Sampling and Inference for Beta Neutral-to-the-Left Models of Sparse Networks

Benjamin Bloem-Reddy, Adam Foster, Emile Mathieu, Yee Whye Teh

Department of Statistics, University of Oxford

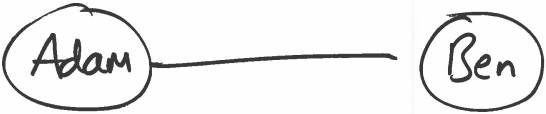
Contents

Temporal networks

Example

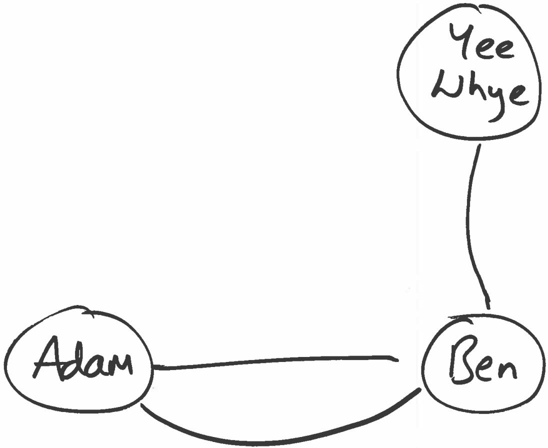
- Messages sent between people over time
- Protein-protein interactions

