
Infinite Hierarchical MMSB Model for Nested Communities/Groups in Social Networks

Qirong Ho

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
qho@cs.cmu.edu

Ankur P. Parikh

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
apparikh@cs.cmu.edu

Le Song

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lesong@cs.cmu.edu

Eric P. Xing

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
exping@cs.cmu.edu

Abstract

Actors in realistic social networks play not one but a number of diverse roles depending on whom they interact with, and a large number of such role-specific interactions collectively determine social communities and their organizations. Methods for analyzing social networks should capture these multi-faceted role-specific interactions, and, more interestingly, discover the latent organization or hierarchy of social communities. We propose a hierarchical Mixed Membership Stochastic Blockmodel 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. Furthermore, to automatically discover these latent structures from social networks, we develop a Gibbs sampling algorithm for our model. We conduct extensive validation of our model 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 the social communities 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 roles and minds of actors which lead to the formation and organization of these communities. We are particularly interested in uncovering the functional/sociological underpinning of network actors, and their influence on the network modularity and hierarchy.

Existing methods for inferring actor roles [10, 16, 9] are limited in that they assume each actor undertakes a single, invariant role when interacting with other actors. This is in stark contrast with real social networks where actors can play multiple roles (or be under multiple cultural influences); the specific role being played depends on whom the actor is interacting with. The *Mixed membership stochastic blockmodel* (MMSB) [2] proposed recently captures the multi-faceted, context-specific nature of an actor's role.

However, the presence of hierarchical structures in real-world social networks raises a significant challenge to current network models. The MMSB formalism described above is intrinsically flat in its structure — all roles of an actor are equal citizens in terms of their relationships to each other — so it does not induce hierarchical structures among actors. A hierar-

chical structure goes beyond simple clustering by explicitly including organization at all scales in a network simultaneously. Taking a corporate email network as example (Figure 1), suppose a finance department employee in the European branch is communicating with an American branch executive. The European employee might not communicate with a (possibly business-oriented) role related to her membership in the European finance department (which normally has no business with the American finance department), but simply with a more generic social role as a member of the European branch of the same company (which has occasional acquaintance with the American colleagues). However, when this same employee interacts with a European human resource employee, she would likely do so with her role as a European finance department member. This notion of *interaction-specific, multi-community membership* accounts for otherwise unusual edges — like the interaction between the European finance employee and the American executive — and motivates an intuitive yet powerful approach that allows one to infer and visualize the semantic underpinnings of every actor and every link between actors.

We propose a tree-structured hierarchical MMSB (hMMSB) to model community-hierarchies and the resultant social network as a function of the community memberships undertaken by actors during different interactions. Our method treats the community structure as a hierarchy of super- and sub-communities, while accounting for the interaction-specific nature of community membership. Moreover, it automatically determines the appropriate number of sub-communities in each community. This sets our method apart from spectral clustering methods, which typically do not produce hierarchies, or agglomerative clustering methods, which are inherently restricted to binary hierarchies. We will use our model to analyze a biological food web and a physics paper citation network, and examine whether the recovered hierarchies agree with the known organizational structure of both networks. Extensive simulation-based validation of both our model’s consistency and the Gibbs-sampling inference algorithm will be conducted to ensure our model’s feasibility for hierarchical structure identification.

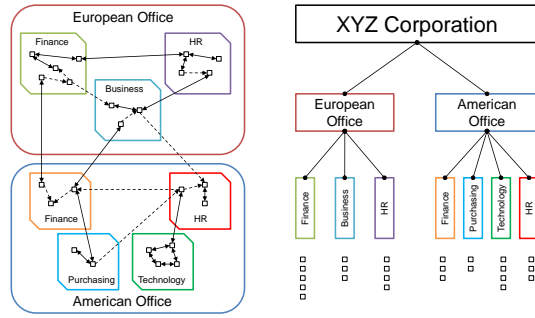


Figure 1: Illustration of a corporate email network/graph as a set of nested communities (**Left**), and the corresponding hierarchy of communities (**Right**). Vertices in the network represent employees, and directed edges represent emails sent on a particular day.

1.1 Related Work

While there has been a great deal of work on graph clustering and community structure inference [7, 14, 18, 5, 8], probabilistic generative models for hierarchical community formation and latent multi-role inference of every actor and link have just started to draw attention.

As mentioned earlier, the MMSB model [2] enables inference of the latent roles of every actor and link in a network, but it cannot capture hierarchical structures of possible communities in the network. The link prediction model in [17] employs an Indian Buffet Process prior over actor positions in an infinite-dimensional latent feature space. In that respect, it may be thought of as a nonparametric extension of the MMSB. However, the goal of their model is missing link prediction rather than inference of latent organizational structure.

Recently, Roy *et. al.* have developed an “annotated hierarchies” model [19] that generalizes an infinite relational model [13] for hierarchical group discovery. Their subsequent work on Mondrian Processes [20] provides an alternative nonparametric take on the same issues. Both the annotated hierarchies model and the Mondrian Process relational model discover *binary* hierarchies. The infinite stochastic blockmodel [12] recovers organizational hierarchies of arbitrary branching factor, among other types of structure like grids or partitions, but it does not offer detailed semantic underpinnings, such as context-dependent role or community membership instantiations, of every network link.

