Thinned random measures for sparse graphs with overlapping communities

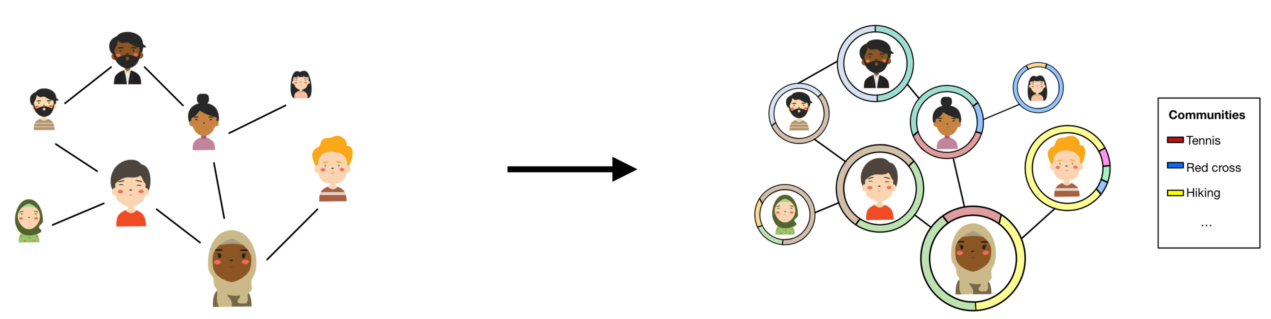
Federica Zoe Ricci
University of California Irvine
fzricci@uci.edu

Michele Guindani
University of California Los Angeles

Erik Sudderth
University of California Irvine

1. Introduction

Motivation:
develop **probabilistic model** for **network data** to discover **latent communities** where nodes interact



Background:

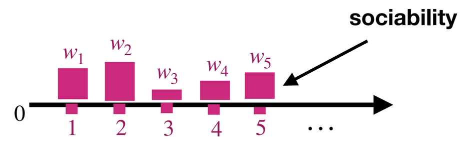
Stochastic block models (Wang et al. 1987)	Dense	No degree heterogeneity	Single community membership
Mixed membership stochastic block models (Airoldi et al. 2008)	Dense	No degree heterogeneity	**Mixed community membership**
Models based on completely random measures (Caron and Fox 2017)	**Sparse**	**Degree heterogeneity**	No community membership

Contributions of proposed model:

Models based on thinned completely random measures (proposed model)	**Sparse** # edges $=\Theta(N^q)$, $1<q<2$	**Degree heterogeneity**	**Mixed community membership**, **learn # communities**

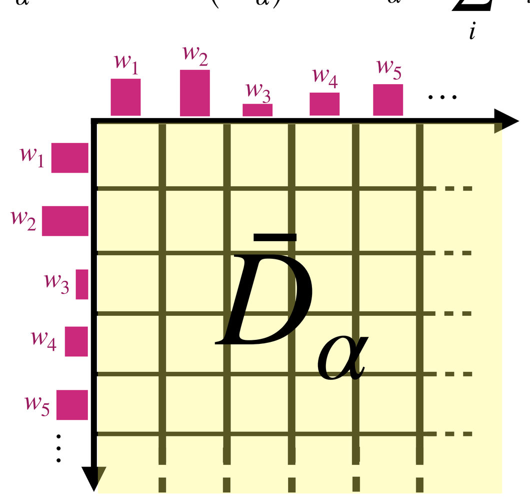
2. Nodes' sociabilities and community memberships

2.1. Draw ***nodes sociabilities*** with the Generalized Gamma Process (as Caron and Fox (2017))



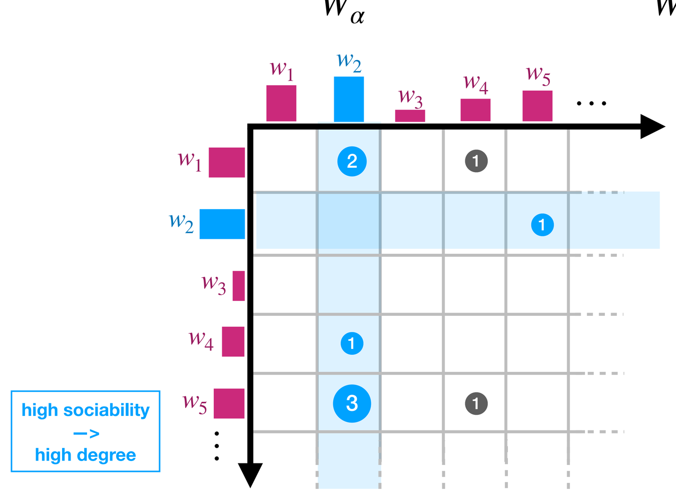
- Draw ***total number of potential edges:***

$$\bar{D}_\alpha \sim \text{Poisson}(\bar{W}_\alpha^2) \qquad \bar{W}_\alpha = \sum_i w_i$$



