative (cross-community) interactions differentiates it from a traditional hierarchical clustering algorithm.

Our experiments explore the effect of different compatibility matrices $\mathbf{B}$. We first explore an on-diagonal $\mathbf{B}$, whose diagonal elements are much larger than the off-diagonals (strong assortative interactions). We also investigate an off-diagonal $\mathbf{B}$, whose off-diagonal elements are larger (strong disassortative interactions). For either $\mathbf{B}$ type, we experimented with maximum hierarchy depths $K = 2$ and 3. For the $K = 2$ simulations, the number of actors N was 150, while for $K = 3$ we used 300 actors. *Additional details and experiments can be found in the Supplemental.*

We compare our approach to hierarchical spectral clustering (denoted HSpectral). For spectral clustering, it is unclear how the number of clusters at each node would be selected, so *we give it the number of 1st-level branches as an advantage* (and then let it do a binary split at the deeper levels). For our model, we fix $m = \pi = \lambda_1 = \lambda_2 = .5$ and search over $\gamma = \{.01, .1, .5, 1, 1.5, 2\}$, picking the value that maximizes the marginal likelihood. We calculate the F1 score at each level k, $\mathrm{F1}_k = \frac{2*Precision*Recall}{Recall+Precision}$ where $Recall = \frac{TP}{TP+FN}$, and $Precision = \frac{TP}{TP+FP}$. TP is true positive count (actors that should be in the same cluster, up to depth k), FP is false positive count, TN is true negative count, and FN is false negative count. The total F1 score is computed by averaging the $\mathrm{F1}_k$ scores for each level.

Figure 4 illustrates our results as a function of the number of branches at the first level of the generated tree. As one can see, in Figure 4(a) and Figure 4(b), when $\mathbf{B}$ is strongly on-diagonal, our algorithm performs well, but a little worse than HSpectral (since HSpectral is given the number of level 1 branches). A specific example for $K = 2$ is shown in Figures 4(e), 4(f), 4(g), 4(h), on which both models performed reasonably well.

However, when $\mathbf{B}$ is strongly off-diagonal (implying strong cross-community interactions), HSpectral performs poorly. This is because, by formulation, spectral clustering cannot recover a disassortative community. On the other hand, our method still gives good results (Figure 4(c), Figure 4(d)).

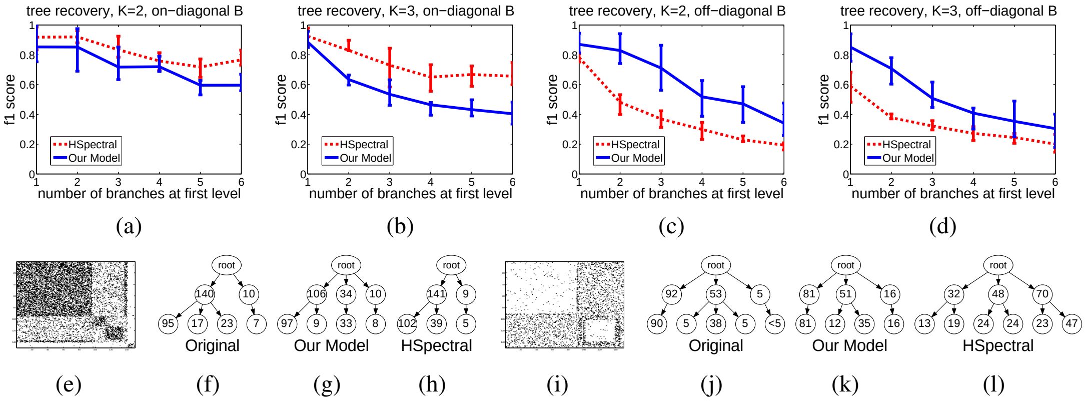

Figure 4: Simulation Results. Figures 4(a), 4(b), 4(c), and 4(d) show quantitative results. Figures 4(e), 4(f), 4(g) 4(h) illustrate results for one on-diagonal (assortative) network, and Figures 4(i), 4(j), 4(k) , 4(l) illustrate results for one off-diagonal (disassortative) network. 4(e) and 4(i) are the original networks for these two cases (black indicates edge). The numbers inside hierarchy nodes are actor counts (nodes of size < 5 are not shown). See text for details.

