Supplemental Material
Multiscale Community Blockmodel for Network Exploration

A Nested Chinese Restaurant Process

The nested Chinese Restaurant Process (nCRP) [1] 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:

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

where $\gamma > 0$ is a "concentration" parameter that controls the probability of drawing new integers, and for conciseness we define $c_{1:(i-1),1} \equiv (c_{11}, \ldots, c_{(i-1)1})$. 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 $i \in N$, we can sample $c_i \sim \mathrm{nCRP}(\gamma)$ using the recursive nCRP definition:

$$\mathrm{P}(c_{ik} = x \mid c_{1:(i-1)}, c_{i,1:(k-1)}) = \\ \begin{cases} \frac{\left|\{j<i \mid c_{j,1:(k-1)}=c_{i,1:(k-1)} \wedge c_{jk}=x\}\right|}{\left|\{j<i \mid c_{j,1:(k-1)}=c_{i,1:(k-1)}\}\right|+\gamma} & x \in \{c_{jk} \mid (j<i) \wedge c_{j,1:(k-1)} = c_{i,1:(k-1)}\} \\ \frac{\gamma}{\left|\{j<i \mid c_{j,1:(k-1)}=c_{i,1:(k-1)}\}\right|+\gamma} & x \text{ is the smallest positive integer not in the above set.} \end{cases} \tag{2}$$

B Stick Breaking Processes

Stick breaking constructions work as follows: Consider a stick of length 1. Draw $V_{i1} \sim Beta(m\pi, (1-m)\pi)$. 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 $V_{i2} \sim Beta(m\pi, (1-m)\pi)$ 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 kth stick that was chopped off. In general, we draw $V_{ik} \sim Beta(m\pi, (1-m)\pi)$ from $k = 1$ to $k = \infty$ and the corresponding $\{\theta_{ik}\}_{k=1}^{\infty}$ is defined below:

$$\theta_{ik} = V_{ik} \prod_{u=1}^{k-1} (1 - V_{iu}) \tag{3}$$

This process is known as the two-parameter GEM distribution [1] (although we refer to it as $\mathrm{Stick}(m, \pi)$) and draws from $\mathrm{Stick}(m, \pi)$ are denoted as $\theta_i \sim \mathrm{Stick}(m, \pi)$. $m > 0$ 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 $\mathrm{Stick}(m, \pi)$ 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).

C 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 $\mathbf{c}$.

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

$$\begin{aligned}
& \mathbb{P}(z_{\rightarrow ij} \mid \mathbf{c}, \mathbf{z}_{-(\rightarrow ij)}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
\propto\ & \mathbb{P}(E_{ij}, z_{\rightarrow ij} \mid \mathbf{c}, \mathbf{z}_{-(\rightarrow ij)}, \mathbf{E}_{-(ij)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
=\ & \mathbb{P}(E_{ij} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(ij)}, \gamma, m, \pi, \lambda_1, \lambda_2)\mathbb{P}(z_{\rightarrow ij} \mid \mathbf{c}, \mathbf{z}_{-(\leftarrow ij)}, \mathbf{E}_{-(ij)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
=\ & \mathbb{P}(E_{ij} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(ij)}, \lambda_1, \lambda_2)\mathbb{P}(z_{\rightarrow ij} \mid \mathbf{z}_{i,(-j)}, m, \pi)
\end{aligned} \tag{4}$$

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

$$\begin{aligned}
\text{First term} &= \begin{cases} \frac{\Gamma(a+b+\lambda_1+\lambda_2)}{\Gamma(a+\lambda_1)\Gamma(b+\lambda_2)} \cdot \frac{\Gamma(a+E_{ij}+\lambda_1)\Gamma(b+(1-E_{ij})+\lambda_2)}{\Gamma(a+b+1+\lambda_1+\lambda_2)} & \mathrm{S}_B^{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases} \\
a &= \left|\left\{(x,y) \mid (x,y) \neq (i,j), \mathrm{S}_B^{xy} = \mathrm{S}_B^{ij}, E_{xy} = 1\right\}\right| \\
b &= \left|\left\{(x,y) \mid (x,y) \neq (i,j), \mathrm{S}_B^{xy} = \mathrm{S}_B^{ij}, E_{xy} = 0\right\}\right|
\end{aligned} \tag{5}$$

