XIDIAN UNIVERSITY

TEXAS
The University of Texas at Austin

Deep Poisson Gamma Dynamical Systems

Dandan Guo[1,2], Bo Chen[1,2], Hao Zhang[1,2], Mingyuan Zhou[3]

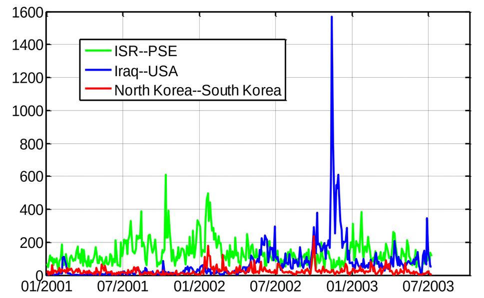
[1]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, Shaanxi, China [2]Collaborative Innovation Center of Information Sensing and Understanding at Xidian University

[3]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA

Motivation

Real world applications:

International Relation Study; Recommender System…



Time varying counts

- Capture complex dependencies across time steps
- Infer interpretable latent structure to analyze and predict

Related Work

Poisson Gamma Dynamical Systems (PGDS) [2]

$$y_v^{(t)} \sim \text{Pois}(\delta^{(t)} \textstyle\sum_{k=1}^K \phi_{vk}\, \theta_k^{(t)}) \text{ and } \theta_k^{(t)} \sim \text{Gam}(\tau_0 \textstyle\sum_{k_2=1}^K \pi_{kk_2}\, \theta_{k_2}^{(t-1)}, \tau_0)$$

✓ Gamma–Poisson construction, and supports expressive latent transition structures

✓ Shallow model may still have shortcomings in capturing long-range temporal dependencies

Deep Temporal Sigmoid Belief Networks [3]

$$p_{\theta}(\mathbf{V},\mathbf{H}) = p(\mathbf{h}_1)p(\mathbf{v}_1|\mathbf{h}_1)\cdot\prod_{t=2}^{T} p(\mathbf{h}_t|\mathbf{h}_{t-1},\mathbf{v}_{t-1})\cdot p(\mathbf{v}_t|\mathbf{h}_t,\mathbf{v}_{t-1})$$

$$p(h_{jt}=1|\mathbf{h}_{t-1},\mathbf{v}_{t-1}) = \sigma(\mathbf{w}_{1j}^\top \mathbf{h}_{t-1} + \mathbf{w}_{3j}^\top \mathbf{v}_{t-1} + b_j),\ p(\mathbf{v}_t|\mathbf{h}_t,\mathbf{v}_{t-1}) = \textstyle\prod_{m=1}^{M} y_{mt}^{v_{mt}}$$

✓ A sequential stack of sigmoid belief networks (SBNs)

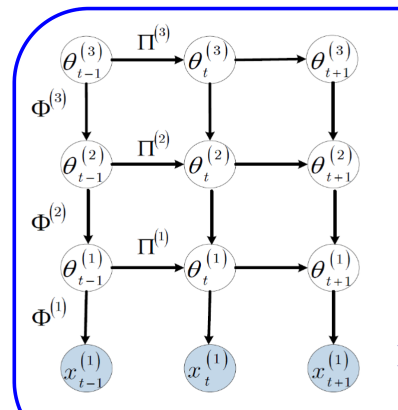
✓ How the layers in DTSBN are related to each other lacks an intuitive interpretation

✓ Modelling count data with Replicated Softmax Model

Contributions

- Improve previously proposed models by mining deep hierarchical latent structure from the data, and capturing long-range temporal dependencies.
- Derive closed-form Gibbs sampling update equations.
- Develop stochastic gradient MCMC inference that is scalable to big dataset.

Deep Poisson–Gamma Dynamical Systems



$$\boldsymbol{\theta}_t^{(L)} \sim \text{Gam}\left(\tau_0 \boldsymbol{\Pi}^{(L)} \boldsymbol{\theta}_{t-1}^{(L)}, \tau_0\right),$$

$$\cdots$$

$$\boldsymbol{\theta}_t^{(l)} \sim \text{Gam}\left(\tau_0 (\boldsymbol{\Phi}^{(l+1)} \boldsymbol{\theta}_t^{(l+1)} + \boldsymbol{\Pi}^{(l)} \boldsymbol{\theta}_{t-1}^{(l)}), \tau_0\right),$$