Temporal networks

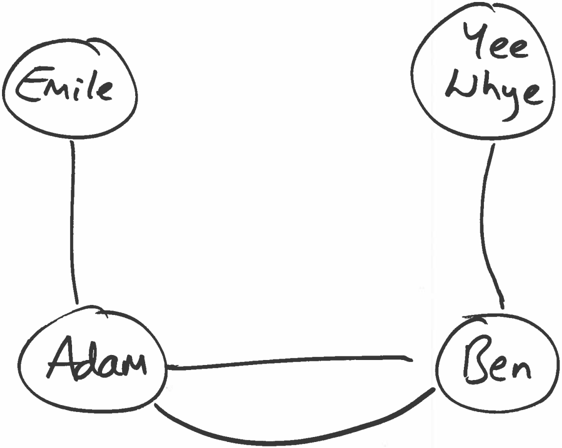


Temporal networks

Temporal networks

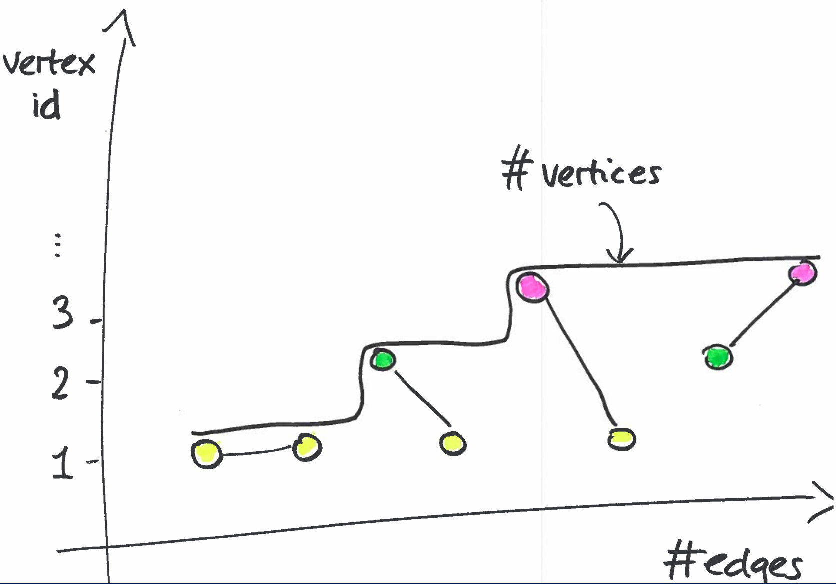


Temporal networks

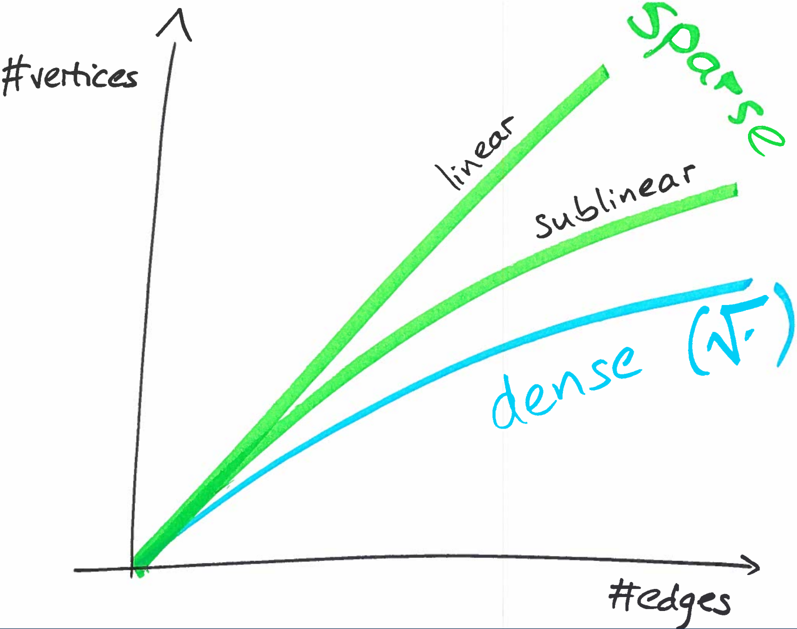


Edges and vertices

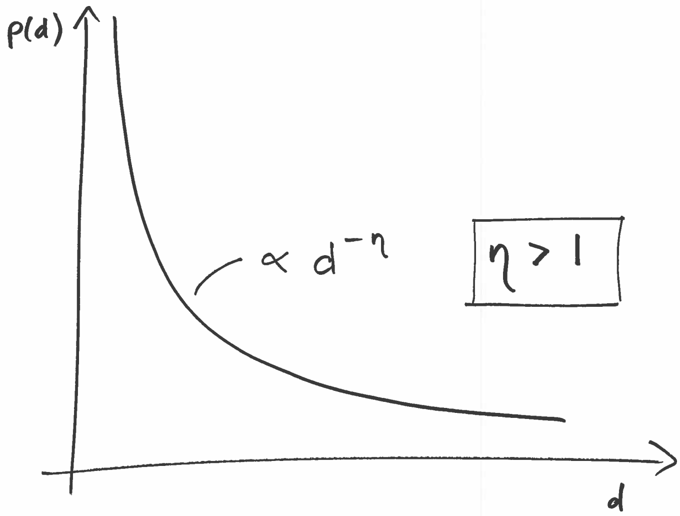
TODO: break picture into 4



Sparsity



Power law degree distribution



Sparsity and power law

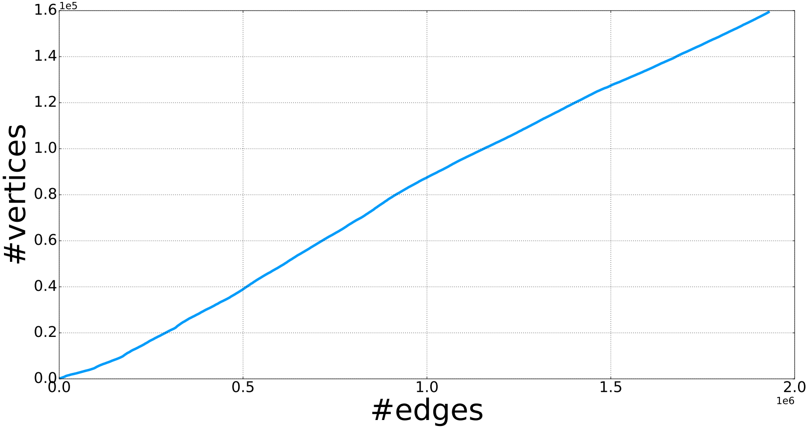
$$\textbf{Sublinear sparsity} \iff \eta \in (1,2)$$

$$\textbf{Linear sparsity} \iff \eta > 2$$

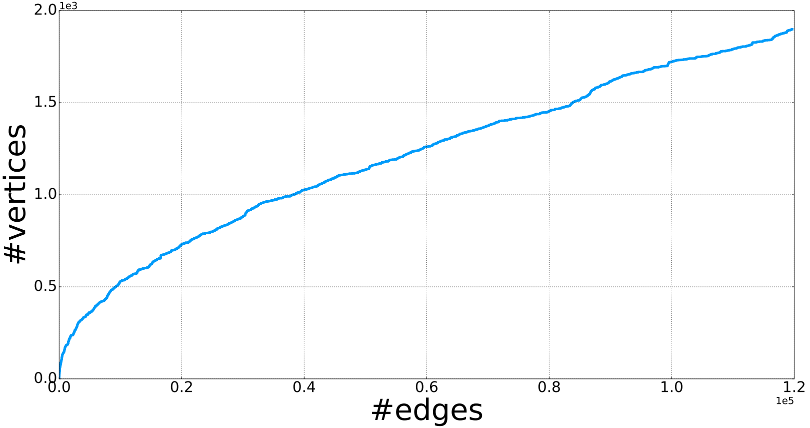
Empirical study

SNAP dataset [2]	# of vertices	# of edges
Ask Ubuntu	159,316	964,437
UCI social network	1,899	20,296
⋮	⋮	⋮

