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

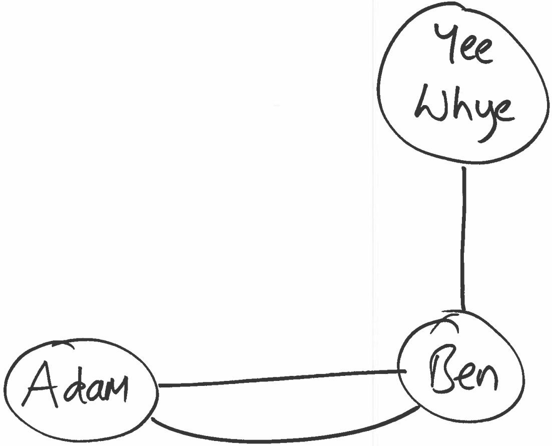
Examples

- Messages between people (email, WhatsApp, ...)
- Posts + replies on StackOverflow

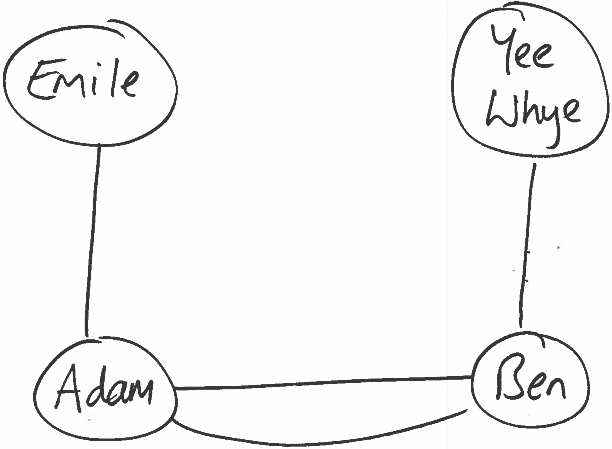
Temporal networks

Temporal networks

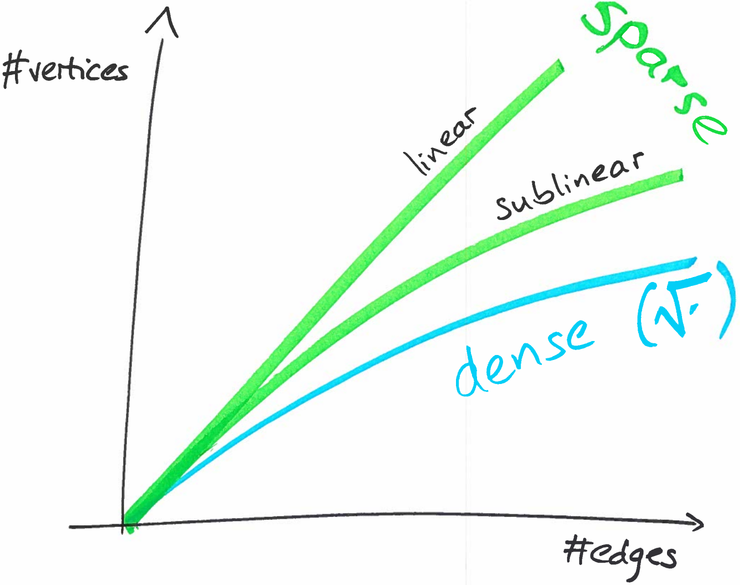
Temporal networks



Temporal networks



Sparsity



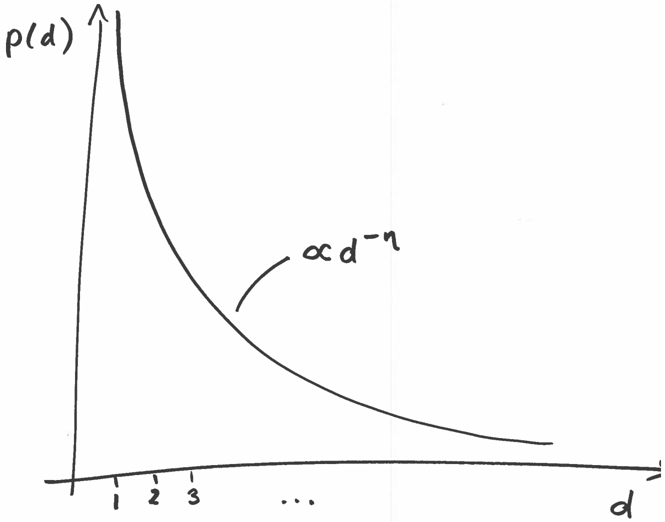
Power law degree distribution

Power law distribution of exponent η

$$p(d) \propto d^{-\eta}$$

where $\eta > 1$

