Scalable MCMC in Degree Corrected Stochastic Block Model

Soumyasundar Pal, Mark Coates
Department of Electrical and Computer Engineering, McGill University

Introduction

- **Community detection** from graphs has many applications in analyzing collaboration networks, protein interaction, and social networks.
- **Community** : dense internal and sparse external connections
- **Earlier approaches** : hierarchical clustering, modularity optimization, spectral clustering, clique percolation etc.
  - Heuristic objective functions
  - Greedy optimization techniques
- **Principled approach** : statistical modelling of community structures
- **Scalable Bayesian inference** using Stochastic Gradient MCMC (SG-MCMC) schemes

We propose a version of a degree corrected stochastic block model and present an MCMC based inference algorithm.

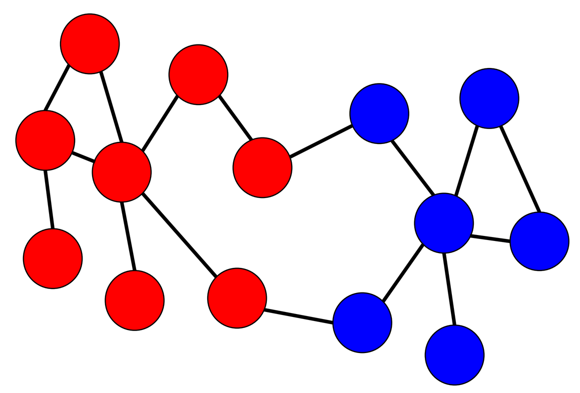
Background

In a stochastic block model, the probability of a link between any two nodes depends on their community memberships.

N : no. nodes, K : no. communities

Stochastic Blockmodel

- $c_i \in \{1, 2, ..., K\}$: membership of node i
- $y_{ab} \in \{0, 1\}$: (a, b)'th entry of the adjacency matrix
- $\beta_{k\ell} \in (0, 1)$: link probability between two nodes in community k and ℓ
- $p(y_{ab} = 1 | c_a = k, c_b = \ell) = \beta_{k\ell}$



A graph with two communities

Overlapping communities [1]

π_{ak} : probability that node a belongs to community k, $\sum_{k=1}^{K} \pi_{ak} = 1$

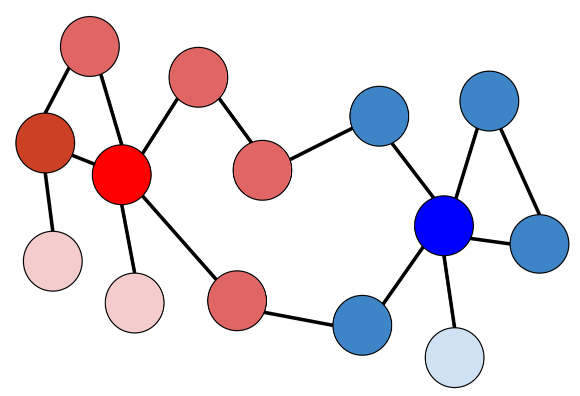
Mixed Membership Stochastic Blockmodel (MMSB)

step 1 : sample $z_{ab} \sim \pi_a$ and $z_{ba} \sim \pi_b$
step 2 : sample $y_{ab} | (z_{ab} = k, z_{ba} = \ell) \sim Bernoulli(\beta_{k\ell})$

- **assortative MMSB (a-MMSB) [2]** : $\beta_{k\ell} = \delta$ for $k \neq \ell$
- State-of-the-art **Bayesian inference** of $p(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{Y})$ is achieved [2] using **Stochastic Gradient Riemannian Langevin Dynamics (SGRLD)** algorithm.

Degree Corrected Blockmodel (DCB)

- Many networks show **heavy tailed degree distributions**.
- **Degree heterogeneity within community** is modelled by considering **the dependence of link probability on incident nodes**.



Non-uniformity of degree within communities

Mixed Membership DCB (MMDCB)

Parameters :

- **Community membership distribution** $\boldsymbol{\pi}_a$ for each node a
- **Degree correction parameter** $r_a \in \mathbb{R}$ for each node a
- **Community specific parameters** $q_k > 0$ for each community k

Prior distributions :

$\pi_a \sim Dir(\alpha)$, $r_a \sim \mathcal{N}(0, \sigma^2)$ and $q_k \sim \mathcal{N}(0, \sigma^2)\mathbf{1}_{\{q_k > 0\}}$

Generative model :

for any two nodes a and b :
  sample $z_{ab} \sim \pi_a$ and $z_{ba} \sim \pi_b$
  if $z_{ab} = z_{ba} = k$:
    sample $y_{ab} \sim Bernoulli(logit^{-1}(q_k + r_a + r_b))$
  else :
    sample $y_{ab} \sim Bernoulli(logit^{-1}(r_a + r_b))$