Ask Ubuntu



UCI social network



Models

Vertex
exchangeable

Models

dense | sparse

Vertex exchangeable

Models

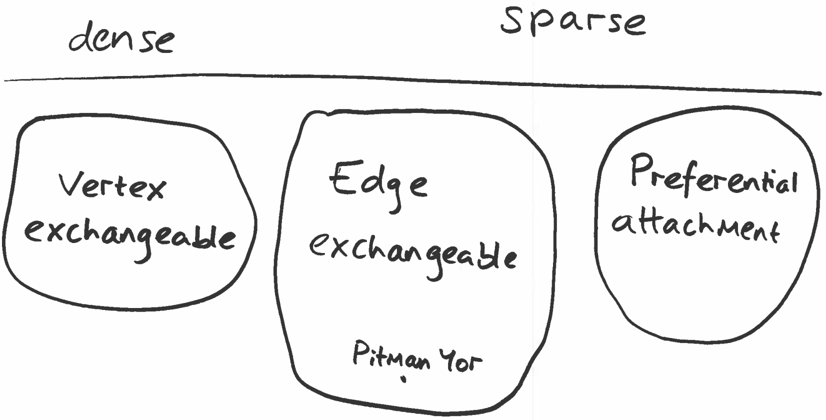
dense

sparse

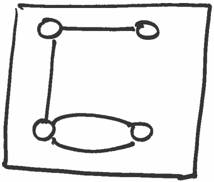
Vertex exchangeable

Preferential attachment

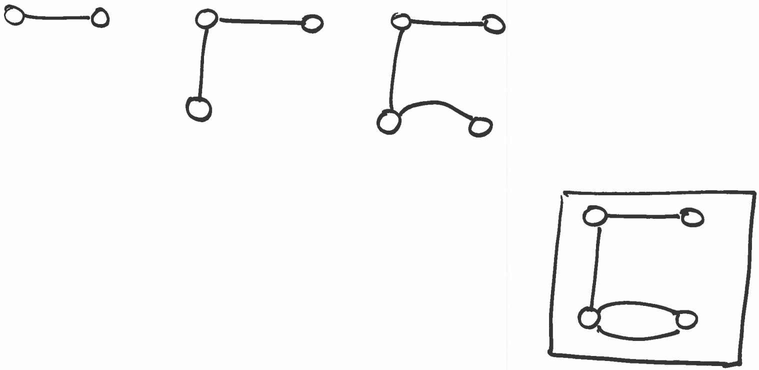
Models


dense
sparse
Vertex exchangeable
Edge exchangeable
Pitman Yor
Preferential attachment

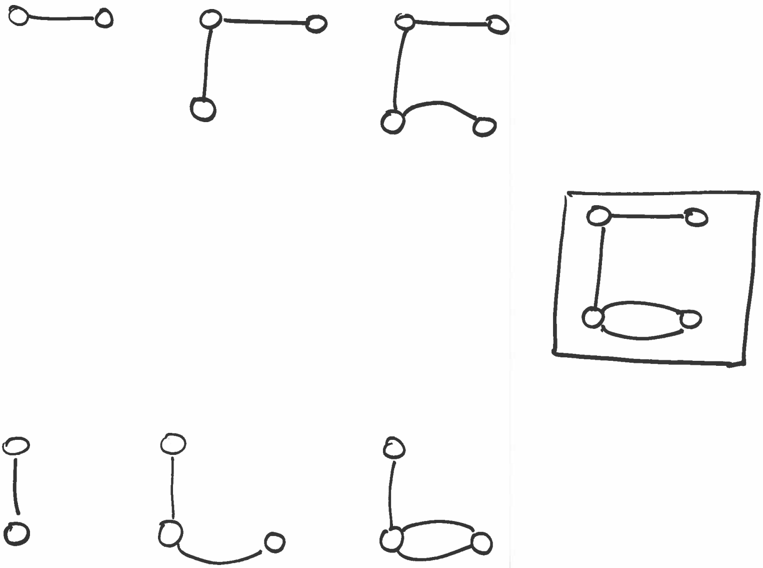
Edge exchangeable models [9], [8]