2 Probabilistic Hierarchical Community Model

Under the MMSB [2], network links can be instantiated by a role-specific *local* interaction mechanism: the link between each pair of actors, say (i, j) , is instantiated according to the latent role specifically undertaken by actor i when it is to interact with j , and that of j when it is to interact with i . Crucial to the goal of role-prediction for network data, is the so-called “role vector” θ of *mixed membership coefficients*, which succinctly captures the probabilities of an actor being involved in different roles when he/she interacts with another actor. In this paper, we leverage and generalize the “role vector” to model context-dependent mixed memberships in communities of different scales when links between actors are formed. We propose a *hierarchical MMSB* (hMMSB) model for nested community structures underlying social networks, thereupon such structures as well as the community implications of every link can be inferred, given the network. In the sequel, we begin with a baseline hMMSB that requires a user to pre-specify the granularity of a tree hierarchy of communities, that is, the branching degree from each group to its subgroup at the next level. We will then relax this constraint by employing a nested Chinese Restaurant Process, described in [3], to allow hierarchies of arbitrary granularities with a fixed depth.

Let $G = \{V, E\}$ denote a directed graph, where $V = \{1, \dots, N\}$ is the set of vertices/actors, and E is the set of edges/interactions. We adopt the convention $E_{ij} = 1$ if $(i, j) \in E$ else $E_{ij} = 0$, and we ignore self-edges E_{ii} . Because the graph is directed, E is not necessarily a symmetric matrix. Given an edge E_{ij} , we refer to actor i as the *donor* and actor j as the *receiver*.

Under an hMMSB, we assume that a link E_{ij} follows a generative process based on two latent variables unique to each actor. The first one is a *community membership-path vector* (or in short, *path*) c_i , which records the community memberships of this actor at different levels of the community hierarchy in the network. Our model requires an integer parameter K , denoting the maximum depth to which the hierarchy should be learnt. Given K , we can represent an actor path c_i as a vector of *positive integers* $(c_{i1}, c_{i2}, \dots, c_{iK})$, where c_{ik} denotes the branch taken by the path at level $k \geq 1$. We adopt the convention that the hierarchy root sits at level 0. For clarity of explanation, we will assume for now that the actor paths c_i are known. Later, we shall relax this assumption by placing a suitable nonparametric prior over c_i , allowing them to be posteriorly inferred.

The second variable is a *mixed membership vector* (in short, *MM*) θ_i , defining the probabilities of an actor identifying himself at different community levels when interacting with other actors. We place a two-parameter stick-breaking prior [3] over θ_i , so that actors are not required to participate in the same number of community levels. Stick breaking constructions work as follows: Consider a stick of length 1. Draw $V_{i1} \sim \text{Beta}(m\pi, (1-m)\pi)$. Let $\theta_{i1} = V_{i1}$ and let $1 - \theta_{i1}$ be the remainder of the stick after chopping off this length V_{i1} . To calculate the length θ_{i2} , draw $V_{i2} \sim \text{Beta}(m\pi, (1-m)\pi)$ and chop off this fraction of the remainder of the stick, giving $\theta_{i2} = V_{i2}(1 - V_{i1})$. Thus V_{ik} is the fraction to chop off from the stick’s remainder, and θ_{ik} is the length of the k th stick that was chopped off. In general, we draw $V_{ik} \sim \text{Beta}(m\pi, (1-m)\pi)$ from $k = 1$ to $k = \infty$ and the corresponding $\{\theta_{ik}\}_{k=1}^{\infty}$ is defined below:

$$\theta_{ik} = V_{ik} \prod_{u=1}^{k-1} (1 - V_{iu}) \quad (1)$$

This process is known as the two-parameter GEM distribution [3] and draws from $\text{GEM}(m, \pi)$ are denoted as $\theta_i \sim \text{GEM}(m, \pi)$. $m > 0$ influences the mean of θ_i , and $\pi > 0$ influences its variance. Because the hierarchy is only learnt up to depth K , we truncate the $\text{GEM}(m, \pi)$ distribution at level K . 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).

To instantiate the edge from actor i to j , we introduce *interaction level indicators* $z_{i \rightarrow j}$ (*donor level*) and $z_{i \leftarrow j}$ (*receiver level*), which follow multinomial distributions defined by the MM vectors θ_i, θ_j respectively. The pair $(c_i, z_{i \rightarrow j})$ specifies actor i ’s interaction-specific position in the community hierarchy, denoted by $c_i[z_{i \rightarrow j}]$, which contributes to determining his interaction probability with actor j . Note that we are *not* proposing a full hierarchical model for a tree over conditionally *iid* entities, as in a coalescent process [11]. Our goal is to model finite-depth nested communities of entities connected by a network; therefore our latent hierarchy is not a full binary tree over all nodes, and our generative process does not output *iid* nodal attributes, but links between nodes.