- Assign edges to **node pairs** (x_{e1}, x_{e2}):

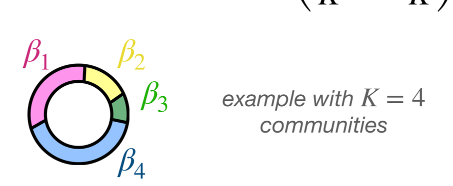
$$P(x_{e1}=i)=\frac{w_i}{\bar{W}_\alpha}, \qquad P(x_{e2}=j)=\frac{w_j}{\bar{W}_\alpha}$$



2.2 Draw nodes ***memberships*** to K ***communities*** (as Airoldi et al. (2008))

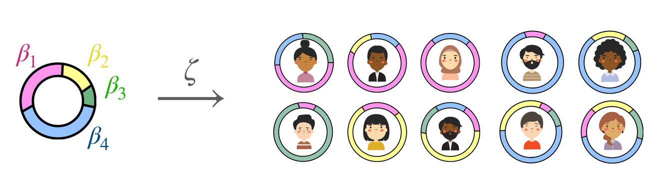
- Draw ***global frequency*** of communities:

$$(\beta_1,\ldots,\beta_K)\sim \text{Dirichlet}\left(\frac{\gamma}{K},\ldots,\frac{\gamma}{K}\right)$$



- Assign ***community memberships*** to nodes:

$$\pi_i=(\pi_{i1},\ldots,\pi_{iK})\mid\beta \overset{\text{ind}}{\sim} \text{Dirichlet}(\zeta\beta_1,\ldots,\zeta\beta_K)$$



Details:

- Learn # communities K by setting:
  - K upper bound to K_{true}
  - $\gamma < K$
- Properties:
  - Approximates Hierarchical Dirichlet Process as $K\to\infty$
  - Encourages learn sparse β
- Set ζ small for heterogeneous memberships
- Nodes are sampled with a 2d point process:

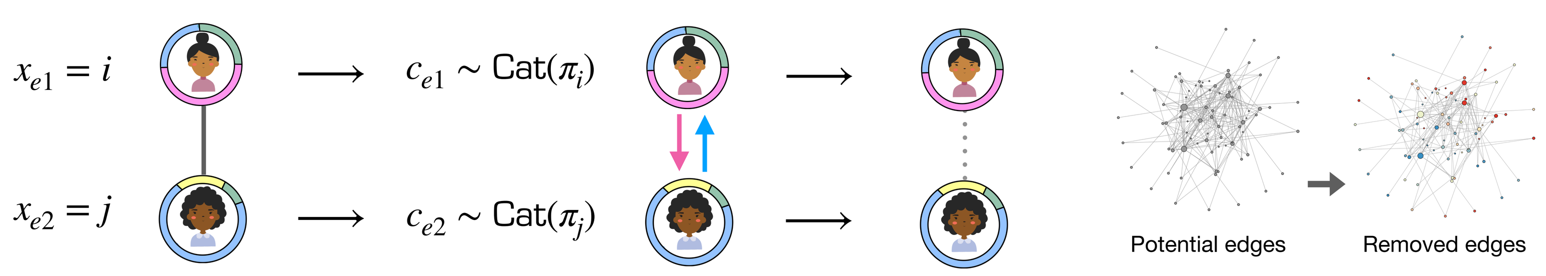
$$W_\alpha=\{(w_i,l_i):\{w_i\}\sim \text{GGP}(\sigma,\tau),\{l_i\}\sim\lambda(dl),l_i<\alpha\}$$

$\sigma\to$ network sparsity
$\tau\to$ decay of degree distribution
$\alpha\to$ network size