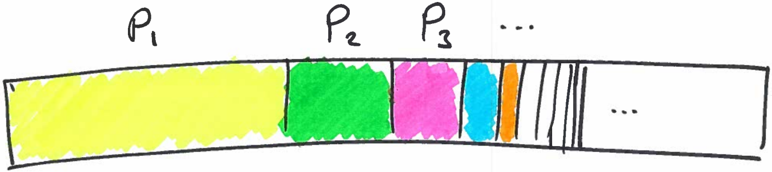
Edge exchangeable models [9], [8]



Edge exchangeable models [9], [8]



Paintbox representation

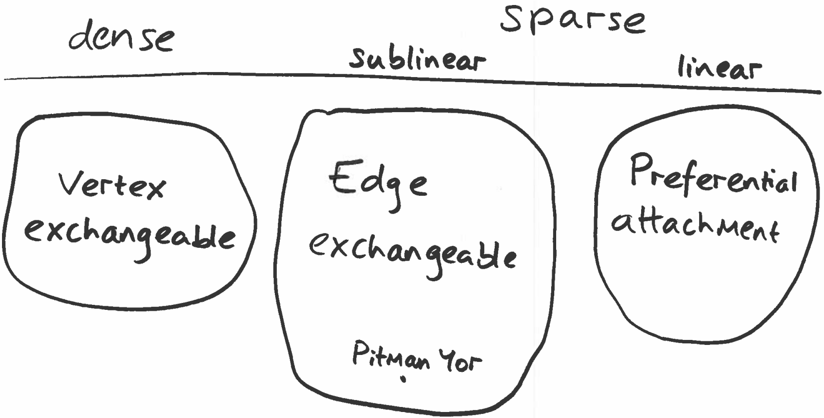


Paintbox representation

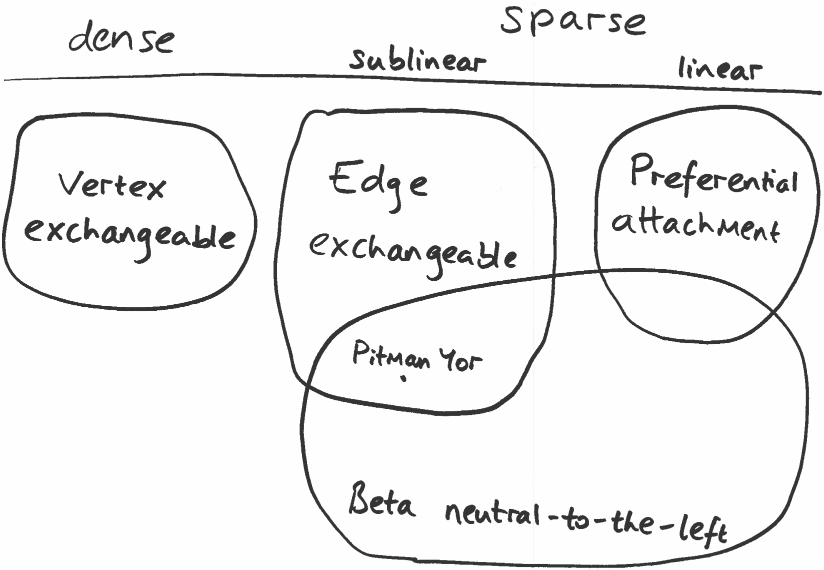
Consequence

- Edge exchangeable models have sublinear sparsity

Models

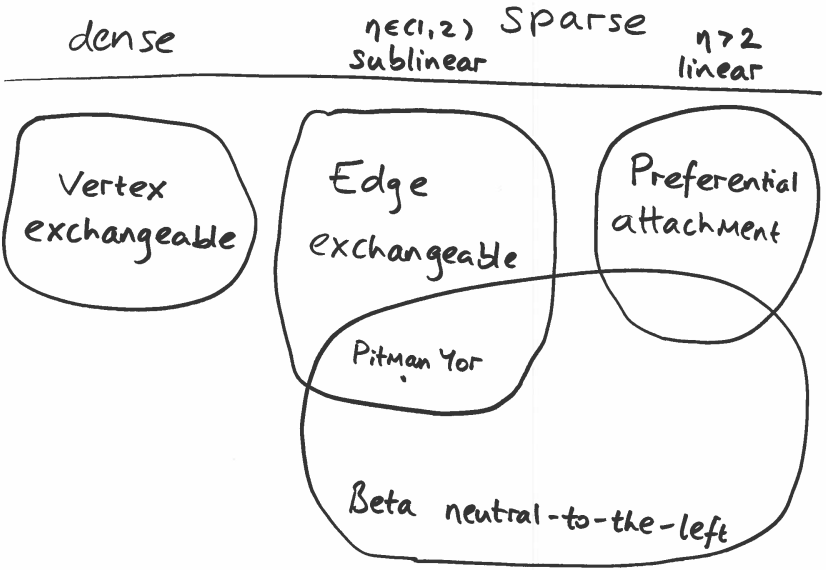

dense
Sparse
sublinear
linear
Vertex exchangeable
Edge exchangeable
Pitman Yor
Preferential attachment