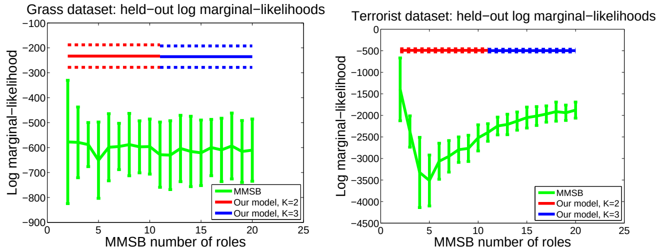

Figure 5: Held-out marginal likelihoods for our model and MMSB. Dotted lines show our model's error bars.

A $K = 2$ example is shown in Figures 4(i), 4(j), 4(k), 4(l) where our model performs accurately while HSpectral essentially divides the actors randomly and performs poorly. Thus our model successfully captures a variety of community interactions that traditional clustering methods cannot. Moreove, it can also recover the actor-specific multi-scale interaction levels for a richer network analysis.

5 HELD-OUT EVALUATION

Having evaluated how our model compares to a traditional hierarchical clustering method without a latent space, we now seek to compare it to latent space models that do not account for hierarchical structure. Since our latent space is integrated with the hierarchy, it is not possible to compare to a "non-hierarchical version" of our model. MMSB (Airoldi, Blei, Fienberg, and Xing 2008) seems the best choice for comparison, since it has analogous (but different) notions to our multi-scale membership and community compatibility in a non-hierarchical setting.

We use two real-world datasets, a 75-species food web of grass-feeding wasps (Dawah, Hawkins, and Claridge 1995; Clauset, Moore, and Newman 2008), and the 9/11 hijacker network consisting of 62 terrorists (Krebs 2002; Clauset, Moore, and Newman 2008). Our choices reflect two common modes of interaction seen in real-world network data: edges in the food web denote predator-prey relationships, while edges in the terrorist network reflect social cohesion. The food web could be represented as a hierarchy where different branches reflect different trophic levels (e.g. parasite, predator or prey), while the terrorist network could be interpreted as an organization chart. In the following experiments, we compare our model to MMSB using held-out marginal likelihood; models with higher likelihood imply a better fit to the data.

For each dataset, we generated 5 sets of training and test subgraphs; each train/test pair was obtained by randomly partitioning the actors into two equal sets, and keeping only the edges within each partition. With each train/test pair, we first used the training subgraph to select an appropriate prior on the community compatibility parameters $\mathbf{B}$, by performing a gridsearch over $(\lambda_1, \lambda_2) \in \{.1, .3, .5, .7, .9\}^2$ according to the log marginal likelihood. The remaining parameters were fixed to $\gamma = 1, m = \pi = 0.5$, as we found our results to be relatively insensitive to them. Using the best gridsearch parameters, we then estimated the log marginal likelihood on the corresponding test subgraphs, averaging over them to obtain our model's average held-out likelihood. This entire procedure was conducted for maximum hierarchy depths $K = 2$ and 3. The procedure for MMSB was similar, except that we used 100 random restarts of the MMSB variational EM algorithm on the training subgraphs to select the best parameters. MMSB also requires the number of latent roles R as a tuning parameter, so we repeated the experiment for each $2 \leq R \leq 20$. For either algorithm, log marginal likelihoods were estimated using 10,000 importance samples.

The results are shown in Figure 5. On both datasets, our model's held-out likelihood for either value of K is superior to MMSB for all R. Notably, MMSB's likelihood peaks on both datasets at $R = 2$, but selecting so few roles will lead to an extremely coarse network analysis. In contrast, our model automatically recovers a suitable level of hierarchical complexity and enables rich interpretations of the data — as we shall demonstrate next.