Power law degree distribution



Sparity and power law

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

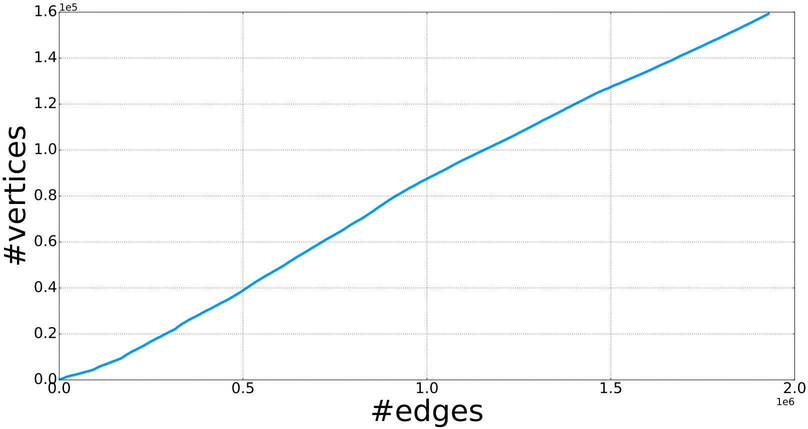
$$\textbf{Linear sparsity} \iff \eta > 2$$

Empirical study

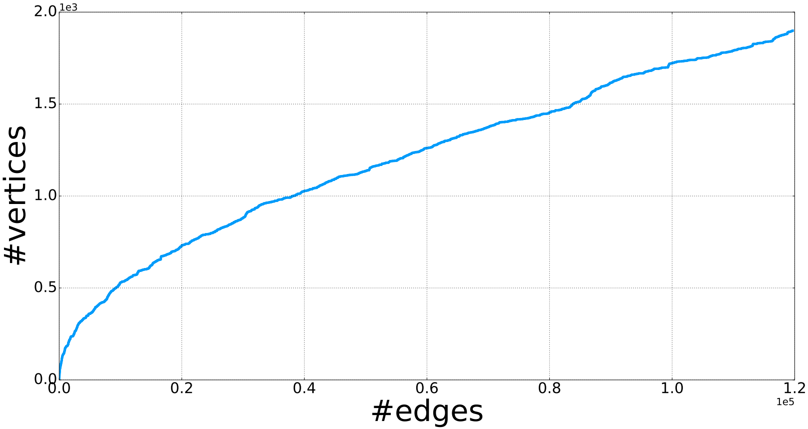
SNAP datasets [2]

Dataset	# of vertices	# of edges
Ask Ubuntu	159,316	964,437
UCI social network	1,899	20,296
EU email	986	332,334
Math Overflow	24,818	506,550
Stack Overflow	2,601,977	63,497,050
Super User	194,085	1,443,339
Wikipedia talk pages	1,140,149	7,833,140