Models



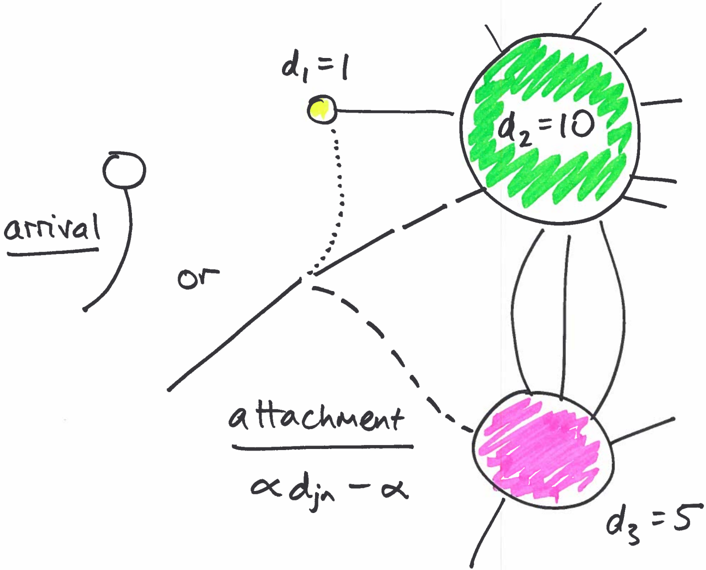
dense
sparse
sublinear
linear
Vertex exchangeable
Edge exchangeable
Preferential attachment
Pitman Yor
Beta neutral-to-the-left

Models

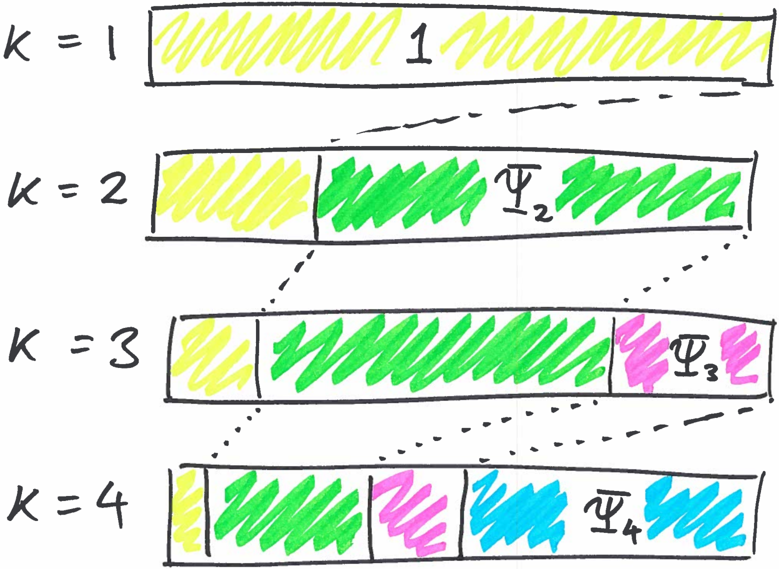


dense
$\eta \in (1,2)$ sparse
sublinear
$\eta > 2$
linear
Vertex exchangeable
Edge exchangeable
Preferential attachment
Pitman Yor
Beta neutral-to-the-left

Beta Neutral-to-the-left Model [10]



Latent representation



Sampling and inference

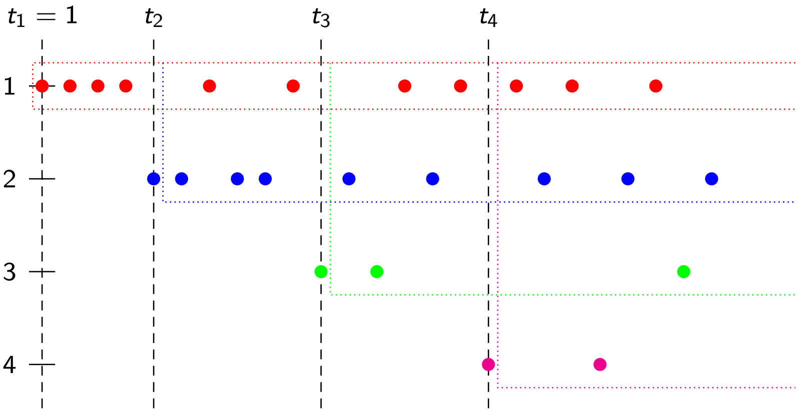
Why a paper on sampling and inference for BNTL models?