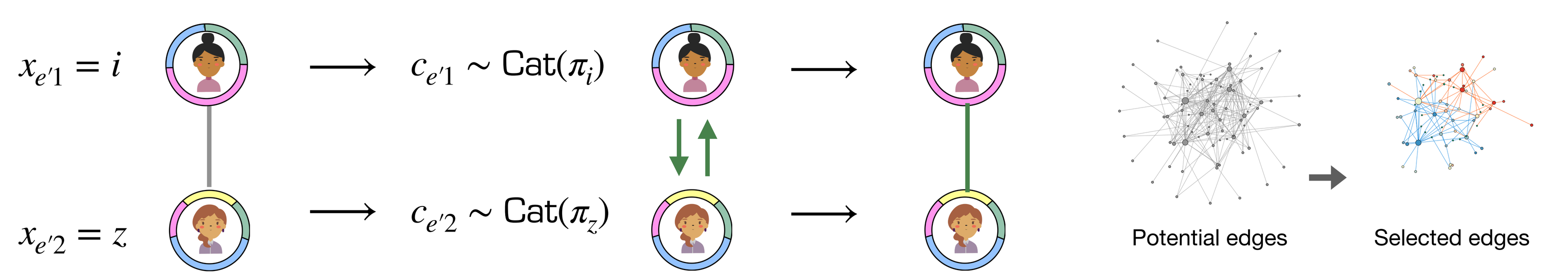
3. Thinning

3.1 **For each edge**, assign nodes to communities:

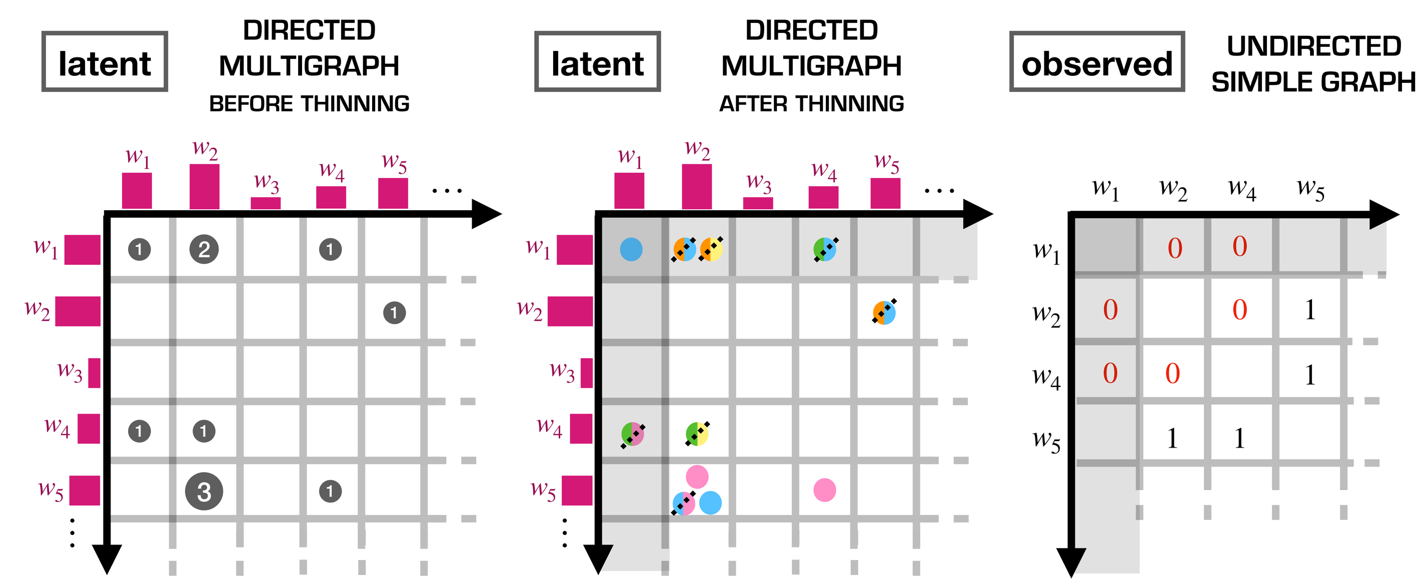
- **Thin** (remove) edges when nodes are assigned to **different communities**:



- **Keep** edges when nodes are assigned to **the same communities**:

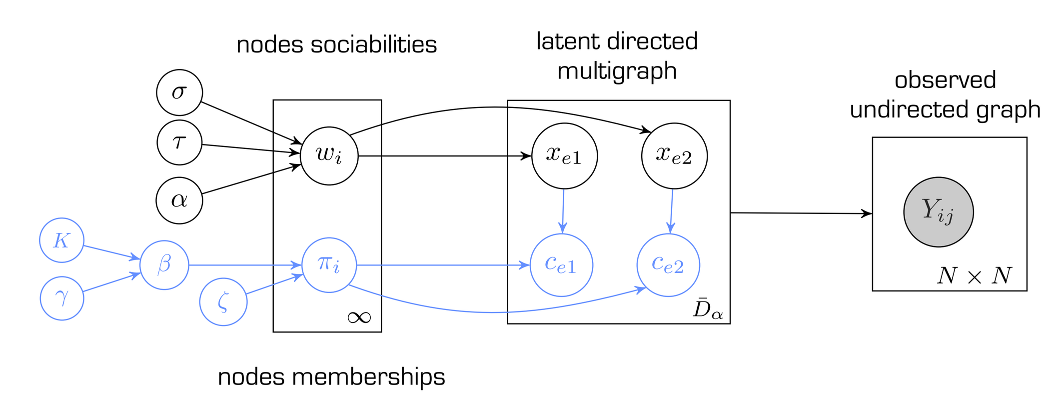


3.2 Transform latent directed multigraph to observed undirected graph:

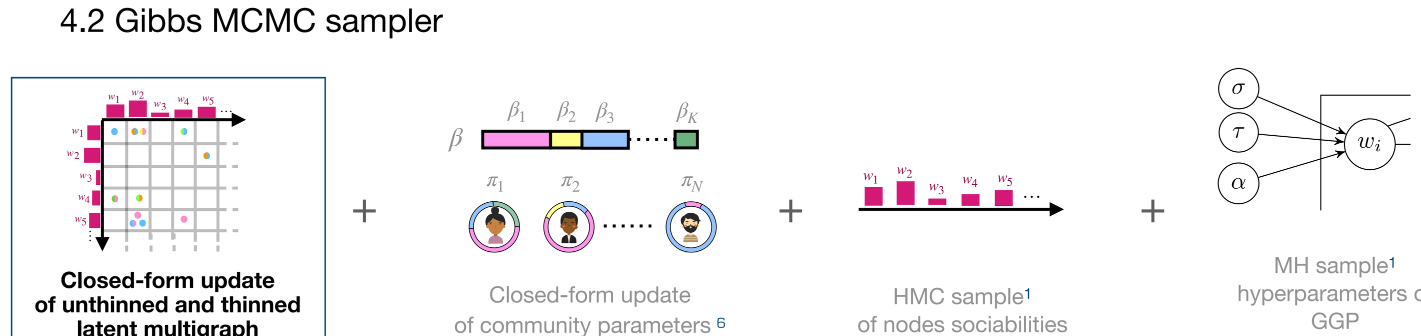


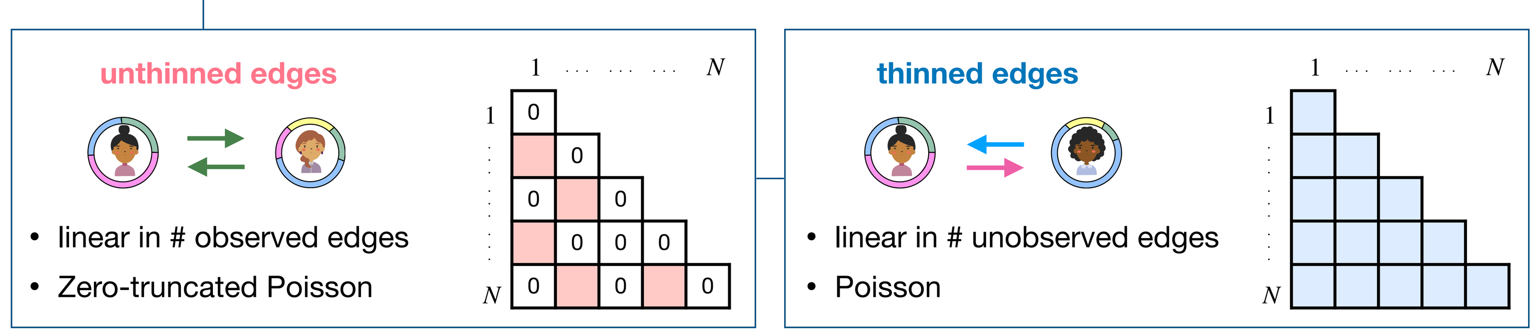
4. Posterior inference

4.1 Graphical representation of proposed model



4.2 Gibbs MCMC sampler





unthinned edges:
- linear in # observed edges
- Zero-truncated Poisson

thinned edges:
- linear in # unobserved edges
- Poisson

5. Related models

Sparse block model (Herlau et. al[2])	**Sparse**	**Degree heterogeneity**	Single community membership
Sparse mixed membership (Todeschini et. al[3])	**Sparse**	**Degree heterogeneity**	Multiple community sociabilities (non-regularized), ~~Learn # communities~~

