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 on WhatsApp
- Posts + replies on StackOverflow

Temporal networks

Examples

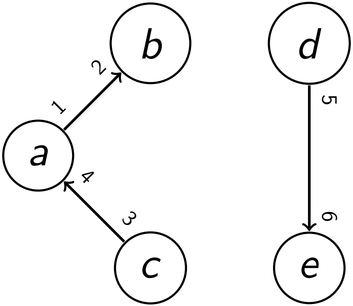
- Messages on WhatsApp
- Posts + replies on StackOverflow

Abstraction

- Graph grows adding one edge (Z_i, Z_{i+1}) at a time
- Vertices enter the graph when connected to

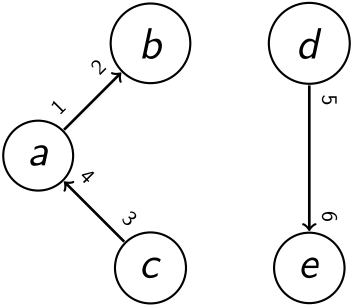
Temporal networks

Ends of edges $\mathbf{Z}_n = Z_1, ..., Z_n$
E.g. $\mathbf{Z}_6 = \underline{a, b}, \underline{c, a}, \underline{d, e}$



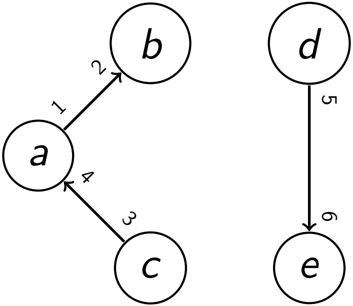
Temporal networks

Number of vertices K_n
E.g. $K_6 = 5$



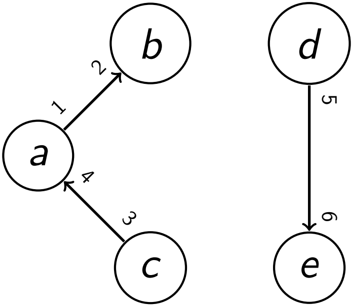
Temporal networks

Arrival time of vertex j is $T_j := \inf\{n : Z_n = j\}$
E.g. $T_e = 6$



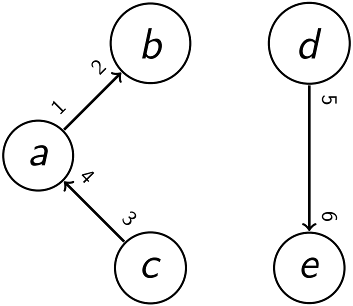
Temporal networks

Degree of vertex j is $d_{j,n}$
E.g. $d_{e,6} = 1$



Temporal networks

Degree counts $m_n(d) := |\{j : d_{j,n} = d\}|$
E.g. $m_6(1) = 4, m_6(2) = 1$



Sparsity

- For a dense graph, $K_n = O(n^{1/2})$
- For a sparse graph,

$$K_n = O(n^{1/(1+\sigma)})$$

  for $0 \leq \sigma < 1$
- Stack Overflow network likely sparse

Power law degree distribution

Power law distribution of exponent η

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

where $\eta > 1$

Power law degree distribution

Power law distribution of exponent η

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

where $\eta > 1$

Asymptotic degree distribution has **power law tail with exponent** η if

$$\frac{m_n(d)}{K_n} \xrightarrow[n\to\infty]{p} L(d)d^{-\eta}\ , \tag{1}$$

for slowly varying $L(d)$

Power law degree distribution

Power law distribution of exponent η

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

where $\eta > 1$
Asymptotic degree distribution has **power law tail with exponent** η if

$$\frac{m_n(d)}{K_n} \xrightarrow[n\to\infty]{p} L(d)d^{-\eta}\ , \tag{1}$$

for slowly varying $L(d)$

Slowly varying function has $\lim_{x\to\infty} L(rx)/L(x) = 1$ for all $r > 0$ [1]

