Supplementary Material
The Nonparametric Metadata Dependent Relational Model

1. MCMC Inference Details

Here we outline the basic MCMC updates performed for all explicitly sampled (e.g. non-collapsed) variables: v (topic advications), η (block-specific metadata weights), s, r (source and receiver block assignments), and hyperparamters λ. Note that block relations matrix W is collapsed and not explicitly represented, as required by our retrospective MCMC technique.

1.1. Independence Sampler for Topic Activations V

The posterior of $v_{:i}$ does not have a closed analytical form due to the nonlinearity associated with the logistic mapping of our stick-breaking weights $\pi_{:i}$. We instead perform a Metropolis-Hastings update with an independence proposal $q(v^*_{:i} \mid v_{:i}, \eta, \phi_{:i}, \lambda_V) = N(v^*_{:i} \mid \eta^T \phi_{:i}, \lambda_V^{-1} I_K)$, which represents a sample from our prior. Since our proposal distributions are equal to the prior we need to evaluate in our acceptance ratio $\mathrm{A}(v^*_{:i}, v_{:i})$, these terms cancel and we are left with a simple ratio between our likelihood terms so that $\mathrm{A}(v^*_{:i}, v_{:i})$ is now:

$$= \prod_{j=1}^{N} \prod_{m=1}^{M} \frac{p(s_{ijm} \mid v^*_{:i}) p(r_{ijm} \mid v^*_{:i})}{p(s_{ijm} \mid v_{:i}) p(r_{ijm} \mid v_{:i})} = \prod_{k=1}^{K} \left(\frac{\pi^*_{ki}}{\pi_{ki}} \right)^{B_{ki}} \tag{1}$$

where we let B_{ki} denote the total number of both source and receiver indicators attached to node i assigned to community k. Formally, $B_{ki} = \sum_{m=1}^{M} \sum_{j=1}^{N} \delta(s_{ijm}, k) + \delta(r_{jim}, k)$. The proposal is then accepted with probability $\min(\mathrm{A}(v^*_{:i}, v_{:i}), 1)$.

1.2. Link-specific membership indicator variables s_{ijm}, r_{ijm}

Recall that for directed edge i to j in relation m, s_{ijm} denotes the community through which node i "sends" of the connection, and r_{ijm} indicates the community through which j "receives" this connection. We note that for fixed i, j, m, we can sample these assignments in an alternating fashion from conditional distributions, or in a blocked fashion from their joint distribution. We find the joint sampling to provide significant runtime speed up in implementation, although little seems gained in terms of mixing efficiency. Thus, we present the alternating sampling scheme here since its exposition is a bit clearer. We update our source community indicator variable s_{ijm} by marginalizing out the beta prior on $W_{::m}$ and fixing its complementary receiver variable r_{ijm} as well as values for π_i and π_j. In general form, we express the conditional posterior on s_{ijm} as

$$p(s_{ijm} | r_{ijm}, s_{-i,j,m}, r_{-i,j,m}, Y, \pi) \propto p(Y_{ijm} | s_{ijm}, r_{ijm}, s_{-i,j,m}, r_{-i,j,m}, Y_{-i,j,m}) p(s_{ijm} | \pi_i) \tag{2}$$

We use several sufficient statistics about observed edges to make this calculation possible. Recall that an edge i' to j' in relation m' is *present* if $Y_{i',j',m'} = 1$, and *absent* if $Y_{i',j',m'} = 0$. Missing/unobserved edges $Y_{i',j',m'} =?$ are ignored during inference.