Now, given interaction level indicators $z_{i \rightarrow j}$ and $z_{i \leftarrow j}$, and paths c_i and c_j of the two actors involved, we can determine the community identities (not necessarily at the same level) underlying an interaction. These community identities specify a community-pair-specific distribution for the link $E_{ij} \sim \text{Bernoulli}(S_B(c_i, c_j, z_{i \rightarrow j}, z_{i \leftarrow j}))$, where

$$S_B(c_i, c_j, z_{i \rightarrow j}, z_{i \leftarrow j}) = \begin{cases} B_{c_i, c_j, z_{\text{coarse}}} & (c_{i,1}, \dots, c_{i, z_{\text{coarse}}-1}) = (c_{j,1}, \dots, c_{j, z_{\text{coarse}}-1}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$z_{\text{coarse}} = \min(z_{i \rightarrow j}, z_{i \leftarrow j}).$$

The notation $B_{c_i, c_j, z_{\text{coarse}}} \in \mathbf{B}$ denotes a parameter specifying the interaction probability from community $c_i[z_{\text{coarse}}]$ to $c_j[z_{\text{coarse}}]$. We shall explain the intuition behind this in more detail below.

With all these details, we arrive at the following generative process of a network $G = \{V, E\}$:

- For each actor $i \in V$:
 - Sample actor i 's path c_i from a distribution over hierarchies (detailed in the next section).
 - Sample actor i 's MM θ_i (distribution over community levels): $\theta_i \sim \text{GEM}(m, \pi)$.
- For each element $B_{\cdot, \cdot, \cdot}$ of \mathbf{B} in the community-compatibility matrix:
 - Sample a value for the element $B_{\cdot, \cdot, \cdot} \sim \text{Beta}(\lambda_1, \lambda_2)$.
- To generate the network, for each directed edge E_{ij} where $i, j \in V, i \neq j$:
 - Sample the donor level $z_{i \rightarrow j} \sim \text{Multinomial}(\theta_i)$.
 - Sample the receiver level $z_{i \leftarrow j} \sim \text{Multinomial}(\theta_j)$.
 - Sample the interaction $E_{ij} \sim \text{Bernoulli}(S_B(c_i, c_j, z_{i \rightarrow j}, z_{i \leftarrow j}))$.

The overall intuition behind hMMSB is that c_i and $z_{i \rightarrow j}$ specify a hierarchy position for the donor, and likewise for the receiver; and we use these positions to determine the interaction probability. There are two different scenarios worth mentioning. First, suppose that the donor and receiver positions share the same immediate parent in the hierarchy. Our model associates a ‘‘compatibility matrix’’ with every parent position in the hierarchy, which gives the probability of interaction between any two of its child positions. The set of all compatibility matrices is denoted by \mathbf{B} . Since the donor and receiver positions share the same parent, we simply look up the appropriate entry in \mathbf{B} to get the interaction probability.

In the second case, the donor and receiver positions do not share the same parent; this always happens when $z_{i \rightarrow j} \neq z_{i \leftarrow j}$ (i.e. the positions are at different depths). We then coarsen both levels to their minimum, z_{coarse} . The idea is that interactions between two hierarchy positions should take place at the level of the higher (coarser) position. For example, when an executive interacts with a sales representative, we reduce the interaction to one between an executive and a generic group of ‘‘other employees’’ below executive rank. (It is in theory possible to explicitly model all possible interactions between communities of arbitrary levels, but this will lead to a severe over-parameterization of our model. We have decided to postpone a more thorough study of this issue to later work.) If the new positions now share the same parent, then we look up the appropriate element of \mathbf{B} as before. Otherwise, we define the interaction probability to be zero.

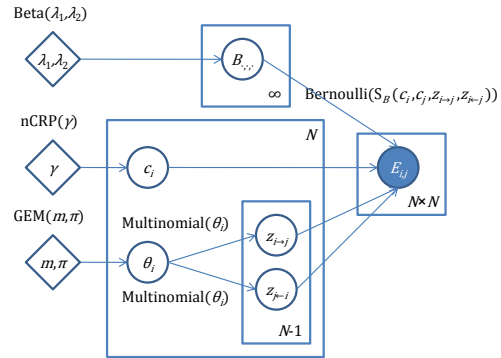


Figure 2: Graphical model for hMMSB.

2.1 Infinite hMMSB

Thus far, we have assumed knowledge of the paths c_i . From a combinatorial standpoint, inferring these paths is difficult. Because the number of children at each hierarchy position is unlimited,

there are infinitely many possible hierarchies or sets of paths $\{c_i\}$, even with the fixed depth K . 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.

Our solution is to place an appropriate nonparametric Bayesian prior over c_i , the Nested Chinese Restaurant Process [3], or nCRP for short. The nCRP is an extension of the regular Chinese Restaurant Process (CRP), a recursively-defined prior over positive integers. For concreteness, we shall use the first level of each actor path, c_{i1} , to define the CRP:

$$P(c_{i1} = x \mid c_{1:(i-1),1}) = \begin{cases} \frac{|\{j < i \mid c_{j1} = x\}|}{i-1+\gamma} & x \in \{c_{1:(i-1),1}\} \\ \frac{\gamma}{i-1+\gamma} & x \text{ is the smallest positive integer not in } \{c_{1:(i-1),1}\} \end{cases} \quad (3)$$

where $\gamma > 0$ is a ‘‘concentration’’ parameter that controls the probability of drawing new integers, and for conciseness we define $c_{1:(i-1),1} \equiv (c_{11}, \dots, c_{(i-1)1})$. The nCRP is essentially a hierarchy of CRP priors, beginning with a single CRP prior at the top level. With each unique integer x seen at the top-level prior, we associate a child CRP prior with $|\{i \mid c_{i1} = x\}|$ observations, resulting in a two-level tree of CRP priors. We can repeat this process *ad infinitum* on the newly-created child priors, resulting in an infinite-level tree of CRP priors, though we only use a K -level nCRP. All CRP priors in the nCRP share the same concentration parameter γ .