Power laws and sparsity

We have

$$K_n = \sum_{d=1}^{n} m_n(d),$$
$$n = \sum_{d=1}^{n} d\, m_n(d).$$

Power laws and sparsity

Suppose $m_n(d)$ is power law distributed

$$K_n = C \sum_{d=1}^{n} d^{-\eta},$$

$$n = C \sum_{d=1}^{n} d^{-\eta+1}$$

$$= K_n \frac{\sum_{d=1}^{n} d^{-\eta+1}}{\sum_{d=1}^{n} d^{-\eta}}.$$

Power laws and sparsity

Letting $n \to \infty$ in
$$\frac{K_n}{n} = \frac{\sum_{d=1}^{n} d^{-\eta}}{\sum_{d=1}^{n} d^{-\eta+1}}$$
we see $K_n = O(n)$ if $\eta > 2$, $K_n = o(n)$ if $\eta \in (1, 2]$.

Power laws and sparsity

Letting $n \to \infty$ in

$$\frac{K_n}{n} = \frac{\sum_{d=1}^{n} d^{-\eta}}{\sum_{d=1}^{n} d^{-\eta+1}}$$

we see $K_n = O(n)$ if $\eta > 2$, $K_n = o(n)$ if $\eta \in (1, 2]$.

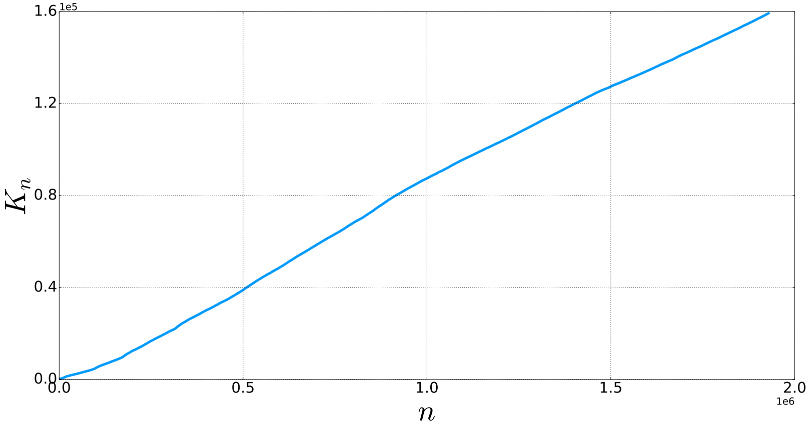
Summary
For sparse graphs, $\sigma = 0 \leftrightarrow \eta > 2$ and $\sigma > 0 \leftrightarrow \eta \in (1, 2]$.

Empirical study

SNAP datasets [2]

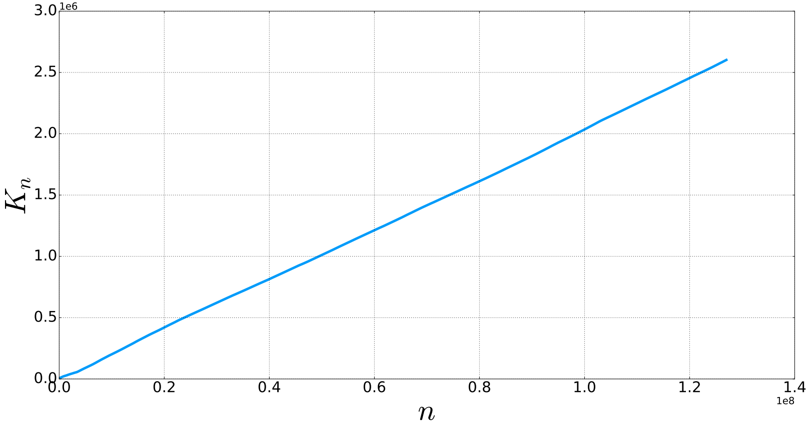
Dataset	# of vertices	# of edges
Ask Ubuntu	159,316	964,437
UCI social network	1,899	20,296
EU email	986	332,334
Math Overflow	24,818	506,550
Stack Overflow	2,601,977	63,497,050
Super User	194,085	1,443,339
Wikipedia talk pages	1,140,149	7,833,140