Suppose $r_{ijm} = \ell$ and consider setting $s_{ijm} = k$. Let $A^{\setminus ij}_{k\ell m}$ be the total number of *present* edges with latent community pairs k, ℓ excluding Y_{ijm}. Furthermore, let $B^{\setminus ij}_{k\ell m}$ be the same for *absent* edges excluding Y_{ijm}. With these sufficient statistics, our posterior reduces to

$$p(s_{ijm} = k | r_{ijm} = \ell, \ldots) \propto \pi_{ki} \left(\frac{(A^{\setminus ij}_{k\ell m} + \gamma_a)^{y_{ijm}} (B^{\setminus ij}_{k\ell m} + \gamma_b)^{1 - y_{ijm}}}{(A^{\setminus ij}_{k\ell m} + B^{\setminus ij}_{k\ell m} + \gamma_a + \gamma_b)} \right) \tag{3}$$

Note that π_{ki} comes from the logistic mapping of our latent Gaussian variable $v_{:i}$ and that the marginalization of W will be required when we implement our retrospective MCMC technique to automatically determine the cardinality of K.

1.3. Block-Specific Metadata Weights η_k

To sample $\eta_{:k}$, the linear function relating metadata to topic k, we condition on all our features ϕ as well as λ_V, μ and Λ, where $\Lambda = \lambda_F I_F$. Columns of η are conditionally independent where the posterior $\eta_{:k}$ is now:

$$\propto N(\eta_{:k} \mid \mu, \Lambda^{-1}) N(v_{k:}^T \mid \phi^T \eta_{:k}, \lambda_V^{-1} I_N) \tag{4}$$

$$\propto N(\eta_{:k} \mid \Sigma_\eta(\lambda_V \phi v_{k:}^T + \Lambda\mu),\ \Sigma_\eta) \tag{5}$$

where $\Sigma_\eta = (\Lambda + \lambda_V \phi\phi^T)^{-1}$, the posterior covariance for $\eta_{:k}$. Conditioning on the feature regression weights η, the mean weight μ_f in our hierarchical prior for each feature f has a Gaussian posterior:

$$\propto N(\mu_f \mid 0, \lambda_S^{-1}) \prod_{k=1}^{K} N(\eta_{fk} \mid \mu_f, \lambda_F^{-1}) \tag{6}$$

$$\propto N(\mu_f \mid \frac{\sum_{k=1}^{K} \eta_{fk}}{K + \lambda_F^{-1}}, (\lambda_S + k\lambda_F)^{-1}) \tag{7}$$

1.4. Hyperparameters

We place a gamma prior to control the variability of the mean associated with η. The gamma prior that we place on λ_S is conjugate so that:

$$\propto \text{Gam}(\lambda_S \mid a_S, b_S) \prod_{f=1}^{F} N(\mu_f \mid 0, \lambda_S^{-1}) \tag{8}$$

$$\propto \text{Gam}(\lambda_S \mid \frac{1}{2}F + a_S, \frac{1}{2}\sum_{f=1}^{F} \mu_f^2 + b_S) \tag{9}$$

Similarly, the precision parameter λ_F comes from a gamma prior and controls the variability of the feature weights associate with each topic.

$$\propto \text{Gam}(\lambda_F \mid a_F, b_F) \prod_{k=1}^{K} \prod_{f=1}^{F} N(\eta_{fk} \mid \mu_f, \lambda_F^{-1}) \tag{10}$$

$$\propto \text{Gam}(\lambda_F \mid \frac{1}{2}KF + a_F, \frac{1}{2}\sum_{k=1}^{K}\sum_{f=1}^{F} (\eta_{fk} - \mu_f)^2 + b_F) \tag{11}$$

Similarly, the precision parameter λ_V has a gamma prior and posterior where $L = \lambda_V I$:

$$\propto \text{Gam}(\lambda_V \mid a_V, b_V) \prod_{i=1}^{N} N(v_{:i} \mid \eta^T \phi_{:i}, L^{-1}) \tag{12}$$

$$\propto \text{Gam}(\lambda_V \mid \frac{1}{2}KN + a_V, \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N} (v_{ki} - \eta_{:k}^T \phi_{:i})^2 + b_V) \tag{13}$$