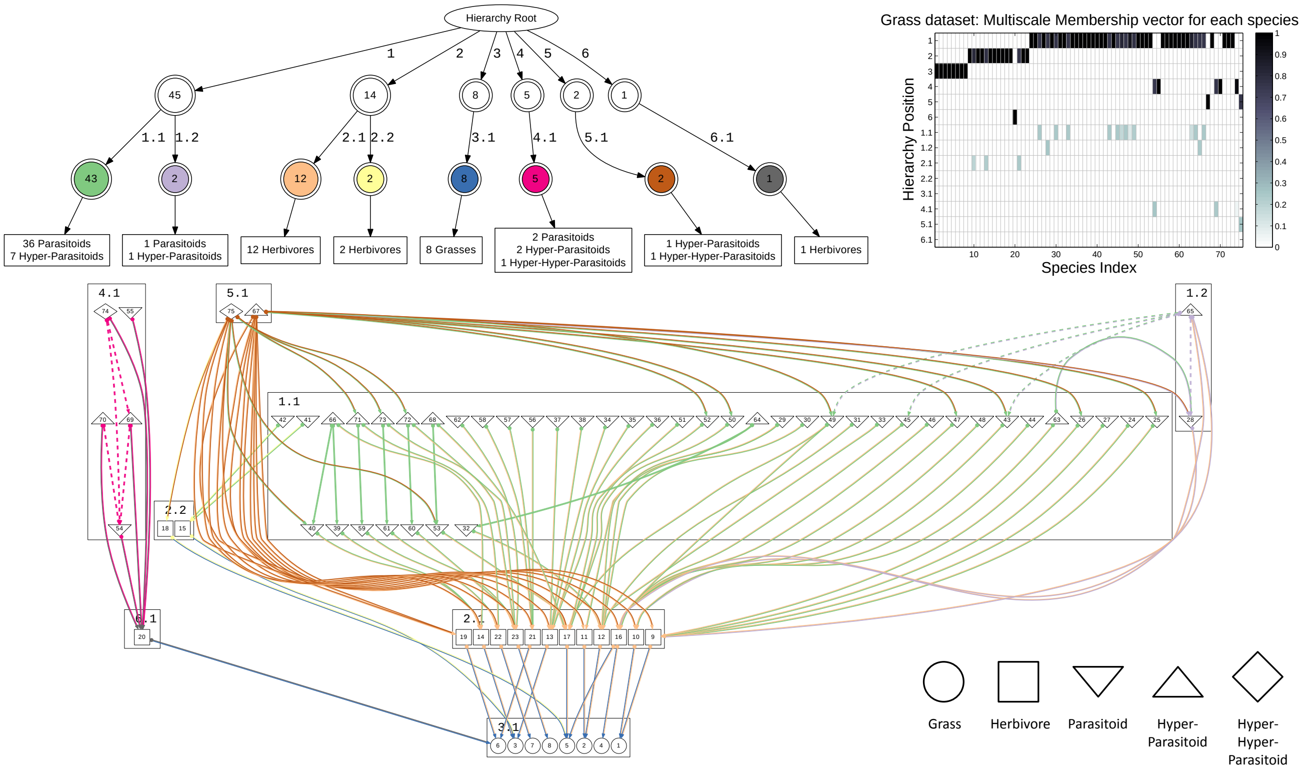

Figure 6: Grass network: **Top Left:** Inferred hierarchy of communities, with community trophic level counts at the bottom. **Top Right:** Multiscale Membership vectors for each actor. **Bottom:** Original network. Edges show interacting communities (edge head/tail colors match assumed hierarchy underlying the interactions) and interaction level (1 = solid, 2 = dashed) inferred by our model. Node shapes represent annotated trophic levels (see legend in bottom right).

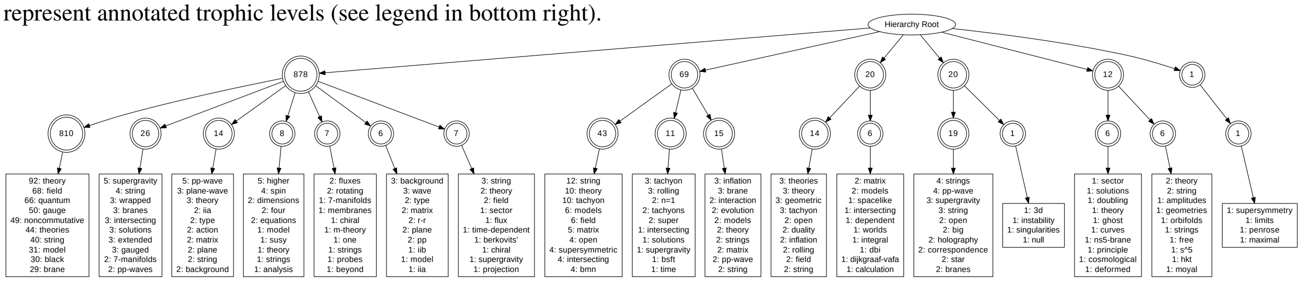

Figure 7: HEP network: Inferred hierarchy of communities, with the most frequent title keywords at the bottom. Community positions (circles) show the number of papers.

