Our Contributions

- Nonparametric mixed membership
  - Unbounded number of communities
  - Retrospective MCMC: No Truncation
- Metadata informs latent structure
  - Upstream inclusion of metadata leads to recovery of interpretable communities

Stochastic Block Models

Unsupervised community discovery [Wang JASA 1987] from observed network edges

- Assign each node to one latent block/community
- Predict edge presence from block assignments of source and receiver nodes

Infinite Relational (IRM)

- Unbounded number of blocks $K$, [Kemp AAI 2006] via Chinese Restaurant Process
- Each node assigned to one block

Mixed Membership (MMSB)

- Finite number of blocks $K$, [Airoldi JMLR 2008] must be specified a priori
- Each node has distribution over blocks

Using Metadata

**Downstream**

Metadata regression in edge likelihood

$$y_{ij} \sim \text{Bern}(W_{s_i, r_j} + \eta^T \phi_i)$$

Recovered communities less interpretable, just explain residual noise

**Upstream**

Metadata informs node membership, creates meaningful communities

Logistic stick breaking allows $\phi_i$ to inform $\pi_i$

$$v_{ki} \sim N(\eta_k^{T} \phi_i, \Lambda^{-1})$$

$$\pi_i = \psi(v_{ki})$$

Link presence probability

$$W_{kl} \sim \text{Beta}(\gamma_a, \gamma_b)$$

Mean metadata weight

$$\mu_f \sim N(0, \Lambda^{-1})$$

Metadata regression weight

$$\eta_{jk} \sim N(\mu_f, \Lambda^{-1})$$

Precision parameters have Gamma priors

$$\lambda_a, \lambda_f, \lambda_b$$

NMDR Graphical Model

Gibbs Sampler

Retrospective MCMC

MH (Independence)

Marginalized

$$\rho_k \propto \pi_k$$

$$\rho_k+1 \propto 1 - \sum_{l=1}^{K} \pi_k$$

$$s_{ij} \sim \text{Mult}(\rho)$$

Sample new params from prior

$$\eta_i, s_{ij} \sim \text{Mult}(\rho)$$

Draw termination indicator

$$\omega \sim \text{Bern}(\psi(v_{s_{ij}}))$$

$$s_{ij} \leftarrow s_{ij} + 1$$

$$\omega = 1$$

$$\text{DONE}$$

Retrospective MCMC

- Dynamically add/delete blocks across iterations
- Only instantiate $K$ blocks actively used
- Create parameters for new blocks as needed

**EXAMPLE:** Sampling $s_{ij} | \pi_i, r_i, r_j = \ell$

Typical Gibbs update, with extra term for new block

Counts of present/absent edges with communities $k,l$ excluding edge $ij$

$$\rho_k \propto \pi_k$$

$$\rho_k+1 \propto 1 - \sum_{l=1}^{K} \pi_k$$

$$s_{ij} \sim \text{Mult}(\rho)$$

remaining “stick” mass marginal likelihood for empty block

Sampson Monastery Analysis

[Miller NIPS 2009]

Lazega Lawyers Analysis

[Stover JASA 2007]

Mixed NMDR

- works well with few nodes
- works well with few blocks

Nonparametric NMDR

- works well with many nodes
- works well with many blocks

MMSB

- works well with few nodes
- works well with few blocks

NMDR

- works well with many nodes
- works well with many blocks

Source Block Assignment

$$s_{ij} \sim \text{Cat}(\pi_i)$$

Receiver Block Assignment

$$r_{ij} \sim \text{Cat}(\pi_j)$$

Binary Edge Indicator

$$y_{ij} \sim \text{Bern}(W_{s_i, r_j})$$

The Nonparametric Metadata Dependent Relational Model

Dae Il Kim, Michael C. Hughes, Erik B. Sudderth  Brown University, Providence, RI, USA