Compound Generalized Gamma Process (CGGP)

Node i has K sociabilities: $(w_{i1},\ldots,w_{iK})$

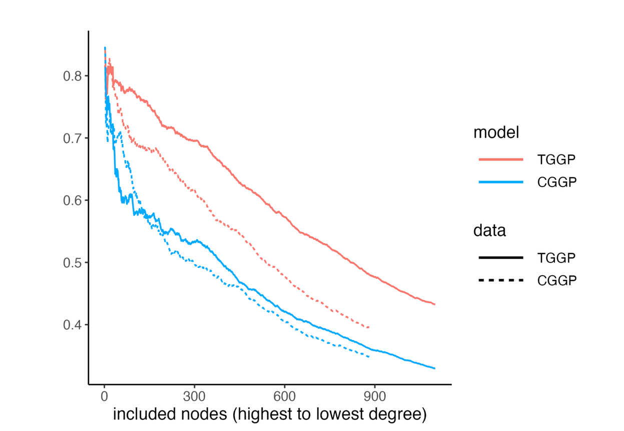
$$w_{ik}=w_{i0}*\eta_{ik}\quad\begin{cases}\text{Base sociability: } \{w_{i0}\}\sim\text{GGP}(\sigma,\tau)\\ \text{Community multiplier: } \eta_{ik}\overset{\text{ind}}{\sim}\text{Gamma}(a_k,b_k)\end{cases}$$

Thinned Generalized Gamma Process (TGGP - proposed model)

Node i has:

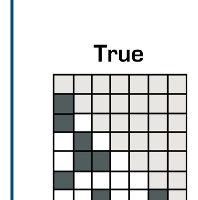
- one sociability: $\{w_i\}\sim\text{GGP}(\sigma,\tau)$
- a vector of community memberships (summing to one): $\pi_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\zeta\beta_1,\ldots,\zeta\beta_K)$