6 REAL-DATA QUALITATIVE ANALYSIS

In this section, we apply our model to interpret two real-world networks. We demonstrate that our model recovers the three network aspects we seek: hierarchy, multiscale granularity, and assortativity/disassortativity.

For both experiments, we use the optimal parameters from a held-out gridsearch similar to the previous section. We then ran our Gibbs sampler for 10,000 burn-in iterations, and took 500 samples. In order to account for posterior spread, we report a "consensus" sample that is analogous to an average. A description of the consensus and other experimental details can be found in the Supplemental.

6.1 Grass-Feeding Wasp Parasitoids Food Web

We begin with the earlier grass dataset, consisting of 75 species in a food web, and in which interactions represent predator-prey relationships. This dataset annotates each species with its position or "trophic level" in the food web: grass, herbivore, parasitoid, hyper-parasitoid (parasites that prey on other parasites), and hyper-hyper parasitoid. Our Gibbs sampler's inferred community hierarchy and Multiscale Membership (MM) vectors are reported in Figure 6. We also show the original network, where each interaction $E_{ij} = 1$ has been augmented with its associated community and interaction level (missing links $E_{ij} = 0$ are not shown). Trophic level annotations are shown in the hierarchy as counts, and in the network as node shapes.

In general, the level 1 super-communities separate the trophic levels. For example, community 3 contains all grass species, 2 contains most herbivores, and 1 contains most parasitoids. Note that the trophic levels form a set of *disassortative* communities, e.g. herbivores feed on grasses, but not on other herbivores. In contrast, Clauset *et al.*'s

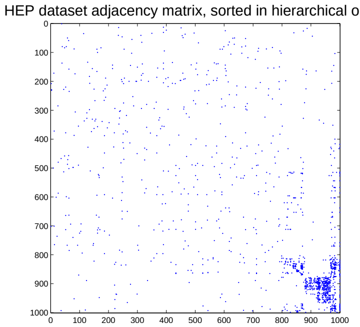

Figure 8: HEP network: Adjacency matrix, permuted according to the communities in Figure 7.

method, which assumes assortative communities, did not recover this structure in their experiments (Clauset, Moore, and Newman 2008).

The smaller super-communities are still more interesting: for instance, the herbivore in super-community 6 is the sole prey of the parasitoids in super-community 4, which justifies its separation from the other herbivores in super-community 2. Moreover, community 5 solely contains the two apex parasitoids with the largest and 2nd-largest range of prey species.

At level 2, the communities are separated by more subtle criteria than just trophic levels. The herbivores in community 2.2 are the sole prey of species 42 and 41 in community 1.1, while community 1.2 contains another apex parasitoid with an especially large range of prey species. In both cases, our model has separated these auxiliary food webs from the main web.

We now investigate the Multiscale Memberships recovered by our model. The MM vectors in Figure 6 show the frequency at which each species interacts as a member of a particular super- or sub-community. Most species identify at the super-community (i.e. generic) level, though some occasionally identify at the sub-community level. Our results show that level 2 interactions occur only within super-communities, hence they account for fine-grained, within-community interactions. For example, the within-community links in community 4, as well as the links from species 65 in sub-community 1.2 to other members of community 1, are all level 2 interactions. Note that we have not shown interaction levels for missing links, and a number of these are accounted for by level 2 interactions (e.g. in community 1).

6.2 High-Energy Physics Citation Network

Finally, we consider a 1,000-paper subgraph of the arXiv high-energy physics citation network, taken from the 2003 KDD Cup (2010). We constructed this subgraph by subsampling papers involved in citations from Jan 2002 through May 2003. Our Gibbs sampler completed 10,000 samples in just under 23 hours on a single processor, demonstrating that our algorithm scales to networks with thousands of actors.