Now we can finish describing our generative process: for each actor $i \in V$, we can sample $c_i \sim \text{nCRP}(\gamma)$ using the recursive nCRP definition:

$$P(c_{ik} = x \mid c_{1:(i-1)}, c_{i,1:(k-1)}) = \begin{cases} \frac{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)} \wedge c_{jk} = x\}|}{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}| + \gamma} & x \in \{c_{jk} \mid (j < i) \wedge c_{j,1:(k-1)} = c_{i,1:(k-1)}\} \\ \frac{\gamma}{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}| + \gamma} & x \text{ is the smallest positive integer not in the above set.} \end{cases} \quad (4)$$

Figure 2 displays our complete generative process as a graphical model. As a side note, the infinite nCRP prior on paths implies that \mathbf{B} contains an infinite number of elements (there could be infinitely many children at each parent, so the compatibility matrices must be infinite-dimensional). Our Gibbs sampler finesses this issue by integrating out \mathbf{B} , while the posterior distribution of each $B_{i,j}$ can be recovered from the path and level samples.

3 Collapsed Gibbs Sampler

Exact inference on our model is intractable, so we derive a collapsed Gibbs sampling scheme for posterior inference. The θ ’s and \mathbf{B} ’s are integrated out for faster mixing, so we only have to sample \mathbf{z} and \mathbf{c} . The sampling equations are provided below.

Sampling levels The distribution of $z_{i \rightarrow j}$ conditioned on all other variables is

$$\begin{aligned} & \mathbb{P}(z_{i \rightarrow j} \mid \mathbf{c}, \mathbf{z}_{-(i \rightarrow j)}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \\ \propto & \mathbb{P}(E_{i,j}, z_{i \rightarrow j} \mid \mathbf{c}, \mathbf{z}_{-(i \rightarrow j)}, \mathbf{E}_{-(i,j)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\ = & \mathbb{P}(E_{i,j} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i,j)}, \gamma, m, \pi, \lambda_1, \lambda_2) \mathbb{P}(z_{i \rightarrow j} \mid \mathbf{c}, \mathbf{z}_{-(i \rightarrow j)}, \mathbf{E}_{-(i,j)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\ = & \mathbb{P}(E_{i,j} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i,j)}, \lambda_1, \lambda_2) \mathbb{P}(z_{i \rightarrow j} \mid \mathbf{z}_{i,(-j)}, m, \pi) \end{aligned} \quad (5)$$

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

$$\begin{aligned} \text{First term} &= \begin{cases} \frac{E_{ij}(a+\lambda_1)+(1-E_{ij})(b+\lambda_2)}{a+b+\lambda_1+\lambda_2} & S_B(E_{ij}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \\ a &= |\{(x,y) \mid (x,y) \neq (i,j) \wedge S_B(E_{xy}) = S_B(E_{ij}) \wedge E_{xy} = 1\}| \\ b &= |\{(x,y) \mid (x,y) \neq (i,j) \wedge S_B(E_{xy}) = S_B(E_{ij}) \wedge E_{xy} = 0\}| \end{aligned} \quad (6)$$

where we have defined the shorthand $S_B(E_{ij}) \equiv S_B(c_i, c_j, z_{i \rightarrow j}, z_{i \leftarrow j})$.

The second term can be computed by iterated expectation, conditioning on actor i 's stick breaking lengths V_{i1}, \dots, V_{ik} :

$$\begin{aligned}
& \mathbb{P}(z_{i \rightarrow j} = k \mid \mathbf{z}_{i,(-j)}, m, \pi) \\
&= \mathbb{E} [\mathbb{I}(z_{i \rightarrow j} = k) \mid \mathbf{z}_{i,(-j)}, m, \pi] \\
&= \mathbb{E} [\mathbb{E} [\mathbb{I}(z_{i \rightarrow j} = k) \mid V_{i1}, \dots, V_{ik}, \mathbf{z}_{i,(-j)}, m, \pi]] \\
&= \mathbb{E} \left[V_{ik} \prod_{u=1}^{k-1} (1 - V_{iu}) \mid \mathbf{z}_{i,(-j)}, m, \pi \right] \\
&= \mathbb{E}[V_{ik} \mid \mathbf{z}_{i,(-j)}, m, \pi] \prod_{u=1}^{k-1} \mathbb{E}[(1 - V_{iu}) \mid \mathbf{z}_{i,(-j)}, m, \pi] \\
&= \frac{m\pi + \#[\mathbf{z}_{i,(-j)} = k]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq k]} \prod_{u=1}^{k-1} \frac{(1 - m)\pi + \#[\mathbf{z}_{i,(-j)} > u]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq u]} \quad (7)
\end{aligned}$$