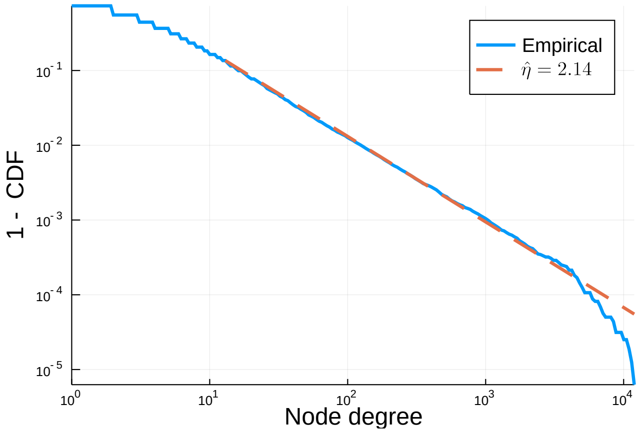
Ask Ubuntu



UCI social network

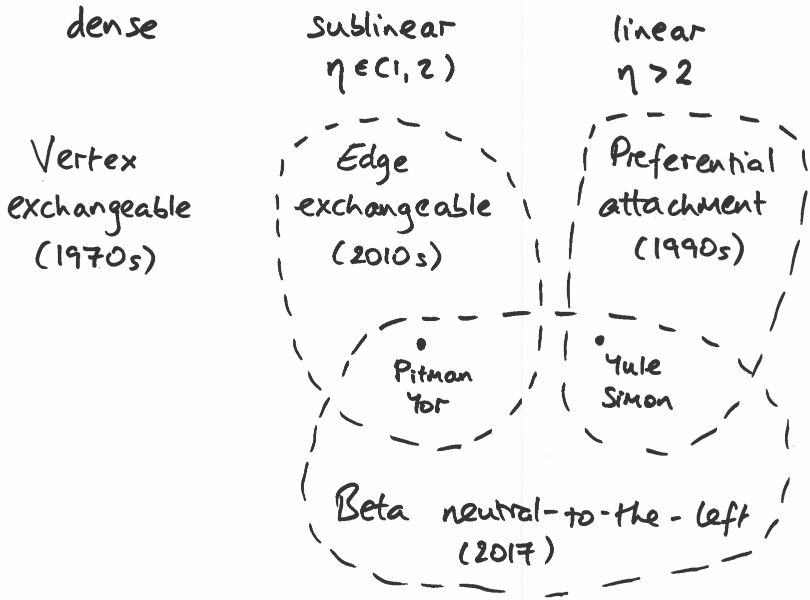


Ask Ubuntu degree distribution

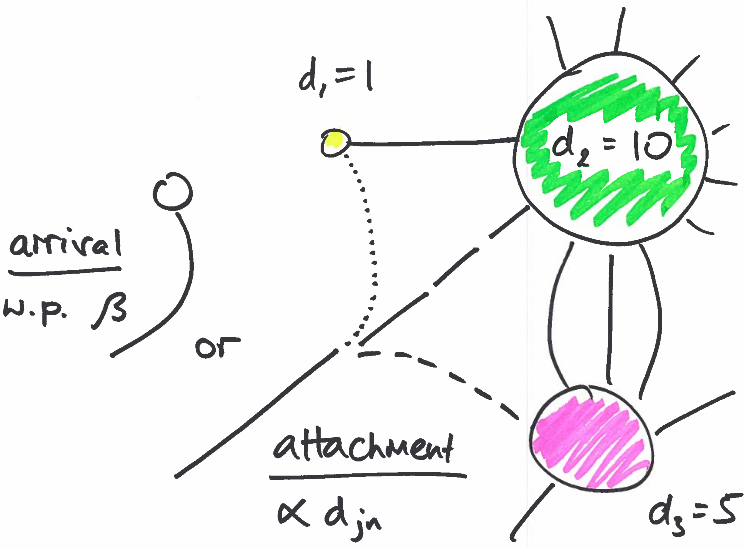


$\hat{\eta} = 2.14$ estimated using technique of [3]