$$\cdots$$

$$\boldsymbol{x}_t^{(1)} \sim \text{Pois}\left(\delta_t^{(1)} \boldsymbol{\Phi}^{(1)} \boldsymbol{\theta}_t^{(1)}\right), \boldsymbol{\theta}_t^{(1)} \sim \text{Gam}\left(\tau_0 (\boldsymbol{\Phi}^{(2)} \boldsymbol{\theta}_t^{(2)} + \boldsymbol{\Pi}^{(1)} \boldsymbol{\theta}_{t-1}^{(1)}), \tau_0\right),$$

priors $\boldsymbol{\pi}_k^{(l)} \sim \text{Dir}(\nu_1^{(l)}\nu_k^{(l)}, \ldots, \nu_{k-1}^{(l)}\nu_k^{(l)}, \xi^{(l)}\nu_k^{(l)}, \nu_{k+1}^{(l)}\nu_k^{(l)} \ldots, \nu_{K_l}^{(l)}\nu_k^{(l)}),\ \ \nu_k^{(l)} \sim \text{Gam}(\frac{\gamma_0}{K_l}, \beta^{(l)})$

$\boldsymbol{\phi}_k^{(l)} = (\phi_{1k}^{(l)}, \ldots, \phi_{K_{l-1}k}^{(l)}) \sim \text{Dir}(\eta^{(l)}, \ldots, \eta^{(l)})$, and $\delta_t^{(1)}, \xi^{(l)}, \beta^{(l)} \sim \text{Gam}(\epsilon_0, \epsilon_0)$

- **Hierarchical structure** $\mathbb{E}\left[\boldsymbol{x}_t^{(1)} \,|\, \boldsymbol{\theta}_t^{(l)}, \{\boldsymbol{\Phi}^{(p)}\}_{p=1}^{l}\right] = \left[\prod_{p=1}^{l} \boldsymbol{\Phi}^{(p)}\right] \boldsymbol{\theta}_t^{(l)}$
- **Long-range temporal dependence**

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_t^{(1)} \,|\, \boldsymbol{\theta}_{t-1}^{(1)}, \boldsymbol{\theta}_{t-2}^{(2)}, \boldsymbol{\theta}_{t-3}^{(3)}]/\delta_t^{(1)} &= \boldsymbol{\Phi}^{(1)}\boldsymbol{\Pi}^{(1)}\boldsymbol{\theta}_{t-1}^{(1)} + \boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}[\boldsymbol{\Pi}^{(2)}]^2\boldsymbol{\theta}_{t-2}^{(2)} \\ &+ \boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}(\boldsymbol{\Pi}^{(2)}\boldsymbol{\Phi}^{(3)} + \boldsymbol{\Phi}^{(3)}\boldsymbol{\Pi}^{(3)})[\boldsymbol{\Pi}^{(3)}]^2\boldsymbol{\theta}_{t-3}^{(3)}\end{aligned}$$

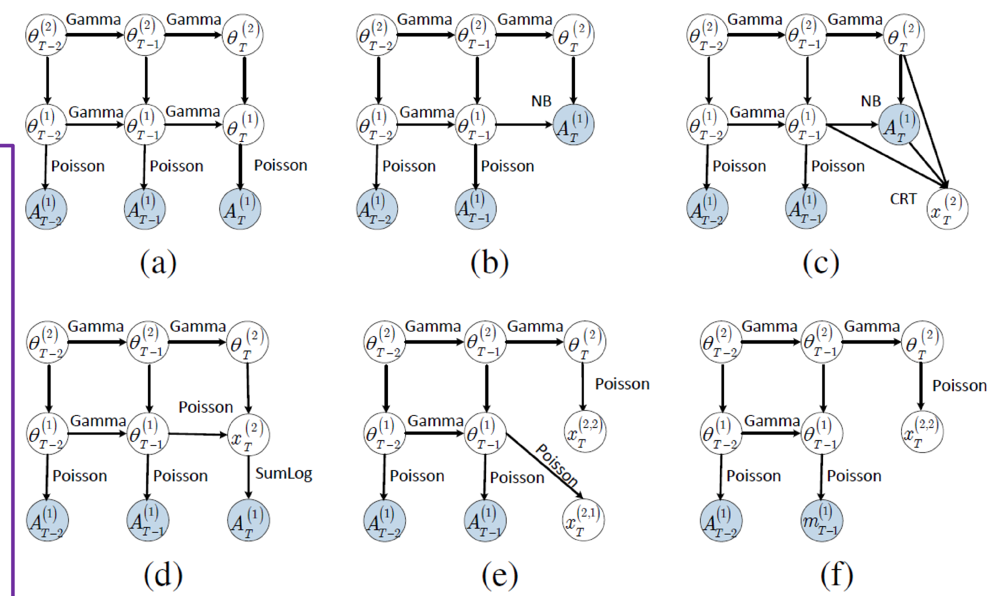
Inference

Backward and upward propagation of latent counts

Techniques:

- **Data augmentation**
- **Marginalization**

Repeat the process



Property 1 (P1): if $y_{\cdot kt} = \sum_{n=1}^{N} y_n$, where $y_n \sim \text{Pois}(\theta_n)$ are independent Poisson-distributed random variables, then $(y_1, \ldots, y_n) \sim \text{Multi}\left(y_{\cdot}, \frac{\theta_1}{\sum_{n=1}^N \theta_n}, \ldots, \frac{\theta_N}{\sum_{n=1}^N \theta_n}\right)$ and $y_{\cdot} \sim \text{Pois}(\sum_{n=1}^N \theta_n)$.

Property 2 (P2): $y \sim \text{Pois}(c\theta)$, where c is a constant, and $\theta \sim \text{Gam}(a,b)$ then $y \sim \text{NB}\left(a, \frac{c}{c+b}\right)$ is a negative binomialCdistributed random variable. We can equivalently parameterize it as $y \sim \text{NB}(a, g(\zeta))$, where $g(\zeta) = 1 - \exp(-\zeta)$ is the BernoulliCPoisson link and $\zeta = \ln\left(1 + \frac{c}{b}\right)$.

Property 3 (P3): if $y \sim \text{NB}(a, g(\zeta))$ and $l \sim \text{CRT}(y, a)$ is a Chinese restaurant table-distributed random variable, then y and l are equivalently jointly distributed as $y \sim \text{SumLog}(l, g(\zeta))$ and $l \sim \text{Pois}(a\zeta)$. **Refer to [1,2]**

- **Working backward for t = T, …, 2 and upward for l = 1,2,…,L, we draw**

$$(A_{k1t}^{(l)}, \ldots, A_{kK_lt}^{(l)}) \sim \text{Multi}\left(x_{kt}^{(l,l)}; \frac{\phi_{k1}^{(l)}\theta_{1t}^{(l)}}{\sum_{k_l=1}^{K_l}\phi_{kk_l}^{(l)}\theta_{k_lt}^{(l)}}, \ldots, \frac{\phi_{kK_l}^{(l)}\theta_{K_lt}^{(l)}}{\sum_{k_l=1}^{K_l}\phi_{kk_l}^{(l)}\theta_{k_lt}^{(l)}}\right),$$

$$x_{kt}^{(l+1)} \sim \text{CRT}\left[A_{\cdot kt}^{(l)} + Z_{\cdot k,t+1}^{(l)}, \tau_0\left(\sum_{k_{l+1}=1}^{K_{l+1}} \phi_{kk_{l+1}}^{(l+1)}\theta_{k_{l+1}t}^{(l+1)} + \sum_{k_l=1}^{K_l} \pi_{kk_1}^{(l)}\theta_{k_1,t-1}^{(l)}\right)\right]$$

$$(x_{kt}^{(l+1,l)}, x_{kt}^{(l+1,l+1)}) \sim \text{Multi}\left(x_{kt}^{(l+1)}, p_1/(p_1+p_2), p_2/(p_1+p_2)\right)$$

$$(Z_{k1t}^{(l)}, \ldots, Z_{kK_lt}^{(l)}) \sim \text{Multi}\left(x_{kt}^{(l+1,l)}; \frac{\pi_{k1}^{(l)}\theta_{1,t-1}^{(l)}}{\sum_{k_l=1}^{K_l}\pi_{kk_l}^{(l)}\theta_{k_l,t-1}^{(l)}}, \ldots, \frac{\pi_{kK_l}^{(l)}\theta_{K_l,t-1}^{(l)}}{\sum_{k_l=1}^{K_l}\pi_{kk_l}^{(l)}\theta_{k_l,t-1}^{(l)}}\right)$$