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

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

$$\begin{aligned}
& \mathbb{P}(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
&\propto \mathbb{P}(c_i, \mathbf{E}_{(i, \cdot), (\cdot, i)} \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}_{-(i, \cdot), -(\cdot, i)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
&= \mathbb{P}(\mathbf{E}_{(i, \cdot), (\cdot, i)} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i, \cdot), -(\cdot, i)}, \gamma, m, \pi, \lambda_1, \lambda_2) \mathbb{P}(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}_{-(i, \cdot), -(\cdot, i)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
&= \mathbb{P}(\mathbf{E}_{(i, \cdot), (\cdot, i)} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i, \cdot), -(\cdot, i)}, \lambda_1, \lambda_2) \mathbb{P}(c_i \mid \mathbf{c}_{-i}, \gamma) \quad (8)
\end{aligned}$$

where $\mathbf{E}_{(i, \cdot), (\cdot, i)} = \{E_{xy} \mid x = i \vee y = i\}$ is the set of all edges E_{ij} whose distributions depend on c_i , and $\mathbf{E}_{-(i, \cdot), -(\cdot, i)}$ is its complement. The second term can be computed using the recursive nCRP definition:

$$\begin{aligned}
& \mathbb{P}(c_{ik} = x \mid c_{1:(i-1)}, c_{i,1:(k-1)}, \gamma) = \\
& \begin{cases} \frac{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)} \wedge c_{jk} = x\}|}{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}| + \gamma} & x \in \{c_{jk} \mid (j < i) \wedge c_{j,1:(k-1)} = c_{i,1:(k-1)}\} \\ \frac{\gamma}{|\{j < i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}| + \gamma} & x \text{ is the smallest positive integer not in the above set.} \end{cases} \quad (9)
\end{aligned}$$

while the first term, for a particular value of c_i , is

$$\begin{aligned}
\text{First term} &= \begin{cases} \prod_{B \in \mathbf{B}_{(i, \cdot), (\cdot, i)}} \frac{\Gamma(g_B + h_B + \lambda_1 + \lambda_2)}{\Gamma(g_B + \lambda_1) \Gamma(h_B + \lambda_2)} \frac{\Gamma(g_B + r_B + \lambda_1) \Gamma(h_B + s_B + \lambda_2)}{\Gamma(g_B + h_B + r_B + s_B + \lambda_1 + \lambda_2)} \quad \forall E_{xy} \in \mathbf{E}_{(i, \cdot), (\cdot, i)}, S_B(E_{xy}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \\
g_B &= |\{(x, y) \mid E_{xy} \in \mathbf{E}_{-(i, \cdot), -(\cdot, i)} \wedge S_B(E_{xy}) = B \wedge E_{xy} = 1\}| \\
h_B &= |\{(x, y) \mid E_{xy} \in \mathbf{E}_{-(i, \cdot), -(\cdot, i)} \wedge S_B(E_{xy}) = B \wedge E_{xy} = 0\}| \\
r_B &= |\{(x, y) \mid E_{xy} \in \mathbf{E}_{(i, \cdot), (\cdot, i)} \wedge S_B(E_{xy}) = B \wedge E_{xy} = 1\}| \\
s_B &= |\{(x, y) \mid E_{xy} \in \mathbf{E}_{(i, \cdot), (\cdot, i)} \wedge S_B(E_{xy}) = B \wedge E_{xy} = 0\}| \quad (10)
\end{aligned}$$

where $\mathbf{B}_{(i, \cdot), (\cdot, i)} = \{B \in \mathbf{B} \mid \exists (i, j), (E_{ij} \in \mathbf{E}_{(i, \cdot), (\cdot, i)} \wedge S_B(E_{ij}) = B)\}$ is the set of all $B \in \mathbf{B}$ associated with some edge in $\mathbf{E}_{(i, \cdot), (\cdot, i)}$ through $S_B(\cdot)$.

4 Simulation

We evaluate our model's ability to recover hierarchies on simulated data. For all simulations, the number of actors N was 150, the max depth was $K = 2$ (2 levels plus root) and $\theta = (.25, .75)$ for all actors, meaning that actors interact at level 1 25% of the time and level 2 75% of the time. Our experiments explore the effect of different compatibility matrices \mathbf{B} . We first explore an on-diagonal \mathbf{B} , where the diagonal elements are much larger than the off-diagonal elements (actors

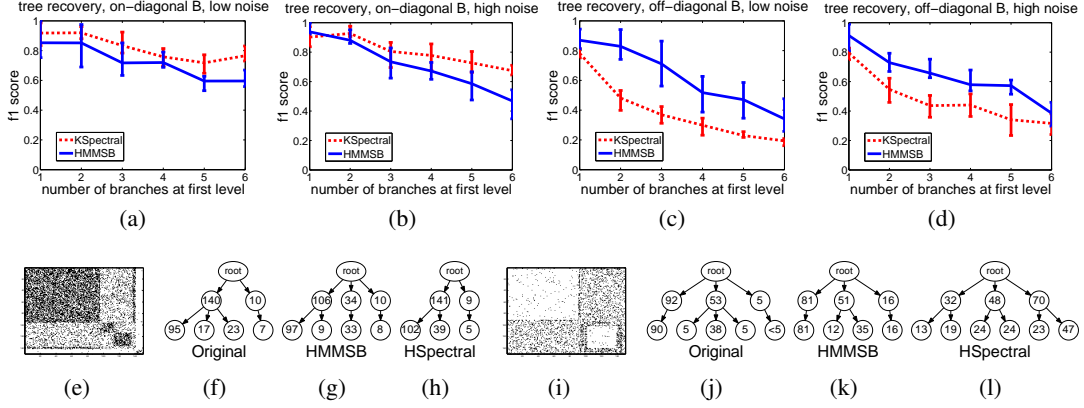


Figure 3: Simulation Results. Figures 3(a), 3(b), 3(c), and 3(d) show quantitative results. Figures 3(e), 3(f), 3(g) 3(h) illustrate results for one on-diagonal network, and Figures 3(i), 3(j), 3(k) , 3(l) illustrate results for one off-diagonal network. 3(e) and 3(i) are the original networks for these two cases (black indicates edge). The numbers inside hierarchy nodes are actor counts (nodes of size < 5 are not shown). See text for details.