Posterior distribution :

$$p(\boldsymbol{\pi}, \mathbf{q}, \mathbf{r} | \mathbf{Y}) \propto p(\boldsymbol{\pi}) p(\mathbf{q}) p(\mathbf{r}) p(\mathbf{Y} | \boldsymbol{\pi}, \mathbf{q}, \mathbf{r}) ,$$
$$= \prod_{a=1}^{N} p(\boldsymbol{\pi}_a) p(r_a) \prod_{k=1}^{K} p(q_k)$$
$$\prod_{1 \leq a < b \leq N} \sum_{z_{ab}, z_{ba}} p(y_{ab}, z_{ab}, z_{ba} | \pi_a, \pi_b, q_{1:K}, r_a, r_b)$$

Inference :

- Analytically intractable $\Rightarrow$ approximation is required
- We design an **MCMC** scheme based on **RLD** to sample from the joint posterior distribution.
- **Computational complexity** per sample : $\mathcal{O}(N^2 K)$ for any gradient based MCMC algorithm, **does not scale well** to large graphs
- **Stochastic Gradient RLD (SGRLD) [3]** provides a **trade-off** between accuracy and complexity.

Experimental Setup

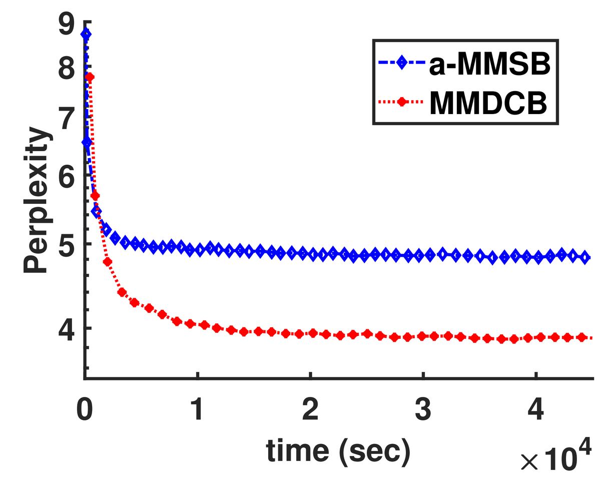
- Evaluation on 4 **academic collaboration networks**

	NETSCIENCE	RELATIVITY	HEP-TH	HEP-PH
Nodes	1589	5242	9877	12008
Edges	2742	14996	25998	118521

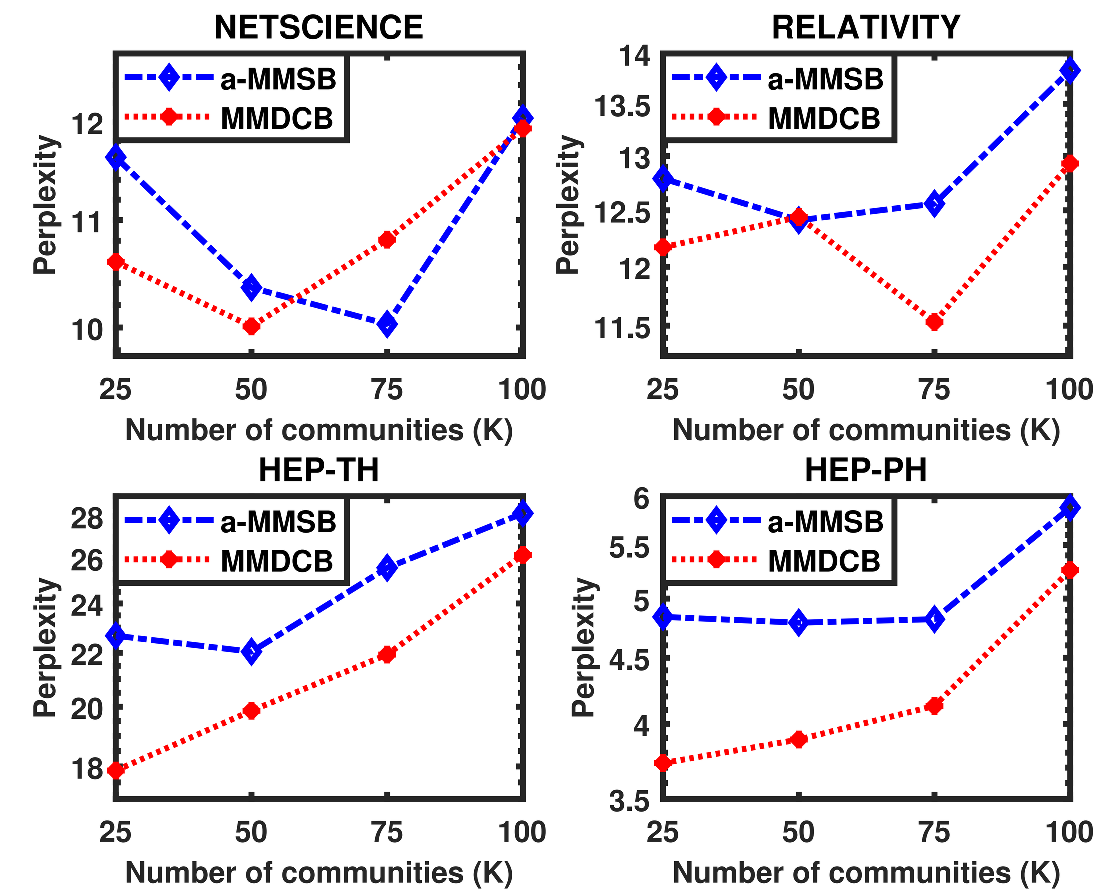
- **Node** : researcher, **edge** : collaboration
- **Held-out test set** : 10% of the links, same number of non-links
- **No. communities** (K) : $25, 50, 75$ and 100
- Comparison with **the SGRLD algorithm on the a-MMSB [2]**.
- **Predictive performance** is measured by **average perplexity**
- **High predictive likelihood** for test set $\Rightarrow$ **low average perplexity**
- Performance metric for **link prediction** : area under the receiver operating characteristic (ROC) curve (**AUC**)