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

$$\begin{aligned}
\mathbb{P}(z_{\rightarrow ij} = k \mid \mathbf{z}_{i,(-j)}, m, \pi) &= \mathbb{E}\left[\mathbb{I}(z_{\rightarrow ij} = k) \mid \mathbf{z}_{i,(-j)}, m, \pi\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(z_{\rightarrow ij} = k) \mid V_{i1}, ..., V_{ik}, \mathbf{z}_{i,(-j)}, m, \pi\right]\right] \\
&= \mathbb{E}\left[V_{ik} \prod_{u=1}^{k-1} (1 - V_{iu}) \mid \mathbf{z}_{i,(-j)}, m, \pi\right] \\
&= \mathbb{E}[V_{ik} \mid \mathbf{z}_{i,(-j)}, m, \pi] \prod_{u=1}^{k-1} \mathbb{E}[(1 - V_{iu}) \mid \mathbf{z}_{i,(-j)}, m, \pi] \\
&= \frac{m\pi + \#[\mathbf{z}_{i,(-j)} = k]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq k]} \prod_{u=1}^{k-1} \frac{(1-m)\pi + \#[\mathbf{z}_{i,(-j)} > u]}{\pi + \#[\mathbf{z}_{i,(-j)} \geq u]}
\end{aligned} \tag{6}$$

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.

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

$$\begin{aligned}
& \mathbb{P}(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
\propto\ & \mathbb{P}(c_i, \mathbf{E}_{(i\cdot),(\cdot i)} \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}_{-(i\cdot),-(\cdot i)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
=\ & \mathbb{P}(\mathbf{E}_{(i,\cdot),(\cdot,i)} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i\cdot),-(\cdot i)}, \gamma, m, \pi, \lambda_1, \lambda_2)\mathbb{P}(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{E}_{-(i\cdot),-(\cdot i)}, \gamma, m, \pi, \lambda_1, \lambda_2) \\
=\ & \mathbb{P}(\mathbf{E}_{(i\cdot),(\cdot i)} \mid \mathbf{c}, \mathbf{z}, \mathbf{E}_{-(i\cdot),-(\cdot i)}, \lambda_1, \lambda_2)\mathbb{P}(c_i \mid \mathbf{c}_{-i}, \gamma)
\end{aligned} \tag{7}$$

where $\mathbf{E}_{(i\cdot),(\cdot i)} = \{E_{xy} \mid x = i \text{ or } y = i\}$ 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

$$\begin{aligned}
&\mathrm{P}(c_{ik} = x \mid c_{1:(i-1)}, c_{i,1:(k-1)}, \gamma) = \\
&\begin{cases} \frac{\left|\{j<i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)} \wedge c_{jk} = x\}\right|}{\left|\{j<i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}\right| + \gamma} & x \in \{c_{jk} \mid (j<i) \wedge c_{j,1:(k-1)} = c_{i,1:(k-1)}\} \\ \frac{\gamma}{\left|\{j<i \mid c_{j,1:(k-1)} = c_{i,1:(k-1)}\}\right| + \gamma} & x \text{ is the smallest positive integer not in the above set.} \end{cases}
\end{aligned} \tag{8}$$

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

$$\begin{aligned}
\text{First term} &= \begin{cases} \prod_{B\in\mathbf{B}_{(i,\cdot),(\cdot,i)}} \frac{\Gamma(g_B+h_B+\lambda_1+\lambda_2)}{\Gamma(g_B+\lambda_1)\Gamma(h_B+\lambda_2)} \cdot \frac{\Gamma(g_B+r_B+\lambda_1)\Gamma(h_B+s_B+\lambda_2)}{\Gamma(g_B+h_B+r_B+s_B+\lambda_1+\lambda_2)} & \forall E_{xy}\in\mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy}\neq 0 \\ 0 & \text{otherwise} \end{cases} \\
g_B &= \left|\left\{(x,y)\mid E_{xy}\in\mathbf{E}_{-(i\cdot),-(\cdot i)},\ \mathrm{S}_B^{xy}=B, E_{xy}=1\right\}\right| \\
h_B &= \left|\left\{(x,y)\mid E_{xy}\in\mathbf{E}_{-(i\cdot),-(\cdot i)},\ \mathrm{S}_B^{xy}=B, E_{xy}=0\right\}\right| \\
r_B &= \left|\left\{(x,y)\mid E_{xy}\in\mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy}=B, E_{xy}=1\right\}\right| \\
s_B &= \left|\left\{(x,y)\mid E_{xy}\in\mathbf{E}_{(i\cdot),(\cdot i)},\ \mathrm{S}_B^{xy}=B, E_{xy}=0\right\}\right|
\end{aligned} \tag{9}$$