tend to interact within their own communities). We also investigate an off-diagonal B , where the off-diagonal elements are larger (actors tend to interact outside their own communities). In the “low noise” setting the off-diagonal and on-diagonal elements are far apart (to clearly distinguish them), while in the “high noise” setting they are closer together. The exact parameters for the 4 types of B ’s are shown below:

1. **on-diagonal, low noise** - $B_{on-diagonal} = (.4, .8)$, $B_{off-diagonal} = (.02, .02)$;
2. **on-diagonal, high noise** - $B_{on-diagonal} = (.3, .6)$, $B_{off-diagonal} = (.1, .1)$;
3. **off-diagonal, low noise** - $B_{on-diagonal} = (.02, .02)$, $B_{off-diagonal} = (.4, .8)$;
4. **off-diagonal, high noise** - $B_{on-diagonal} = (.1, .1)$, $B_{off-diagonal} = (.3, .6)$.

$B_{on-diagonal} = (a, b)$ means that actors interacting in the same level-1 community do so with probability a , while actors interacting in the same level-2 community do so with probability b (and analogously for $B_{off-diagonal}$).

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 each level-2 node). For hMMSB, 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.

The Gibbs sampler was run for 1,500 iterations on each experiment. We calculate the F1 score at each level k , $F1_k = \frac{2 * Precision * Recall}{Recall + Precision}$ where $Recall = \frac{TP}{TP + FN}$, and $Precision = \frac{TP}{TP + FP}$. TP is true positive count (actors that are in the same cluster and should be 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 $F1_k$ scores for each level.

Figure 3 illustrates the results as a function of the number of branches at the first level of the generated tree. The “number of branches at level 1” refers to number of branches of size ≥ 5 (since there are often branches of size 1 or 2). For fairness, the “correct” number of level-1 branches given to spectral clustering is also the number of branches of size ≥ 5 .

As one can see, in Figure 3(a) and Figure 3(b), when 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 is shown in Figures 3(e), 3(f), 3(g), 3(h) where both models perform reasonably well. However, when B is strongly off-diagonal (implying that actors tend to interact with those in different communities and rather than within their own community), HSpectral performs poorly. Our method still gives good results (Figure 3(c), Figure 3(d)). An example is shown in

Figures 3(i), 3(j), 3(k), 3(l) where our model performs accurately while HSpectral essentially divides the actors randomly and performs poorly. Thus HMMSB successfully models a variety of community interactions that traditional clustering methods cannot. It can also recover the actor-specific interaction levels for a richer network analysis.

5 Held-out Evaluation

We compare our algorithm to MMSB [2] on two real-world datasets, a 75-species food web of grass-feeding wasps [6, 4], and the September 11th, 2001 hijacker terrorist network [15, 4]. 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.

For each dataset, we generated 5 sets of training and test subgraphs, each obtained by randomly partitioning the actors into two equal sets. For each training subgraph, we performed a gridsearch over the \mathbf{B} parameters $(\lambda_1, \lambda_2) \in \{.1, .3, .5, .7, .9\}^2$ according to the log marginal likelihood. The remaining parameters were fixed to $\gamma = 1, m = \pi = 0.5$ and $K = 2$. We then computed the log marginal likelihood on the corresponding 5 test subgraphs, and averaged them to obtain hMMSB’s average held-out likelihood. The procedure for MMSB was similar, except that we used 100 random restarts of the MMSB variational EM algorithm [2] on the training subgraphs. 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 iterations of importance sampling.

The results are shown in Figure 4. On either dataset, our model’s held-out likelihood 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. We note that the “annotated hierarchies” model [19] and infinite stochastic blockmodel [12] are good candidates for evaluation, but the authors did not make code available for computing marginal likelihoods, so we do not include those models in our evaluation.

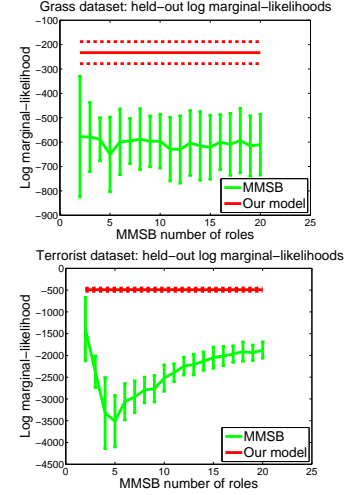


Figure 4: Marginal likelihood for MMSB and hMMSB (top: grass, bottom: terrorist). Dotted lines show hMMSB’s error bars.