- **Working forward for t = 1,…,T and downward for l = L,…,1, we sample**

$$\theta_{kt}^{(l)} \sim \text{Gamma}\Big[A_{\cdot kt}^{(l)} + Z_{\cdot k(t+1)}^{(l)} + \tau_0\Big(\sum_{k_{l+1}=1}^{K_{l+1}} \phi_{kk_{l+1}}^{(l+1)}\theta_{k_{l+1}t}^{(l+1)} + \sum_{k_l=1}^{K_l} \pi_{kk_l}^{(l)}\theta_{k_2,t-1}^{(l)}\Big), \tau_0\big(1 + \zeta_t^{(l-1)} + \zeta_{t+1}^{(l)}\big)\Big],$$

Stochastic gradient MCMC inference for simplex

$$\left(\boldsymbol{\pi}_k^{(l)}\right)_{n+1} = \left[\left(\boldsymbol{\pi}_k^{(l)}\right)_n + \frac{\varepsilon_n}{M_k^{(l)}}\left[\left(\rho\tilde{\boldsymbol{z}}_{:k\cdot}^{(l)} + \boldsymbol{\eta}_{:k}^{(l)}\right) - \left(\rho\tilde{z}_{\cdot k\cdot}^{(l)} + \eta_{\cdot k}^{(l)}\right)\left(\boldsymbol{\pi}_k^{(l)}\right)_n\right] + \mathcal{N}\left(0, \frac{2\varepsilon_n}{M_k^{(l)}}\left[\text{diag}(\boldsymbol{\pi}_k^{(l)})_n - (\boldsymbol{\pi}_k^{(l)})_n(\boldsymbol{\pi}_k^{(l)})_n^T\right]\right)\right]_{\angle},$$ **Refer to [4]**

Algorithm 2 Stochastic-gradient MCMC for DPGDS

Input: Data mini-batches; Output: Global parameters of DPGDS.

for $i = 1, 2, \cdots$ **do**

  \ ⋆ *Collect local information*

  Backward-upward Gibbs sampling on the ith mini-batch for $\{A_{vkt}^{(l)}\}_{v,k,t}$; $\{x_{kt}^{(l+1)}\}_{k,t}$; $\{x_{kt}^{(l+1,l)}\}_{k,t}$; $\{x_{kt}^{(l+1,l+1)}\}_{k,t}$; $\{Z_{k_1k_2t}^{(l)}\}_{k_1,k_2,t}$ with (9)(10) (11)(12);

  Backward-upward calculating for the $\{\zeta_t^{(l)}\}_t$;

  Forward-downward Gibbs sampling for the $\{\boldsymbol{\theta}_t^{(l)}\}_t$ with (13);

  Sampling $\boldsymbol{\delta}^{(1)}$ with (17) or (18);

  \ ⋆ *Update global parameters*

  for $l = 1, 2, \cdots, L$ *and* $k = 1, 2, \cdots, K_L$ **do**

    Update $M_k^{(l)}$ according to [28]; then $\{\boldsymbol{\phi}_k^{(l)}\}_k$ with (35); Update $M_k^{(l)}$ according to [28]; then $\{\boldsymbol{\pi}_k^{(l)}\}_k$ with (34);

  end for

  Update $\xi^{(l)}$, $\{\nu_k^{(l)}\}_k$, and $\beta^{(l)}$ with SGNHT [21]