Ask Ubuntu arrival process



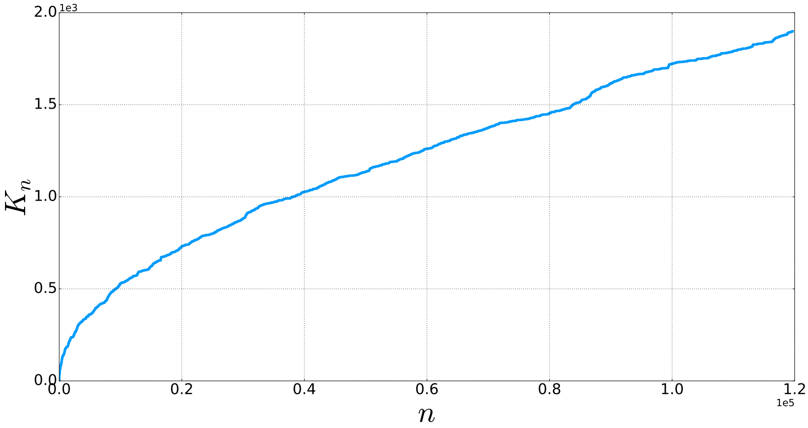
$\hat{\sigma} = -0.099$

Stack Overflow arrival process



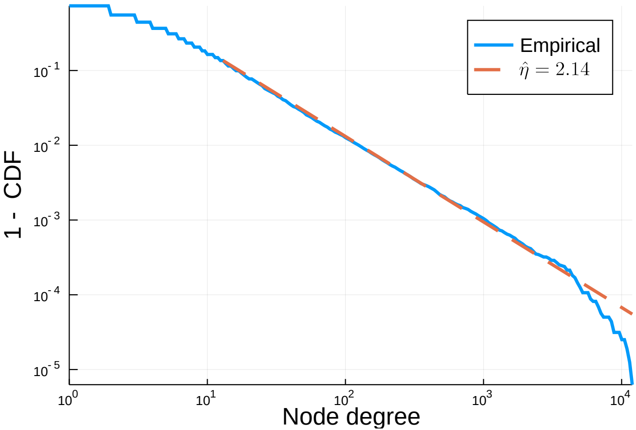
$\hat{\sigma} = 0.015$

UCI social network arrival process



$\hat{\sigma} = 0.952$

Ask Ubuntu degree distribution



Estimation used technique of [3]

Models

- Vertex exchangeable models do not give sparsity [4] [5]
- Exchangeable point process models [6] have an independent notion of time

Models

- Vertex exchangeable models do not give sparsity [4] [5]
- Exchangeable point process models [6] have an independent notion of time
- **Preferential attachment models** [7]
- **Edge exchangeable models** [8] [9]

Yule-Simon Process

Parameter $\beta \in (0, 1)$.

Arrivals

$$T_{j+1} - T_j \overset{\text{i.i.d.}}{\sim} \text{Geom}\,(\beta)$$

Size-biased reinforcement

$$Z_{n+1} | \mathbf{Z}_n, \mathbf{T} = \begin{cases} K_{n+1} \text{ w.p. } 1 & \text{if } n+1 = T_{K_{n+1}} \\ j \text{ w.p. } \propto d_{j,n} & \text{otherwise} \end{cases}$$

Yule-Simon Process

Asymptotic power law degree distribution with

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

and $K_n = O(n)$

Pitman-Yor Process

Parameters $\tau \in (0, 1), \theta > -\tau$.

Urn process

$$Z_{n+1}|\mathbf{Z}_n = \begin{cases} K_{n+1} \text{ w.p. } \dfrac{\theta + K_n\tau}{\theta + n} \\ \\ j \text{ w.p. } \dfrac{d_{j,n} - \tau}{\theta + n} \end{cases}$$

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"

For edge exchangeable models, $K_n = o(n)$
Sketch proof

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

Enough to show $\mathbb{E}[K_n|\text{paintbox}]/n \to 0$, which we can do directly from P_j of the paintbox.

Rewriting the Pitman-Yor Process

Parameters $\tau \in (0,1), \theta > -\tau$.