6 Real-data Qualitative Analysis

6.1 Grass-feeding Wasp Parasitoids Food Web

We now use hMMSB to interpret real-world social networks, beginning with the grass dataset from earlier. We ran our Gibbs sampler on the full network to infer the community hierarchy structure and actor latent community MMs. The model parameters were chosen on the full network via grid search over $(\lambda_1, \lambda_2) \in \{.1, .3, .5, .7, .9\}^2$, according to the log marginal likelihood (estimated using 10,000 importance samples). The remaining parameters were fixed to $\gamma = 1, m = 0.5, \pi = 0.5$ and $K = 2$. We ran our Gibbs sampler on the optimal parameters $\lambda_1 = 0.1, \lambda_2 = 0.5$ for 10,000 iterations of burn-in, and took 100 samples with a lag time of 5 iterations. Convergence was determined from a plateauing log complete likelihood plot.

The Gibbs samples represent a posterior distribution over paths c_i . In order to represent the “average” of this posterior, we generated a consensus sample by counting the number of times each pair of actors shared the same community hierarchy position, over all samples. Actors that shared positions in $> 50\%$ of all samples were assigned to the same path in the consensus. For levels $z_{i \rightarrow j}$ and $z_{i \leftarrow j}$,

we simply took the mode over all samples. In a final post-processing step to reduce visual clutter, we merged bottom-level (i.e. level-2) communities with ≤ 5 actors into one community under the same parent.

The inferred community hierarchy and MM vectors from our Gibbs sampler are reported in Figure 6. We also show the original network, where links $E_{ij} = 1$ have been augmented with their communities and interaction levels (missing links $E_{ij} = 0$ are not shown). The dataset contains trophic level annotations, shown in the hierarchy as counts and in the network as node shapes.

We see that community 3 contains all grass species, 2 contains most herbivores, and 1 contains most parasitoids; in contrast, an assortative clustering algorithm would not discover these multi-partite divisions. The “outlier” communities are still more interesting — the herbivore in community 6 is the sole prey of the parasitoids in community 4; hMMSB has separated this sub-web from the main food-web. Moreover, the herbivores in community 2.2 are the sole prey of species 42 and 41 in community 1.1. We also observe that community 5’s two apex parasitoids have the largest and 2nd-largest range of prey species, while community 1.2 contains the apex parasitoid with the third-largest range. As to why these communities are separated, we note that the parasitoid in sub-community 1.2 preys on only two herbivores, compared to at least six herbivores for either parasitoid in community 5.

The community MM vectors in Figure 6 show the frequency at which each species identifies itself with a particular community. Most species identify at the super-community level, though some occasionally identify at the sub-community level. In our model, 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. Although we have not shown interaction levels for missing links, the latter are sometimes 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 [1], which we constructed by subsampling papers involved in citations from Jan 2002 through May 2003.

We applied the same parameter selection, convergence criteria, and post-processing as the previous dataset, and the optimal parameters were $(\lambda_1 = 0.7, \lambda_2 = 0.5)$. Each of the 25 parameter combinations required less than 6 hours to test on a single processor core. We ran our Gibbs sampler for 10,000 iterations of burn-in, and took 10 samples with a lag time of 50 iterations. The entire Gibbs sampling procedure completed in just under 23 hours on a single processor core.

The inferred community hierarchy is shown in Figure 7, where each sub-community has been annotated with its papers’ most frequent title words¹. We also show the adjacency matrix in Figure 5, permuted to match the order of inferred communities. We observe that the 810-paper sub-community has a sparse citation pattern, implying that its papers are not specific to a particular research topic. This is confirmed by the top 3 keywords: “theory”, “field” and “quantum”, which are general to physics research. The other sub-communities from the same super-community are more focused, with top keywords like “supergravity”, “string” and “pp-wave” being more specific physical concepts. This is also reflected in the adjacency matrix, which is denser among these sub-communities. The remaining super-communities form a very dense sub-network that is mostly separated from the rest of the graph, suggesting that these papers might come from a specific community of researchers, working on a narrow set of research topics. In particular, three

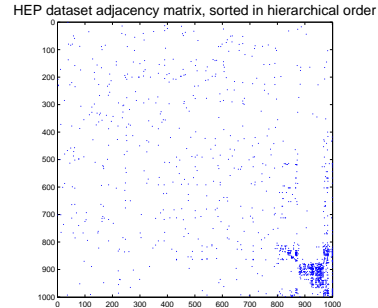


Figure 5: HEP network: Adjacency matrix, permuted according to the communities in Figure 7.

¹While this output is reminiscent of topic models, we stress that hMMSB is *not* a topic model — the actor paths are learnt solely from the citation network, independent of the paper contents.