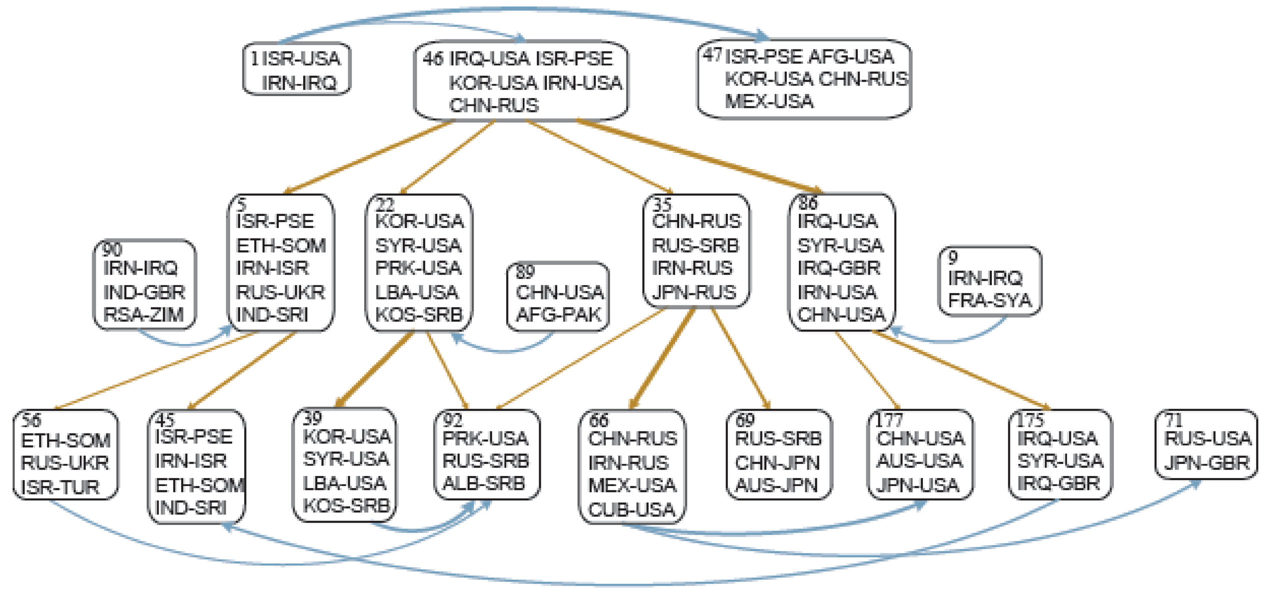
end for

Experiments

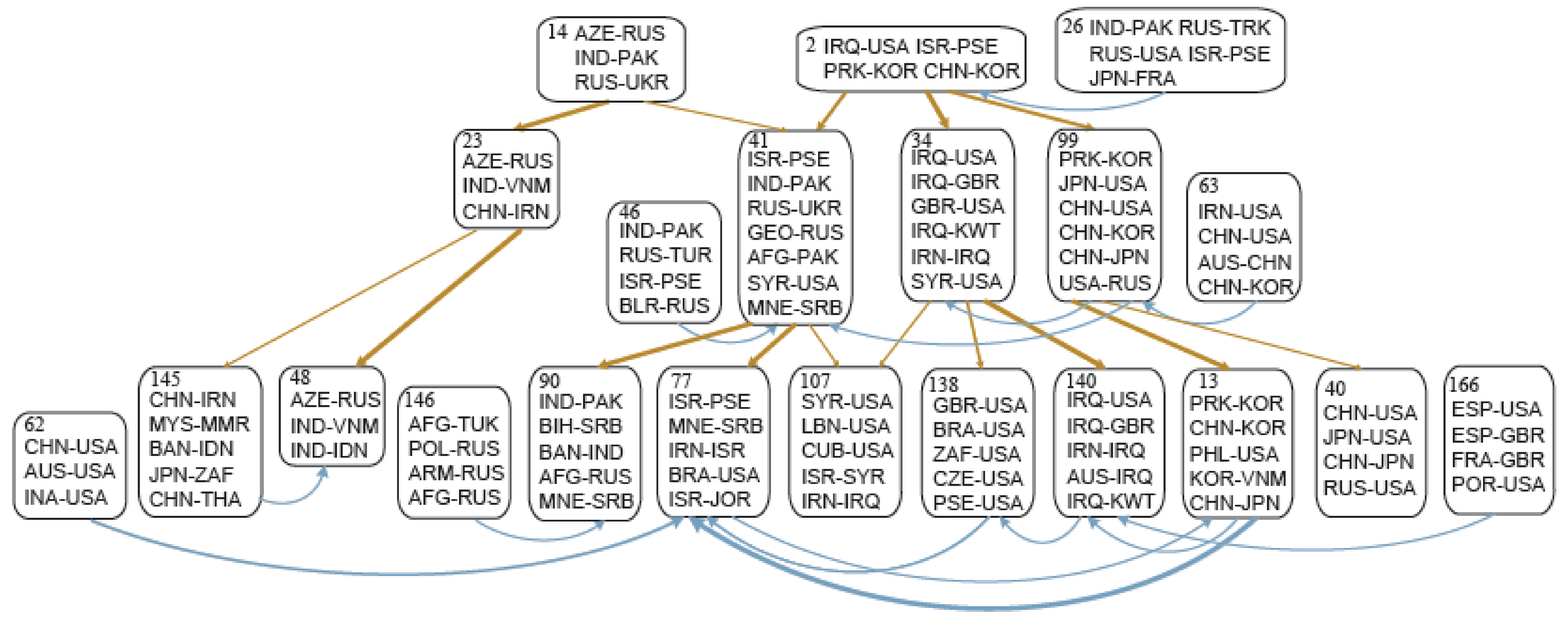
Top@M results on real-world text data

Model	Top@M	Data: GDELT ($T=365, V\approx 9000$)	ICEWS ($T=365, V\approx 3000$)	SOTU ($T=225, V=7518$)	DBLP ($T=14, V=1771$)	NIPS ($T=17, V=9836$)
GPDPFA	MP	0.611 ±0.001	0.607 ±0.002	0.379 ±0.002	**0.435** ±0.009	0.843 ±0.005
	MR	0.145 ±0.002	0.235 ±0.005	0.369 ±0.002	0.254 ±0.005	0.050 ±0.001
	PP	0.447 ±0.014	0.465 ±0.008	0.617 ±0.013	0.581 ±0.011	0.807 ±0.006
PGDS	MP	0.679 ±0.001	0.658 ±0.001	0.375 ±0.002	0.419 ±0.004	0.864 ±0.004
	MR	**0.150** ±0.001	**0.245** ±0.005	0.373 ±0.002	0.252 ±0.004	0.050 ±0.001
	PP	0.420 ±0.017	0.455 ±0.008	0.612 ±0.018	0.566 ±0.008	0.802 ±0.020
GPDM	MP	0.520±0.001	0.530 ±0.002	0.274 ±0.001	0.388 ±0.004	0.355 ±0.008
	MR	0.141 ±0.001	0.234 ±0.001	0.261 ±0.002	0.146 ±0.005	0.050 ±0.001
	PP	0.362±0.021	0.185±0.017	0.587 ±0.016	0.509 ±0.008	0.384 ±0.028
TSBN	MP	0.594 ±0.007	0.471 ±0.001	0.360 ±0.001	0.403 ±0.012	0.788 ±0.005