Results

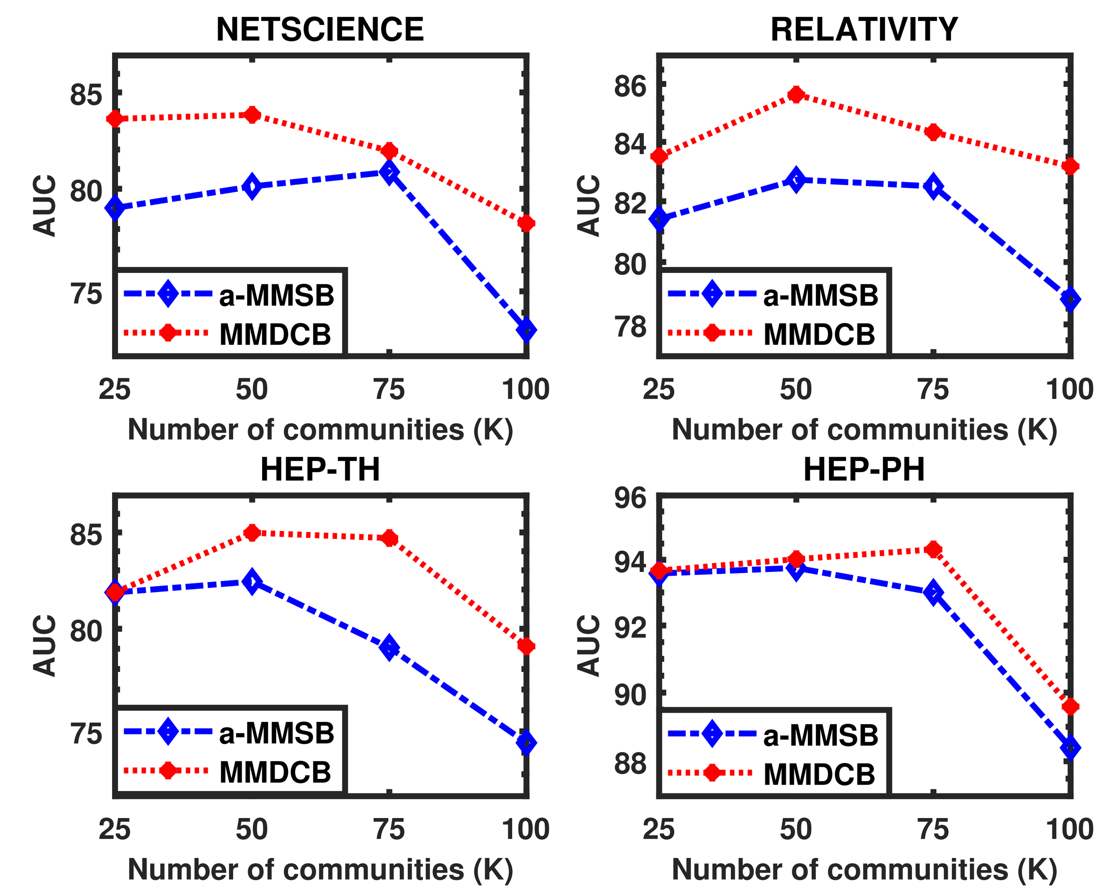
Convergence of perplexity for the HEP-PH dataset with $K = 50$



The MMDCB achieves lower perplexity compared to the a-MMSB



The MMDCB obtains higher AUC compared to the a-MMSB



Conclusion

- MMDCB models the networks better than a-MMSB.
- SG-MCMC algorithms scale well to large networks.
- **Future work** : better graph models, advanced SG-MCMC schemes

References

[1] Airoldi, Edoardo M. et al. 2008. Mixed membership stochastic blockmodels. In *JMLR*.
[2] Li, Wenzhe et al. 2016. Scalable MCMC for mixed membership stochastic blockmodels. In *Proc. AISTATS*.
[3] Welling, Max and Teh, Yee W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. ICML*.