Similarity of true and estimated memberships



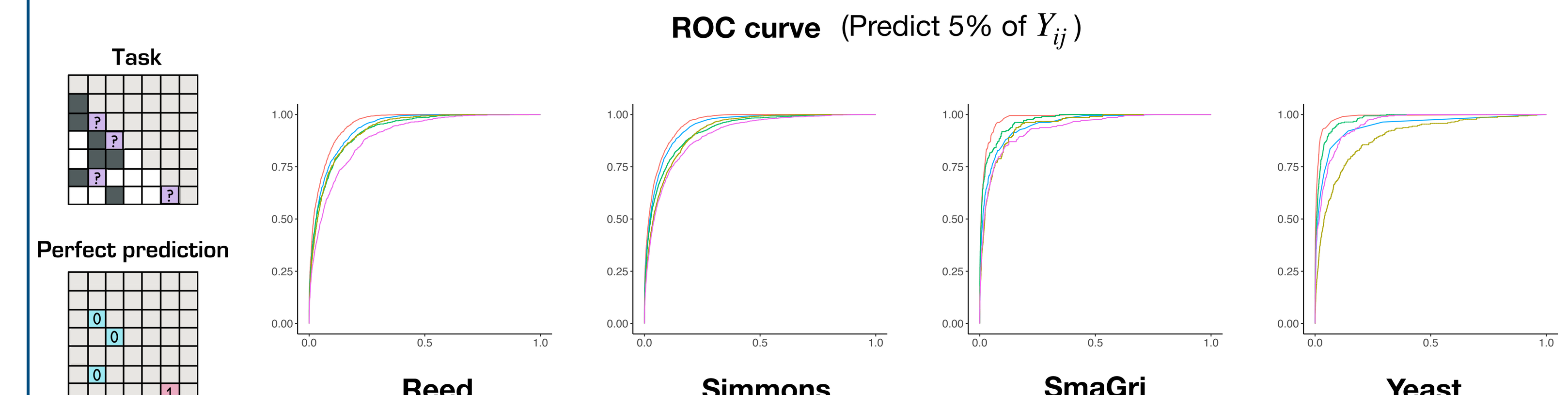
6. Results

Real-world data: posterior predictive accuracy

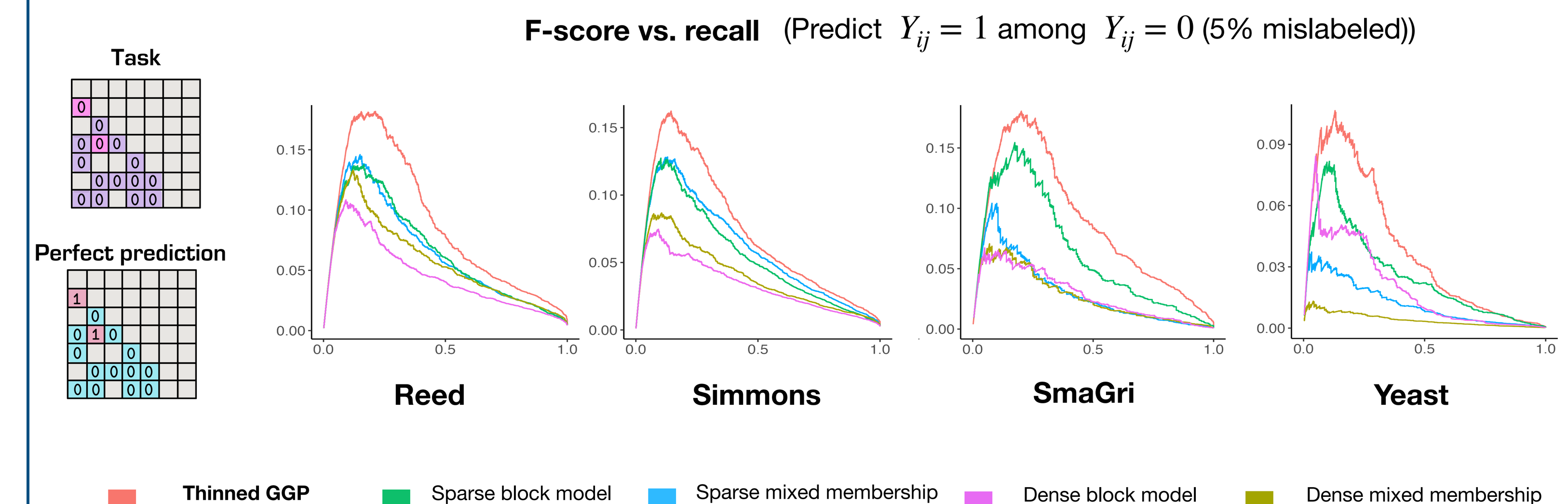


1. Fit model on fully observed data
2. Learn node-specific interaction parameters (e.g. nodes sociabilities and community memberships)
3. Use node-specific interaction parameters to predict edges (two prediction tasks)

ROC curve (Predict 5% of Y_{ij})



F-score vs. recall (Predict $Y_{ij}=1$ among $Y_{ij}=0$ (5% mislabeled))



References

1. Caron, François, and Emily B. Fox. "Sparse graphs using exchangeable random measures." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79, no. 5 (2017): 1295-1366.
2. Herlau, Tue, Mikkel N. Schmidt, and Morten Mørup. "Completely random measures for modelling block-structured sparse networks." Advances in Neural Information Processing Systems 29 (2016)
3. Todeschini, Adrien, Xenia Miscouridou, and François Caron. "Exchangeable random measures for sparse and modular graphs with overlapping communities." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82, no. 2 (2020): 487-520.
4. Wang, Yuchung J., and George Y. Wong. "Stochastic blockmodels for directed graphs." Journal of the American Statistical Association 82, no. 397 (1987): 8-19.
5. Airoldi, Edo M., David Blei, Stephen Fienberg, and Eric Xing. "Mixed membership stochastic blockmodels." Advances in neural information processing systems 21 (2008).
6. Yee Whye Teh, Michael I Jordan, Matthew J Beal & David M Blei. "Hierarchical Dirichlet Processes." Journal of the American Statistical Association 101, no. 476 (2006): 1566-1581.