	MR	0.124 ±0.001	0.158 ±0.001	0.275 ±0.001	0.194 ±0.001	0.050 ±0.001
	PP	0.418 ±0.019	0.445 ±0.031	0.611 ±0.001	0.527 ±0.003	0.692 ±0.017
DTSBN-2	MP	0.439 ±0.001	0.475 ±0.002	0.370 ±0.004	0.407 ±0.003	0.756 ±0.001
	MR	0.134 ±0.001	0.208 ±0.001	0.361 ±0.001	0.248 ±0.007	0.050 ±0.001
	PP	0.391 ±0.001	0.446 ±0.001	0.587 ±0.027	0.522 ±0.005	0.737 ±0.004
DTSBN-3	MP	0.411 ±0.001	0.431 ±0.001	**0.450** ±0.008	0.390 ±0.002	0.774 ±0.002
	MR	0.141 ±0.001	0.189 ±0.001	0.274 ±0.001	0.252 ±0.004	0.050 ±0.001
	PP	0.367 ±0.011	0.451 ±0.026	0.548 ±0.013	0.510 ±0.006	0.715 ±0.009
DPGDS-2	MP	0.688 ±0.002	0.659 ±0.001	0.379 ±0.002	0.430 ±0.009	0.867 ±0.008
	MR	0.149 ±0.001	0.242 ±0.007	0.373 ±0.001	0.254 ±0.005	0.050 ±0.001
	PP	0.443 ±0.025	0.473 ±0.012	0.622 ±0.014	0.582 ±0.007	0.814 ±0.035
DPGDS-3	MP	**0.689** ±0.002	0.660 ±0.001	0.380 ±0.001	0.431 ±0.012	**0.887** ±0.002
	MR	**0.150** ±0.001	0.244 ±0.003	**0.374** ±0.002	**0.255** ±0.004	0.050 ±0.001
	PP	**0.456** ±0.015	**0.478** ±0.024	**0.628** ±0.021	**0.600** ±0.001	**0.839** ±0.007