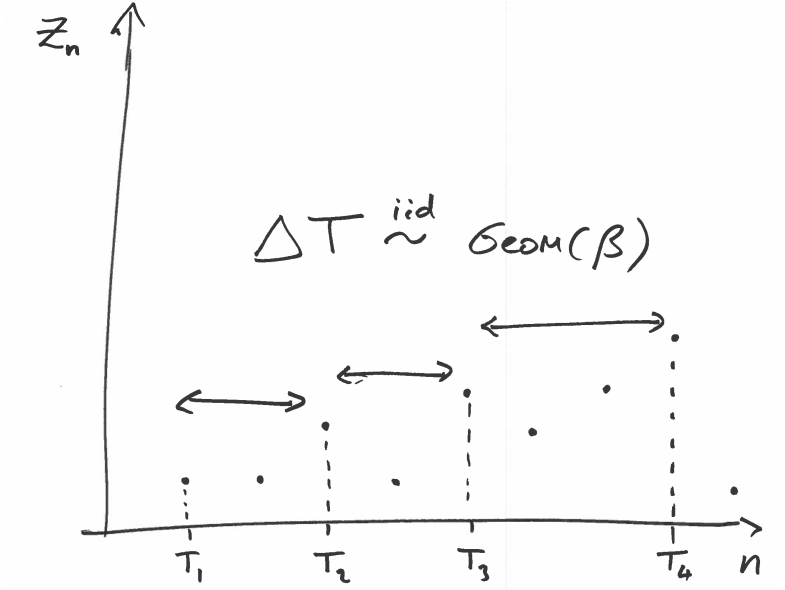
Models


dense
sublinear
$\eta \in (1, 2)$
linear
$\eta > 2$
Vertex exchangeable (1970s)
Edge exchangeable (2010s)
Preferential attachment (1990s)
Pitman Yor
Yule Simon
Beta neutral-to-the-left (2017)