Arrivals

$$\mathbb{P}(T_{j+1} - T_j > t \mid T_j) = \prod_{i=1}^{t} \frac{T_j + t - j\tau}{T_j + t + \theta}$$

Size-biased reinforcement

$$Z_{n+1} | \mathbf{Z}_n, \mathbf{T} = \begin{cases} K_{n+1} \text{ w.p. } 1 & \text{if } n+1 = T_{K_{n+1}} \\ j \text{ w.p.} \propto (d_{j,n} - \tau) & \text{otherwise} \end{cases}$$

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

Parameters $\alpha \in (-\infty, 1)$ and Λ_ϕ a law on $\mathbb{N}^\infty$.

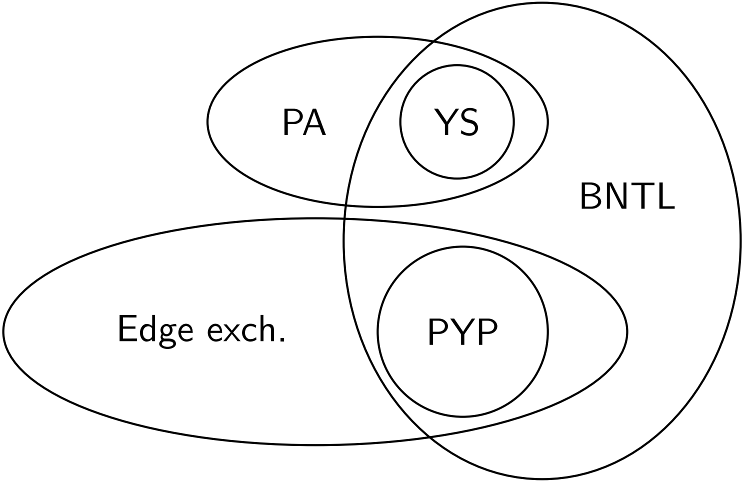
Arrivals

$$\mathbf{T} \sim \Lambda_\phi$$

Size-biased reinforcement

$$Z_{n+1} | \mathbf{Z}_n, \mathbf{T} = \begin{cases} K_{n+1} \text{ w.p. } 1 & \text{if } n+1 = T_{K_{n+1}} \\ j \text{ w.p. } \propto (d_{j,n} - \alpha) & \text{otherwise} \end{cases}$$

Relationship with other model classes



Hierarchical representation of BNTL process

Arrivals

$$\mathbf{T} \sim \Lambda_\phi$$

Latent sociabilities

$$\Psi_j | T_j \sim \mathrm{Beta}(1-\alpha, T_j - 1 - (j-1)\alpha) \text{ for } j \geq 1$$

Left-neutral resampling probabilities

$$P_{j,k+1} = \begin{cases} P_{j,k}(1-\Psi_{k+1}), & j \in \{1,\ldots,k\} \\ \Psi_{k+1}, & j = k+1 \end{cases}$$

Sampling rule

$$Z_{n+1} | \mathbf{P}_{K_n}, \mathbf{T} = \begin{cases} K_{n+1} \text{ w.p. } 1 & \text{if } n+1 = T_{K_{n+1}} \\ j \text{ w.p. } P_{j,K_n} & \text{otherwise} \end{cases}$$

BNTL properties

- Collapsed sampler
- Latent representation

BNTL properties

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

Sampling and inference

- Sampling posterior on latents
- Point estimation of latents
- Sampling predictive distribution

Sampling and inference

- Sampling posterior on latents **Condition on what?**
- Point estimation of latents
- Sampling predictive distribution

Observation cases

Observation	Unobserved variables
End of edge sequence $\mathbf{Z}_n$	$\alpha, \phi, \mathbf{\Psi}_{K_n}$
Vertex arrival-ordered graph $\mathbf{d}_{K_n}$	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}$
Unlabeled graph	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}, \sigma[K_n]$