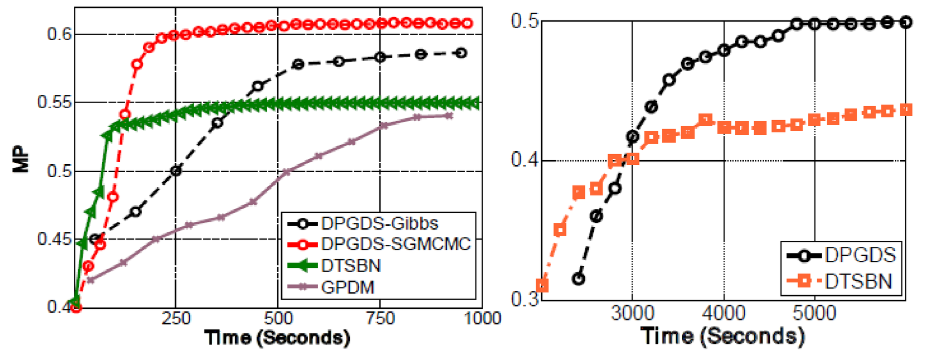
Hierarchical topics learned from ICEWS 2007-2009



Hierarchical topics learned from ICEWS 2001-2003



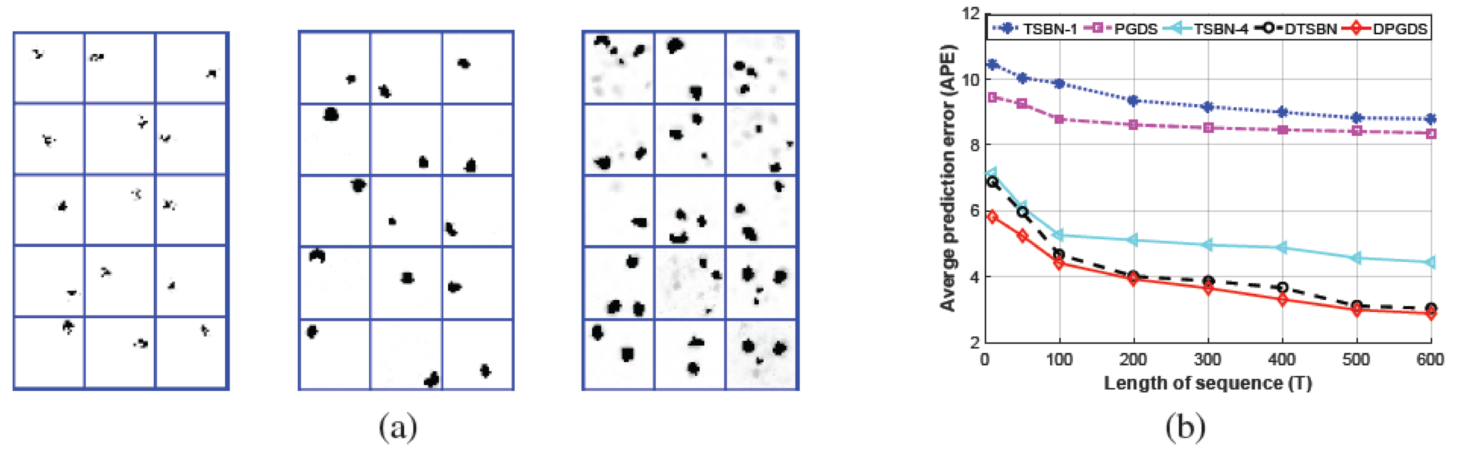
Scalability



GDELT 2001-2005, temporal granularity of 24 hrs, T=1825.