Yule-Simon Process



Yule-Simon Process

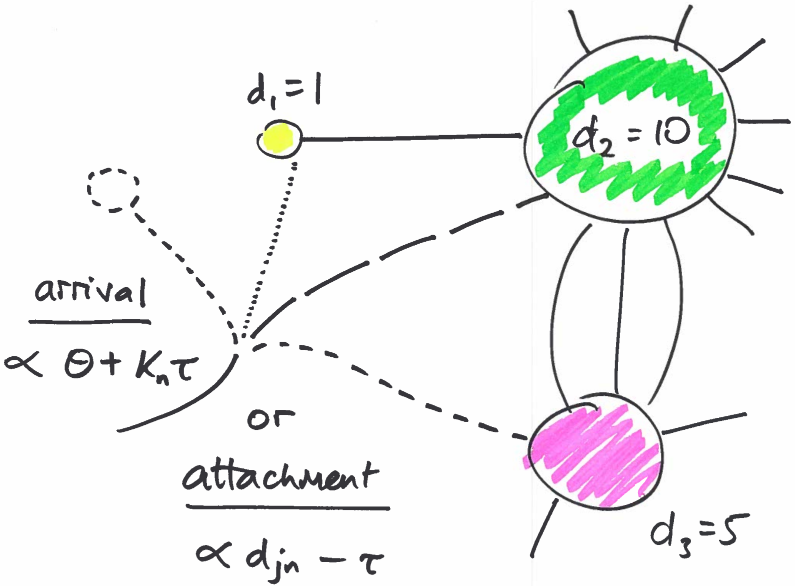


Yule-Simon Process

Asymptotic power law degree distribution with

$$\eta = 1 + \frac{1}{1-\beta} > 2$$

Pitman-Yor Process



Pitman-Yor Process

Asymptotic power law degree distribution with

$$\eta = 1 + \tau \in (1, 2)$$

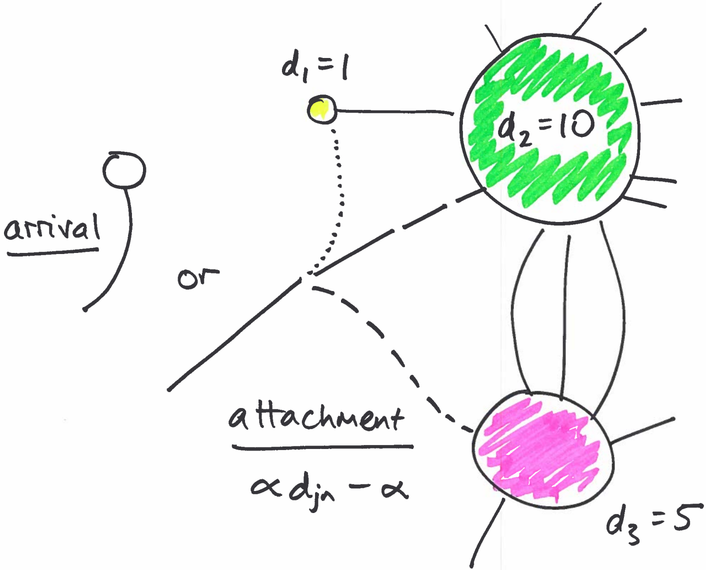
and $K_n = o(n)$

Edge exchangeable models [9], [8]

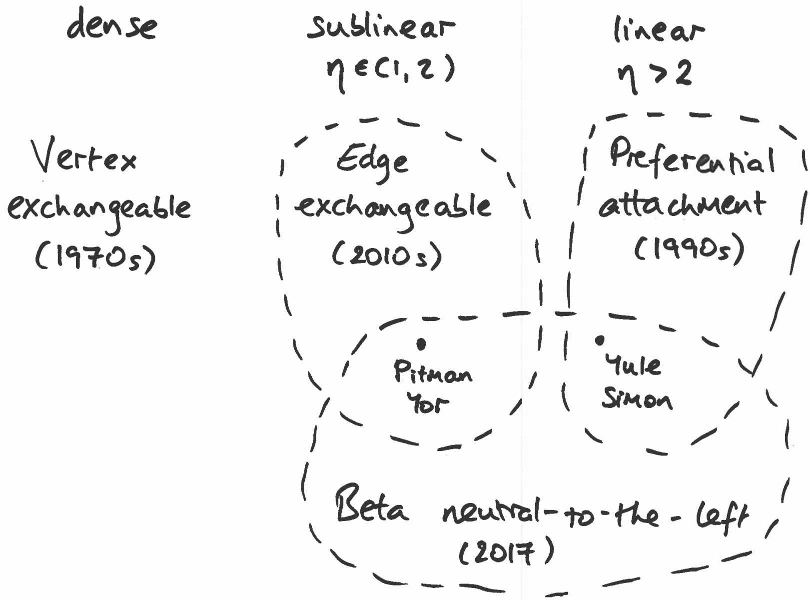
"The probability of all orderings of edge arrivals is the same"

- Sublinear sparsity
- $\eta \in (1, 2)$

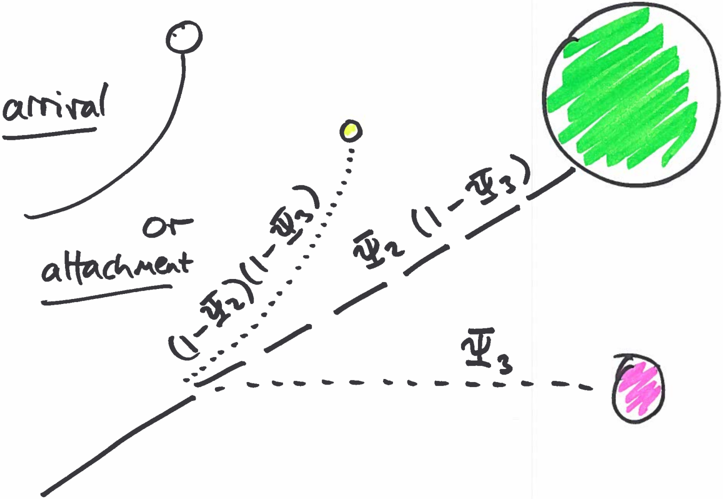
Beta Neutral-to-the-left Process [10]