Sampling $\boldsymbol{\Psi}$

If $\mathbf{Z}_n$ or $\mathbf{d}_{K_n}$ observed

$$p_{\alpha,\phi}(\boldsymbol{\Psi}_{K_n}, \mathbf{Z}_n | \mathbf{T}_{K_n}, \mathbf{d}_{K_n}) \propto \prod_{j=1}^{K_n} \Psi_j^{-\alpha}(1-\Psi_j)^{T_j-(j-1)\alpha-1}$$

$$\cdot \prod_{j=1}^{K_n} \Psi_j^{d_{j,n}-1}(1-\Psi_j)^{\bar{d}_{j-1,n}-T_j}$$

$$\propto \prod_{j=1}^{K_n} \Psi_j^{d_{j,n}-\alpha-1}(1-\Psi_j)^{\bar{d}_{j-1,n}-(j-1)\alpha-1}$$

where

$$\bar{d}_{j,n} = \sum_{i=1}^{j} d_{j,n}.$$

Sampling $\boldsymbol{\Psi}$

Spot a closed form for $\boldsymbol{\Psi}$

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

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

Sampling α, ϕ

- Place priors on α, ϕ
- Left with one-dimensional unnormalized density for α and MCMC is applicable
- For ϕ, depends on Λ_ϕ. Our experiments used conjugacy or slice sampling.

Sampling $\mathbf{T}$

Assume

$$\Lambda^{\phi}(\mathbf{T}_k) = \delta_{T_1}(1) \prod_{j=2}^{k} \Lambda_j^{\phi}(\Delta_j | T_{j-1}) \, ,$$

support of $T_j - T_{j-1} | T_{\setminus j}$ is

$$\{1, ..., \min(T_{j+1} - T_{j-1} - 1, \bar{d}_{j-1} - T_{j-1} + 1)\}$$

and we can compute each probability

$$p_{\alpha,\phi}(T_j - T_{j-1} = s | \mathbf{T}_{\setminus j}, \mathbf{d}_K) \propto \Lambda_j^{\phi}(s | T_{j-1}) \Lambda_{j+1}^{\phi}(T_{j+1} - T_{j-1} - s) \cdot \binom{\bar{d}_j - T_{j-1} - s}{d_j - 1}.$$

Sampling $\sigma[K_n]$

- Use Metropolis-Hastings with swap proposal $\sigma_j \leftrightarrow \sigma_{j+1}$
- Ratio of joints can be easily computed in terms of Γ function.

Point estimation

- Factorization $p_{\alpha,\phi}(\mathbf{Z}_n) = p_\alpha(\mathbf{Z}_n|\mathbf{T}_{K_n})\Lambda_\phi(\mathbf{T}_{K_n})$
- Learn α separately from ϕ using standard optimization
- We have explicit formulae for MLE/MAP estimates for $\boldsymbol{\Psi}$

Experiments

- Synthetic data – parameter recovery
- Scaling in n
- Point estimation with massive graphs

Synthetic data

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

Gibbs sampler results

Gen. arrival distn.	Inference model	$\lvert\hat{\alpha} - \alpha^*\rvert$	$\lvert\hat{\mathbf{S}} - \mathbf{S}^*\rvert$	Pred. log-lik.
$\mathcal{PYP}(1.0, 0.75)$	$(\tau, \mathcal{PYP}(\theta, \tau))$	$\mathbf{0.046 \pm 0.002}$	$\mathbf{28.5 \pm 0.7}$	$\mathbf{-2637.0 \pm 0.1}$
$\mathcal{PYP}(1.0, 0.75)$	$(\alpha, \text{Geom}(\beta))$	0.049 ± 0.004	66.8 ± 1.2	-2660.5 ± 0.7
Geom(0.25)	$(\tau, \mathcal{PYP}(\theta, \tau))$	0.086 ± 0.002	56.6 ± 1.3	-2386.8 ± 0.1
Geom(0.25)	$(\alpha, \text{Geom}(\beta))$	$\mathbf{0.043 \pm 0.003}$	$\mathbf{24.8 \pm 0.8}$	$\mathbf{-2382.6 \pm 0.2}$

where $\mathbf{S} := \frac{1}{K_n - 1} \sum_{j>1} (\bar{d}_{j-1} - T_j)$

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 $\mathbf{T}$

Large scale real data experiments

- MLE point estimation for SNAP datasets
- Predictive log-likelihood

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

MLEs for SNAP datasets

These datasets may lack sparsity

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
  - Metroplis-Hastings
  - variational inference [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

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