Sampling and inference

Why a paper on sampling and inference for BNTL models?

- Inference is notoriously difficult for *non-exchangeable* structures
- Need to identify *exchangeable substructures*

Exchangeable substructure



Gibbs structure

The joint density has **Gibbs structure**

$$p(\text{graph}|\mathbf{T}) = \frac{\Gamma(d_1 - \alpha)}{\Gamma(n - K\alpha)} \prod_{j=2}^{K} \frac{\Gamma(T_j - j\alpha)\Gamma(d_j - \alpha)}{\Gamma(T_j - 1 - (j-1)\alpha)\Gamma(1 - \alpha)}$$

- T_j = arrival time of vertex j
- d_j = degree of vertex j
- K = #vertices
- n = #edges

Available data

Observation	**Unobserved variables**
Entire history	$\alpha, \phi, \mathbf{\Psi}$
Degrees in arrival order	$\alpha, \phi, \mathbf{\Psi}, \mathbf{T}$
Snapshot	$\alpha, \phi, \mathbf{\Psi}, \mathbf{T}, \sigma$

- $\alpha =$ BTNL parameter $\in (-\infty, 1)$
- $\phi =$ arrival distribution parameters
- $\mathbf{\Psi} =$ latent sociabilities
- $\mathbf{T} =$ arrival times
- $\sigma =$ arrival order

Sampling $\boldsymbol{\Psi}$

Beta prior on Ψ_j, plus Gibbs structure, give

$$\Psi_j \mid \text{graph}, \boldsymbol{\Psi}_{\setminus j} \sim \text{Beta}(d_j - \alpha, \bar{d}_{j-1} - (j-1)\alpha) \, ,$$

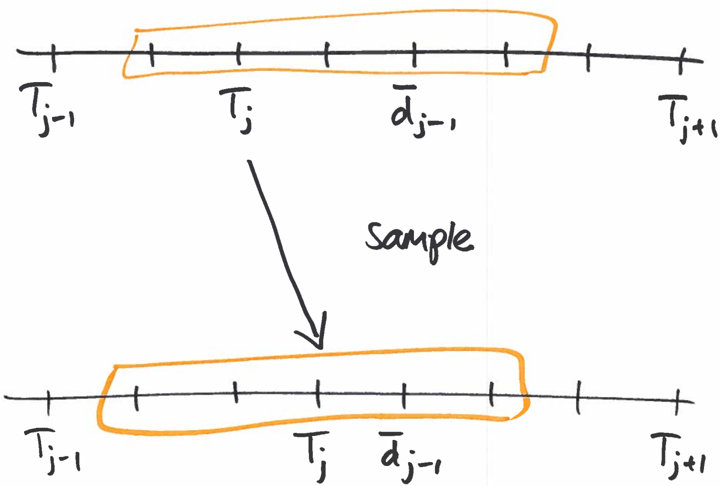
where $\bar{d}_j = \sum_{i=1}^{j} d_i$

Sampling α, ϕ

- One-dimensional unnormalized density for α
- For ϕ depends on arrival distribution family

Sampling **T**

TODO remove $\bar{d}$ from this picture



Sampling σ

- Initialise in descending degree order
- Use Metropolis-Hastings with adjacent swap proposal $\sigma_j \leftrightarrow \sigma_{j+1}$

Point estimation

- Decompose $p_{\alpha,\phi}(\text{graph}) = p_\phi(\mathbf{T})p_\alpha(\text{graph}|\mathbf{T})$
- MLE/MAP estimation for α by optimizing density

Experiments

- I DON'T KNOW
- Synthetic data – parameter recovery
- Scaling in n
- Point estimation with massive graphs

Synthetic data

- Simulate 500 edges from the prior with fixed α
- Arrivals either $\mathcal{PYP}$ or Geom
- Observe final snapshot of the graph only

Gibbs sampler results

Gen. arrival distn.	Inference model	$\lvert\hat{\alpha} - \alpha^*\rvert$	Pred. log-lik.
$\mathcal{PYP}(1.0, 0.75)$	$(\tau, \mathcal{PYP}(\theta, \tau))$	**0.046** $\pm$ **0.002**	-**2637.0** $\pm$ **0.1**
$\mathcal{PYP}(1.0, 0.75)$	$(\alpha, \text{Geom}(\beta))$	0.049 $\pm$ 0.004	-2660.5 $\pm$ 0.7
Geom(0.25)	$(\tau, \mathcal{PYP}(\theta, \tau))$	0.086 $\pm$ 0.002	-2386.8 $\pm$ 0.1
Geom(0.25)	$(\alpha, \text{Geom}(\beta))$	**0.043** $\pm$ **0.003**	-**2382.6** $\pm$ **0.2**

Scalability of Gibbs sampler

- Do we learn from all data?
- How does performance scale?