2. Link Prediction Experiments

After training the model on the observed edges (e.g. both *present* and *absent*), we wish to predict values for missing edges (those marked $Y_{ijm} =?$). Our MCMC inference provides samples for π, S, R and K (cardinality of the community set) at each iteration. After running the sampler for sufficiently many iterations, we discard the first half of samples as burn in and train on the remaining samples. Given T samples indexed by t for each hidden variable, we predict the value Y^*_{ij} of the link between i and j given observed edges Y for a particular relation as follows

$$\mathbb{E}(Y^*_{i,j}|Y, \{\pi, S, R\}_{t=1}^T) = \frac{1}{T}\sum_{t=1}^{T} f(Y^*_{i,j}|\pi_t, S_t, R_t, K_t, Y) \tag{14}$$

$$f(Y^*_{i,j}|\pi, S, R, K, Y) = \sum_{k=1}^{K}\sum_{\ell=1}^{L} p(s_{i,j} = k)p(r_{i,j} = l)p(Y^*_{i,j} = 1|Y, s_{i,j} = k, r_{i,j} = l) \tag{15}$$

$$= \sum_{k=1}^{K}\sum_{\ell=1}^{L} \pi_{i,k}\pi_{j,\ell}\frac{A_{k\ell} + \gamma_a}{A_{k\ell} + B_{k\ell} + \gamma_a + \gamma_b} \tag{16}$$

where A and B are sufficent statistic counts over the observed edges Y defined above.

3. Computational Complexity

Here, we compare the per-iteration cost of inference for MMSB and NMDR models with K active communities. For both models, the cost of resampling the s, r assignments for all edges is comparable, scaling as $\mathcal{O}(KN^2)$. NMDR's non-conjugate community membership prior requires an additional $\mathcal{O}(KN)$ operations to resample the node-specific community weights v; in practice this dominates updates to η and other parameters. This additional cost scales linearly with the network size, is parallelizable, and allows the NMDR to capture metadata and avoid model selection issues. Our non-optimized Matlab implementation can be applied, with reasonable computational time, to networks containing a few thousand nodes.

4. Generating Metadata Graphs

Let $\tilde{\phi}_{:i}$ be an $Fx1$ feature vector representing one possible organism indexed i. We then generate $T = 1000$ samples of $v_{:it} \sim N(\eta^T\tilde{\phi}_{:i}, \lambda_V^{-1}I)$ where η and λ_V are samples from the last iteration of our MCMC chain. We then obtain an estimate for $\tilde{\pi}_{:i} = \frac{1}{T}\sum_{t=1}^{T}\pi_{:it}$ where $\pi_{:it}$ is calculated by the deterministic function in (1) which determines the logistic mapping of $v_{:it}$ into its stick breaking weights. We repeat this task for all organisms (42 for the figure we show) to generate a set of $\tilde{\pi}$ vectors. We then generate potential links between these metadata derived organism types so that a link $x_{ijt} \sim \text{Bern}(\hat{\pi}_{:it}^T\hat{W}_{::t}\hat{\pi}_{:jt})$ where $\hat{W}_{k\ell t} \sim \text{Beta}(A_{k\ell}+\gamma_A, B_{k\ell}+\gamma_b)$ is a sample of our stochastic block matrix generated from the last sample of our MCMC chain. Finally, we take a Monte Carlo estimation for this edge so that $\tilde{x}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{ijt}$, which represents the likelihood of a predator prey relationship between predator type i and prey type j.

5. Lazega Lawyers Experiments

For all the NMDR link prediction tasks, we ran our models for 10000 MCMC samples across two chains for each experiment type. The NMDR model used a sequential initialization scheme that incorporates our likelihood terms to find a useful configuration for our community assignments s and r. We then chose the model that resulted in the highest log probability for Y between 2 chains and utilized the last 1000 MCMC samples to calculate $\mathbb{E}(Y)$ as described in section 2 of the supplementary material.