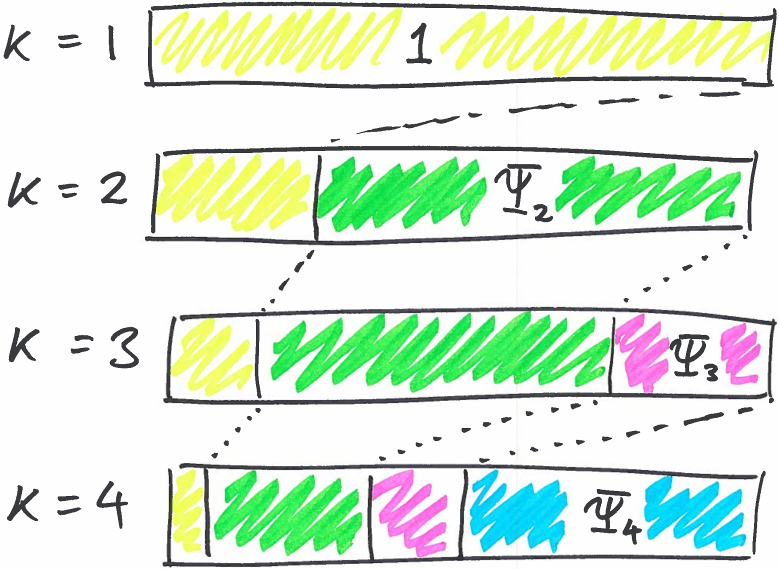
Models – again



Hierarchical representation of BNTL process



Recursive scaling of BNTL latents



BNTL properties

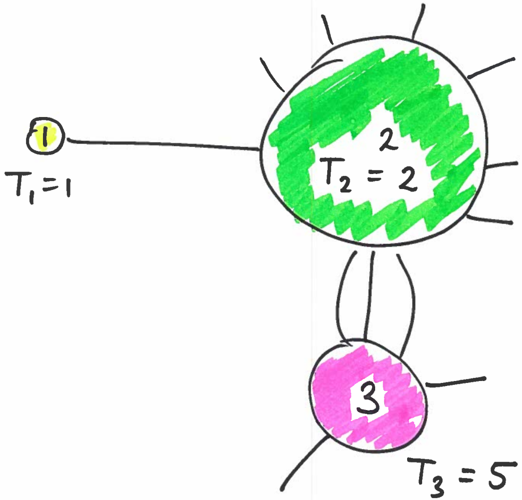
- Collapsed sampler
- Latent representation **not** from de Finetti