The inferred community hierarchy is shown in Figure 7, where each sub-community has been annotated with its papers' most frequent title words[2]. We also show the adjacency matrix in Figure 8, permuted to match the order of inferred communities.

As expected, our model learns communites reflecting specific areas of study (an assortative network). The giant 810-paper level 2 community has a sparse citation pattern, implying that its papers are not specific to any research topic. This is confirmed by the top 3 keywords: 'theory', 'field' and 'quantum', which are general to physics research. The other level 2 communities under the same parent are more focused, with specific physical concepts like 'supergravity', 'string' and 'pp-wave'. This is also reflected in the adjacency matrix, which is denser among these communities. The remaining super-communities form a dense subnetwork mostly separated from the rest, implying narrower research foci. In particular, three of the sub-communities involve the title keyword "tachyon", which is absent from the giant level 1 community.

7 CONCLUSION

We have developed a nonparametric Multiscale Community Blockmodel (MSCB) that models social networks in terms of the hierarchical community memberships that actors undertake during interactions. Our model automatically infers the structure of the hierarchy while simultaneously recovering the Multiscale Memberships of every actor, setting it apart from hierarchy-discovering methods that are restricted to binary hierarchies and/or single-community-memberships for actors. Moreover, our model is expressive enough to account for both assortative (within-community) and disassortative (cross-community) interactions, as we have demonstrated through our simulation and real dataset experiments. We believe these aspects make our model suitable for exploring and understanding real-world network phenomena.

Acknowledgements

This paper is based on work supported by NSF IIS-0713379, NSF DBI-0546594 (Career), ONR N000140910758, AFOSR FA9550010247, NIH 1R01GM093156, and an Alfred P. Sloan Research Fellowship to Eric P. Xing. Qirong Ho is supported by a graduate fellowship from the Agency for Science, Technology And Research, Singapore. Le Song is supported by a Stephenie and Ray Lane Fellowship.

[2] While this output is reminiscent of topic models, our model is *not* a topic model. The hierarchy is learnt only from the citation network, without the paper contents.

References

Airoldi, E., D. Blei, S. Fienberg, and E. Xing (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research 9*, 1981–2014.

Blei, D., T. Griffiths, and M. Jordan (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM) 57*(2), 1–30.

Blei, D., A. Ng, and M. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Clauset, A., C. Moore, and M. Newman (2008). Hierarchical structure and the prediction of missing links in networks. *Nature 453*(7191), 98–101.

Clauset, A., M. Newman, and C. Moore (2004). Finding community structure in very large networks. *Physical Review E 70*(6), 66111.

Dawah, H., B. Hawkins, and M. Claridge (1995). Structure of the parasitoid communities of grass-feeding chalcid wasps. *Journal of animal ecology 64*(6), 708–720.

Girvan, M. and M. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821.

Guimera, R. and L. Amaral (2005). Functional cartography of complex metabolic networks. *Nature 433*, 895–900.

Handcock, M., A. Raftery, and J. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society-Series A 170*(2), 301–354.

Hoff, P., A. Raftery, and M. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*, 1090–1098.

KDD (2010, June). KDD Cup 2003 - Datasets. `http://www.cs.cornell.edu/projects/kddcup/datasets.html`.

Kemp, C. and J. Tenenbaum (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences 105*(31), 10687.

Kemp, C., J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence*, Volume 21, pp. 381. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Krause, A., K. Frank, D. Mason, R. Ulanowicz, and W. Taylor (2003). Compartments revealed in food-web structure. *Nature 426*(6964), 282–285.

Krebs, V. (2002). Mapping networks of terrorist cells. *Connections 24*(3), 43–52.

Miller, K., T. Griffiths, and M. Jordan (2009). Nonparametric Latent Feature Models for Link Prediction. *Advances in Neural Information Processing Systems (NIPS)*.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto, and D. Parisi (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences 101*(9), 2658.

Roy, D., C. Kemp, V. Mansinghka, and J. Tenenbaum (2007). Learning annotated hierarchies from relational data. *Advances in neural information processing systems 19*, 1185.

Teh, Y. and D. Roy (2009). The Mondrian Process. *Advances in neural information processing systems*.

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