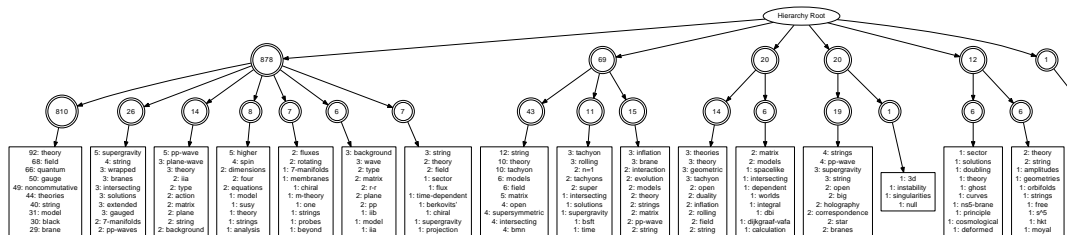
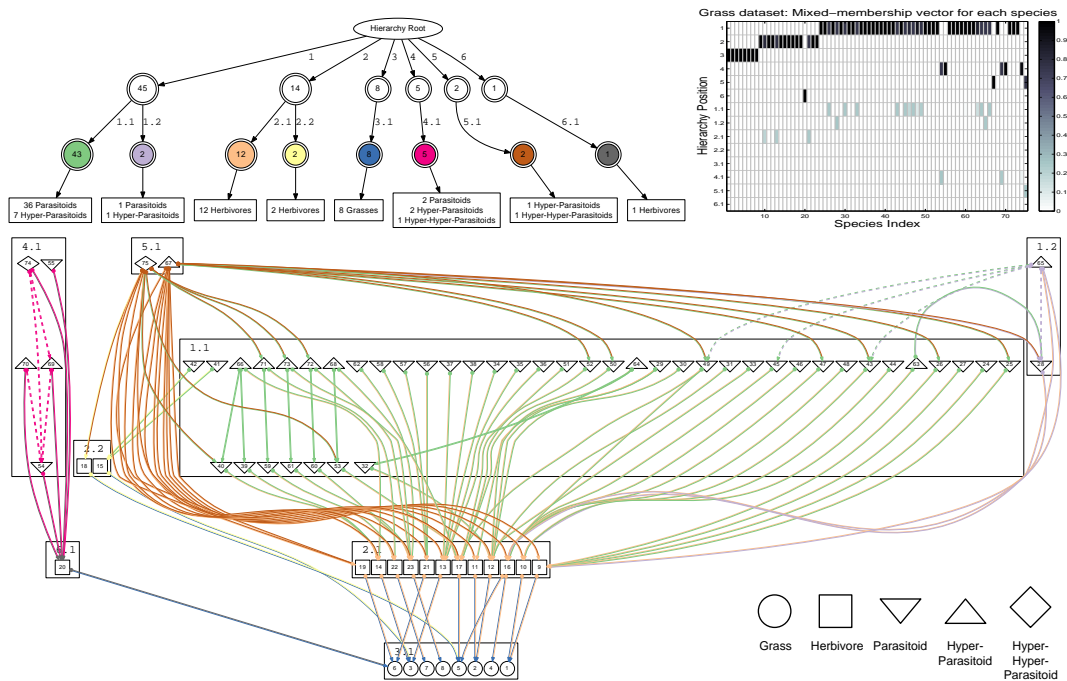


Figure 7: HEP network: Inferred hierarchy of communities, with the most frequent title keywords at the bottom. Community positions (circles) show the number of papers.

of the sub-communities involve the title keyword “tachyon”, which is absent from the large super-community. To this point, we have only scratched the surface of the citation network — a deeper analysis would involve the abstracts/contents of the papers in the hMMSB-inferred communities.

7 Conclusion

We have developed a tree-structured hierarchical Mixed-Membership Stochastic Blockmodel (hMMSB) that models social networks in terms of the multiple, hierarchical community memberships that actors undertake during interactions. Our model automatically infers the number of sub-communities in each community while simultaneously recovering the community-mixed-memberships of every actor, setting it apart from hierarchy-discovering methods that are restricted to binary hierarchies and/or single-role-memberships for actors. Moreover, hMMSB is expressive enough to account for non-diagonal community-compatibility matrices, as we have demonstrated

through our simulation and grass dataset experiments. On real networks, we show that hMMSB recovers intuitive, mixed-membership community organization. Finally, our collapsed Gibbs sampler efficiently scales to medium-sized datasets of around 1,000 actors, completing inference in a single day on a single processor core.

Acknowledgements

This paper is based on work supported by NSF IIS-0713379, AFOSR FA9550010247, ONR N000140910758, DARPA NBCH1080007, 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.

References

- [1] KDD 2003. KDD Cup 2003 - Datasets. <http://www.cs.cornell.edu/projects/kddcup/datasets.html>, June 2010.
- [2] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic block-models. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.
- [3] D.M. Blei, T.L. Griffiths, and M.I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30, 2010.
- [4] A. Clauset, C. Moore, and MEJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
- [5] A. Clauset, MEJ Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):66111, 2004.
- [6] H.A. Dawah, B.A. Hawkins, and M.F. Claridge. Structure of the parasitoid communities of grass-feeding chalcid wasps. *Journal of animal ecology*, 64(6):708–720, 1995.
- [7] M. Girvan and MEJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.
- [8] R. Guimera and L.A.N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.
- [9] M.S. Handcock, A.E. Raftery, and J.M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society-Series A*, 170(2):301–354, 2007.
- [10] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.
- [11] R.R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44, 1990.
- [12] C. Kemp and J.B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687, 2008.
- [13] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 381. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [14] A.E. Krause, K.A. Frank, D.M. Mason, R.E. Ulanowicz, and W.W. Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, 2003.
- [15] V.E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.
- [16] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, page 584. ACM, 2006.
- [17] K.T. Miller, T.L. Griffiths, and M.I. Jordan. Nonparametric Latent Feature Models for Link Prediction. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

- [18] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658, 2004.
- [19] D.M. Roy, C. Kemp, V.K. Mansinghka, and J.B. Tenenbaum. Learning annotated hierarchies from relational data. *Advances in neural information processing systems*, 19:1185, 2007.
- [20] Y.W. Teh and D. Roy. The Mondrian Process. *Advances in neural information processing systems*, 2009.