Scalability of Gibbs sampler

- Do we learn from all data?
- How does performance scale?

	$n = 200$	$n = 20000$
$\lvert\hat{\alpha} - \alpha^*\rvert$	0.12 ± 0.01	0.01 ± 0.00
$\lvert\hat{\beta} - \beta^*\rvert$	0.02 ± 0.00	0.00 ± 0.00
ESS	0.90 ± 0.04	0.75 ± 0.08
Runtime (s)	21 ± 0	2267 ± 2

- Most expensive Gibbs update is for $\mathbf{T}$

MLEs for SNAP datasets

Fitted point estimates

Dataset	Coupled $\mathcal{PYP}(\theta, \alpha)$			Uncoupled $\mathcal{PYP}(\theta, \tau)$			Geom(β)		
	$(\hat{\theta}, \hat{\alpha})$	$\hat{\eta}$	Pred. l-l.	$\hat{\alpha}$	$(\hat{\theta}, \hat{\tau})$	Pred. l-l.	$\hat{\beta}$	$\hat{\eta}$	Pred. l-l.
Ask Ubuntu	(18080, 0.25)	1.25	-3.707e6	-2.54	(-0.99, 0.99)	-3.678e6	0.083	2.32	**-3.678e6**
UCI social network	(320.4, 4.4e-11)	–	-1.600e5	-4.98	(5.50, 0.52)	**-1.595e6**	0.016	2.10	-1.596e5
EU email	(113.6, 2.5e-14)	–	**-8.06e5**	-1.86	(113.6, 9.2e-10)	**-8.06e5**	0.001	2.00	-8.07e5
Math Overflow	(2575, 0.15)	1.15	-1.685e6	-6.62	(-0.97, 0.997)	-1.670e6	0.025	2.19	**-1.670e6**
Stack Overflow	(297600, 0.11)	1.11	-3.358e8	-8.94	(-1.0, 1.0)	-3.333e8	0.020	2.21	**-3.333e8**
Super User	(20640, 0.24)	1.24	-5.855e6	-4.19	(-0.996, 1.0)	**-5.775e6**	0.067	2.37	-5.775e6
Wikipedia talk pages	(14870, 0.54)	1.54	-3.074e7	-0.25	(-1.0, 1.0)	**-3.066e7**	0.073	2.10	-3.066e7

MLEs for SNAP datasets

$\mathcal{PYP}$ parameter estimates vary coupled and uncoupled

Dataset	Coupled $\mathcal{PYP}(\theta, \alpha)$				Uncoupled $\mathcal{PYP}(\theta, \tau)$		Geom(β)		
	$(\hat{\theta}, \hat{\alpha})$	$\hat{\eta}$	Pred. l-l.	$\hat{\alpha}$	$(\hat{\theta}, \hat{\tau})$	Pred. l-l.	$\hat{\beta}$	$\hat{\eta}$	Pred. l-l.
Ask Ubuntu	(18080, 0.25)	1.25	-3.707e6	-2.54	(-0.99, 0.99)	-3.678e6	0.083	2.32	**-3.678e6**
UCI social network	(320.4, 4.4e-11)	–	-1.600e5	-4.98	(5.50, 0.52)	**-1.595e6**	0.016	2.10	-1.596e5
EU email	(113.6, 2.5e-14)	–	**-8.06e5**	-1.86	(113.6, 9.2e-10)	**-8.06e5**	0.001	2.00	-8.07e5
Math Overflow	(2575, 0.15)	1.15	-1.685e6	-6.62	(-0.97, 0.997)	-1.670e6	0.025	2.19	**-1.670e6**
Stack Overflow	(297600, 0.11)	1.11	-3.358e8	-8.94	(-1.0, 1.0)	-3.333e8	0.020	2.21	**-3.333e8**
Super User	(20640, 0.24)	1.24	-5.855e6	-4.19	(-0.996, 1.0)	**-5.775e6**	0.067	2.37	-5.775e6
Wikipedia talk pages	(14870, 0.54)	1.54	-3.074e7	-0.25	(-1.0, 1.0)	**-3.066e7**	0.073	2.10	-3.066e7