Sampling and inference

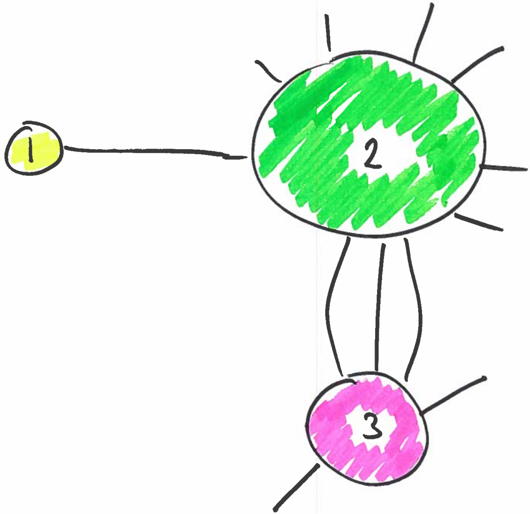
Three observation cases

- Entire history
- Vertex order
- Snapshot

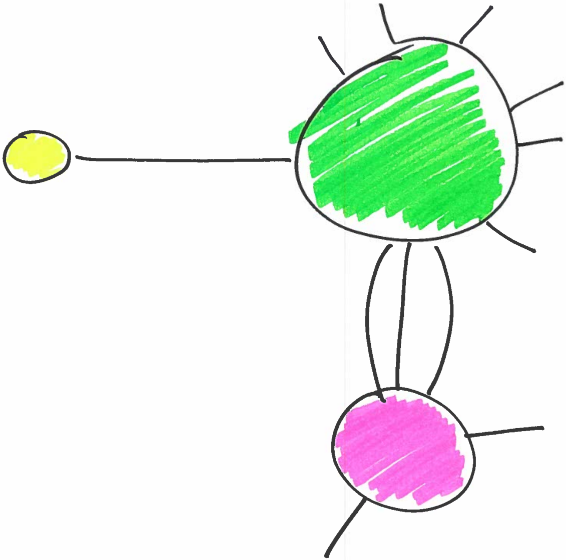
Entire history



Vertex order



Snapshot



Observation cases

Observation	**Unobserved variables**
Entire history	$\alpha, \phi, \mathbf{\Psi}_{K_n}$
Vertex order	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}$
Snapshot	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}, \sigma[K_n]$

Sampling $\mathbf{\Psi}$

Beta prior on Ψ_j, plus recursive scaling –

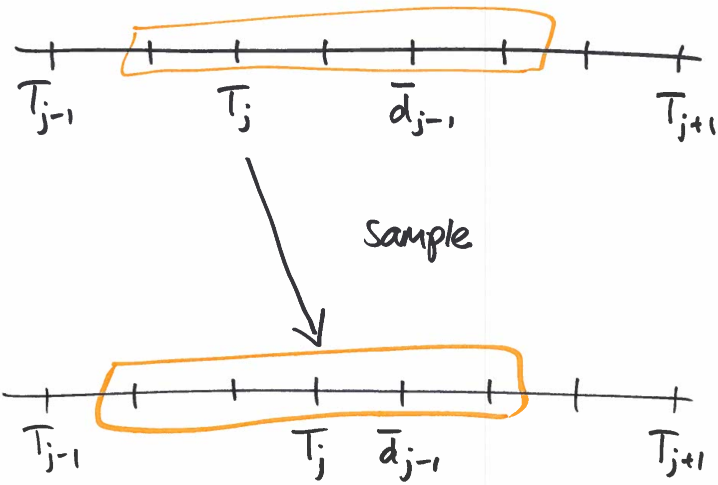
$$\Psi_j \mid \mathbf{Z}_n, \mathbf{\Psi}_{\setminus j} \sim \mathrm{Beta}(d_{j,n} - \alpha, \bar{d}_{j-1,n} - (j-1)\alpha) \ ,$$

- For fixed α, we have our posterior
- Learning other variables, we have a Gibbs update

Sampling α, ϕ

- For α, one-dimensional unnormalized density
- For ϕ, depends on family. Our experiments used conjugacy or slice sampling.

Sampling **T**



Sampling $\sigma[K_n]$

- Use Metropolis-Hastings with swap proposal $\sigma_j \leftrightarrow \sigma_{j+1}$

Point estimation

- MLE/MAP estimation for α, ϕ by optimizing unnormalized density

Experiments

- 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**
$\mathcal{PYP}(1.0, 0.75)$	$(\alpha, \mathrm{Geom}(\beta))$	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)	$(\alpha, \mathrm{Geom}(\beta))$	**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 Gibss 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 $\mathbf{T}$ [11]
- Recency-weighted preferential attachment

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.

Theorem Under exchangeable models, $K_n = o(n)$.
Proof

$\mathbb{P}(K_n/n \to 0) = \mathbb{E}[\mathbb{P}(K_n/n \to 0 \mid \text{paintbox})]$ by Paintbox / de Finetti

Enough to show $\mathbb{E}[K_n|\text{paintbox}]/n \to 0$. We have

$$\begin{aligned}
\mathbb{E}[K_n] &= \mathbb{E}\left[\sum_j \mathbf{1}(\text{visited } j \text{ by } n)\right] \\
&= \sum_j \mathbb{P}(\text{visited } j \text{ by } n) \text{ by Monotone Convergence} \\
&\leq \sum_{j:P_j>1/\sqrt{n}} 1 + \sum_{j:P_j\leq 1/\sqrt{n}} nP_j \text{ by Bonferroni} \\
&\leq \sqrt{n} + n \sum_{j:P_j\leq 1/\sqrt{n}} P_j \\
&= o(n)
\end{aligned}$$