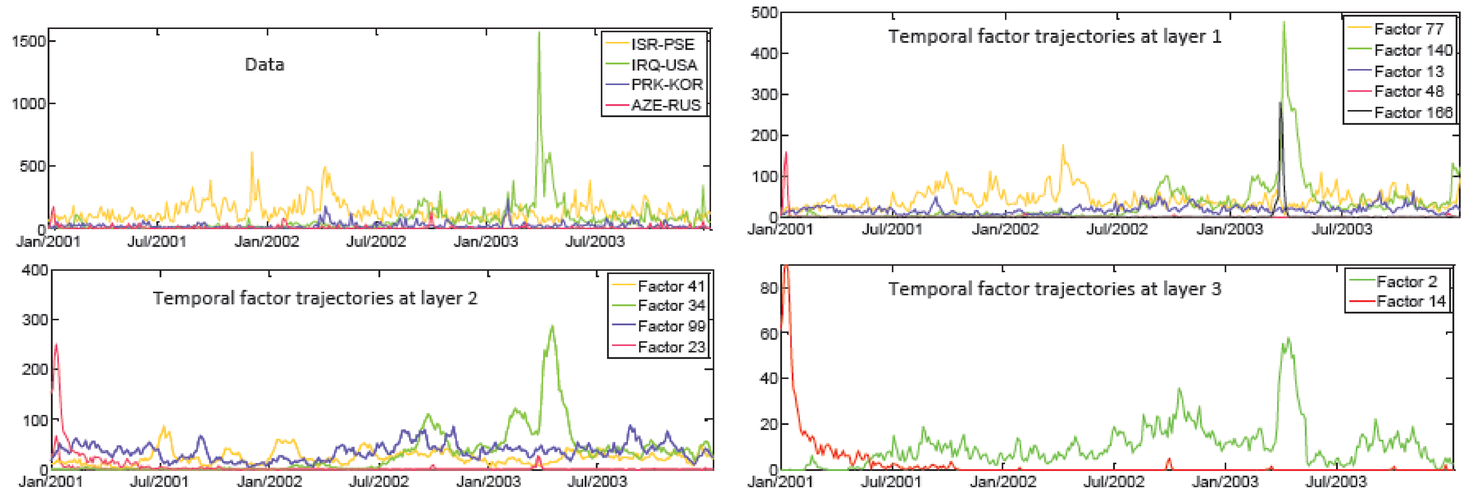
GDELT 2015-2018, temporal granularity of 15 mins, T=120000.

Shown in left and right are MP, as the function of time for GDELT 2001-2005, 2015-2018.

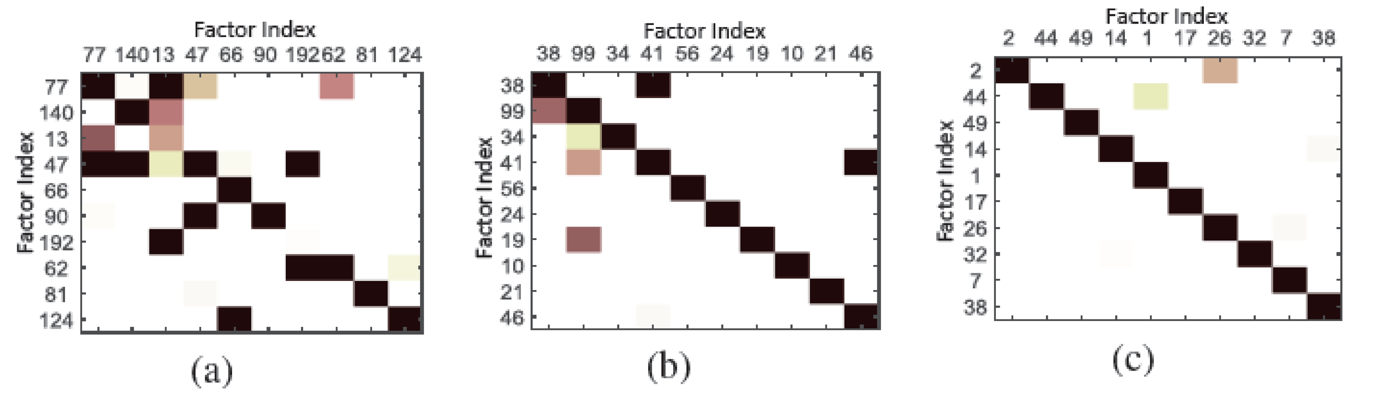


Results on the bouncing ball data set. (a) Top fifteen topics learned by three layer DPGDS at the layer 1, 2, 3. (b) APE as a function of the sequence length for various algorithms.

Temporal trajectories at different layers for ICEWS 2001-2003



Transition structure on ICEWS 2001-2003



Shown in (a)-(c) are transition matrices for layers 1, 2 and 3, respectively

[1] M. Zhou and L. Carin, "Negative binomial process count and mixture modeling," TPAMI, 2015.

[2] A. Schein, M. Zhou, and H., Wallach "Poisson–gamma dynamical systems," in NIPS, 2016.

[3] Z. Gan, C. Li, R. Henao, D. E. Carlson, and L. Carin, "Deep temporal sigmoid belief networks for sequence modeling," in NIPS, 2015.

[4] Y. Cong, B. Chen, H. Liu, and M. Zhou, "Deep latent Dirichlet allocation with topic-layer adaptive stochastic gradient Riemannian MCMC," in ICML, 2017.