MLEs for SNAP datasets

Edge exchangeable models likely misspecified

Dataset	Coupled $\mathcal{PYP}(\theta, \alpha)$				Uncoupled $\mathcal{PYP}(\theta, \tau)$		Geom(β)		
	$(\hat{\theta}, \hat{\alpha})$	$\hat{\eta}$	Pred. l-l.	$\hat{\alpha}$	$(\hat{\theta}, \hat{\tau})$	Pred. l-l.	$\hat{\beta}$	$\hat{\eta}$	Pred. l-l.
Ask Ubuntu	(18080, 0.25)	1.25	-3.707e6	-2.54	(-0.99, 0.99)	-3.678e6	0.083	2.32	**-3.678e6**
UCI social network	(320.4, 4.4e-11)	–	-1.600e5	-4.98	(5.50, 0.52)	**-1.595e6**	0.016	2.10	-1.596e5
EU email	(113.6, 2.5e-14)	–	**-8.06e5**	-1.86	(113.6, 9.2e-10)	**-8.06e5**	0.001	2.00	-8.07e5
Math Overflow	(2575, 0.15)	1.15	-1.685e6	-6.62	(-0.97, 0.997)	-1.670e6	0.025	2.19	**-1.670e6**
Stack Overflow	(297600, 0.11)	1.11	-3.358e8	-8.94	(-1.0, 1.0)	-3.333e8	0.020	2.21	**-3.333e8**
Super User	(20640, 0.24)	1.24	-5.855e6	-4.19	(-0.996, 1.0)	**-5.775e6**	0.067	2.37	-5.775e6
Wikipedia talk pages	(14870, 0.54)	1.54	-3.074e7	-0.25	(-1.0, 1.0)	**-3.066e7**	0.073	2.10	-3.066e7

MLEs for SNAP datasets

Though better than Geom for some datasets

Dataset	Coupled $\mathcal{PYP}(\theta, \alpha)$				Uncoupled $\mathcal{PYP}(\theta, \tau)$		Geom(β)		
	$(\hat{\theta}, \hat{\alpha})$	$\hat{\eta}$	Pred. l-l.	$\hat{\alpha}$	$(\hat{\theta}, \hat{\tau})$	Pred. l-l.	$\hat{\beta}$	$\hat{\eta}$	Pred. l-l.
Ask Ubuntu	(18080, 0.25)	1.25	-3.707e6	-2.54	(-0.99, 0.99)	-3.678e6	0.083	2.32	**-3.678e6**
UCI social network	(320.4, 4.4e-11)	–	-1.600e5	-4.98	(5.50, 0.52)	**-1.595e6**	0.016	2.10	-1.596e5
EU email	(113.6, 2.5e-14)	–	**-8.06e5**	-1.86	(113.6, 9.2e-10)	**-8.06e5**	0.001	2.00	-8.07e5
Math Overflow	(2575, 0.15)	1.15	-1.685e6	-6.62	(-0.97, 0.997)	-1.670e6	0.025	2.19	**-1.670e6**
Stack Overflow	(297600, 0.11)	1.11	-3.358e8	-8.94	(-1.0, 1.0)	-3.333e8	0.020	2.21	**-3.333e8**
Super User	(20640, 0.24)	1.24	-5.855e6	-4.19	(-0.996, 1.0)	**-5.775e6**	0.067	2.37	-5.775e6
Wikipedia talk pages	(14870, 0.54)	1.54	-3.074e7	-0.25	(-1.0, 1.0)	**-3.066e7**	0.073	2.10	-3.066e7

Conclusion

- BNTL models are *flexible*
- BNTL models are *tractable*

Future work

- Scalability of inference
  - Metropolis-Hastings to update $\mathbf{T}$ altogether
  - Variational inference for σ

References

[1] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27. Cambridge University Press, 1989.

[2] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[3] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[4] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[5] Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979.

[6] François Caron and Emily B Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

[7] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[8] Harry Crane and Walter Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, (just-accepted), 2017.

[9] Diana Cai, Trevor Campbell, and Tamara Broderick. Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, pages 4249–4257, 2016.

[10] Benjamin Bloem-Reddy and Peter Orbanz. Preferential attachment and vertex arrival times. *arXiv preprint arXiv:1710.02159*, 2017.

[11] Scott W Linderman, Gonzalo E Mena, Hal Cooper, Liam Paninski, and John P Cunningham. Reparameterizing the birkhoff polytope for variational permutation inference. *arXiv preprint arXiv:1710.09508*, 2017.