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

D Simulation Framework Details and Additional Experiments

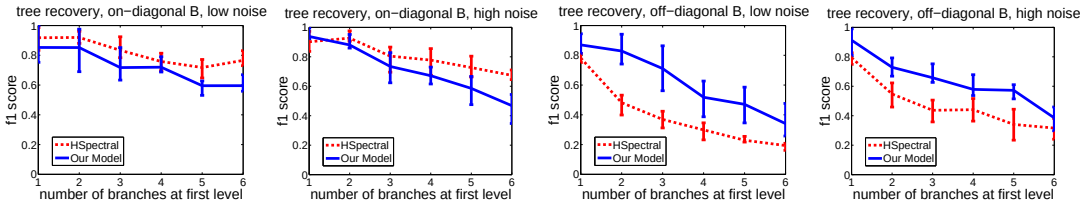
D.1 K=2 experiment details

For $K=2$, the 4 types of $\mathbf{B}$'s explored are:

1. **on-diagonal, low noise** - $B_{on-diagonal}=(.4,.7)$, $B_{off-diagonal}=(.02,.02)$;
2. **on-diagonal, high noise** - $B_{on-diagonal}=(.3,.6)$, $B_{off-diagonal}=(.1,.1)$;
3. **off-diagonal, low noise** - $B_{on-diagonal}=(.02,.02)$, $B_{off-diagonal}=(.4,.7)$;
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. $B_{off-diagonal}$ gives analogous interaction probabilities for *different* communities on the same level.

The experiment had $N=150$ actors and $\theta=(.25,.75)$ for all actors.



(a) (b) (c) (d)

Figure 1: $K=2$ experiments

The Gibbs sampler was run for 1,500 burn-in iterations on each experiment. 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. Note that spectral clustering was given the the number of first level clusters as an advantage (with binary splits in the deeper levels).

D.2 K=3 experiment results

For $K=3$, the 4 types of $\mathbf{B}$'s explored are

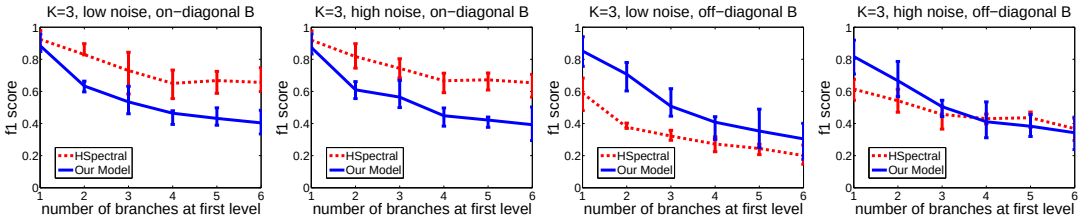
1. **on-diagonal, low noise** - $B_{on-diagonal}=(.5,.7,.9)$, $B_{off-diagonal}=(.02,.02,.02)$;
2. **on-diagonal, high noise** - $B_{on-diagonal}=(.5,.7,.9)$, $B_{off-diagonal}=(.2,.2,.2)$;

3. **off-diagonal, low noise** - $B_{on-diagonal} = (.02, .02, .02)$, $B_{off-diagonal} = (.5, .7, .9)$;

4. **off-diagonal, high noise** - $B_{on-diagonal} = (.2, .2, .2)$, $B_{off-diagonal} = (.5, .7, .9)$.

Similar to the $K = 2$ experiments, $B_{on-diagonal} = (a, b, c)$ means that actors interacting in the same level-1, 2 and 3 communities do so with probabilities a, b and c respectively. $B_{off-diagonal}$ gives analogous interaction probabilities for *different* communities on the same level.

The experiment had $N = 300$ actors and $\theta = (.15, .3, .55)$ for all actors.



(a) (b) (c) (d)

Figure 2: $K = 3$ experiments

The Gibbs sampler was run for 3,000 burn-in iterations on each experiment. As in the previous case, spectral clustering was given the the number of first level clusters as an advantage (with binary splits in the deeper levels). Note that spectral clustering actually performs better in the "high-noise off-diagonal" case than the "low noise off-diagonal" case, since the higher noise means more assortativity for the off-diagonal case.

E Qualitative analysis details

E.1 Grass Food Web Dataset

We ran our Gibbs sampler on the full network to infer the community hierarchy and actor Multiscale Memberships. The model parameters were chosen via gridsearch over $(\lambda_1, \lambda_2) \in \{.1, .3, .5, .7, .9\}^2$, according to the marginal log likelihood (estimated using 10,000 importance samples). In line with the held-out experiments, we fixed the remaining parameters to $\gamma = 1, m = 0.5, \pi = 0.5$. Finally, the hierarchy depth was set to $K = 2$. We ran our Gibbs sampler using 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. A plateauing log complete likelihood plot revealed that our sampler coverged well before the last iteration.

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 $z_{\rightarrow ij}$ and $z_{\leftarrow ij}$, 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.

E.2 High Energy Physics dataset

We applied the same parameter selection and post-processing as the previous dataset; the optimal gridsearch 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.

References

[1] 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.