SCALABLE INFERENCE OF TOPIC EVOLUTION VIA MODELS FOR LATENT GEOMETRIC STRUCTURES

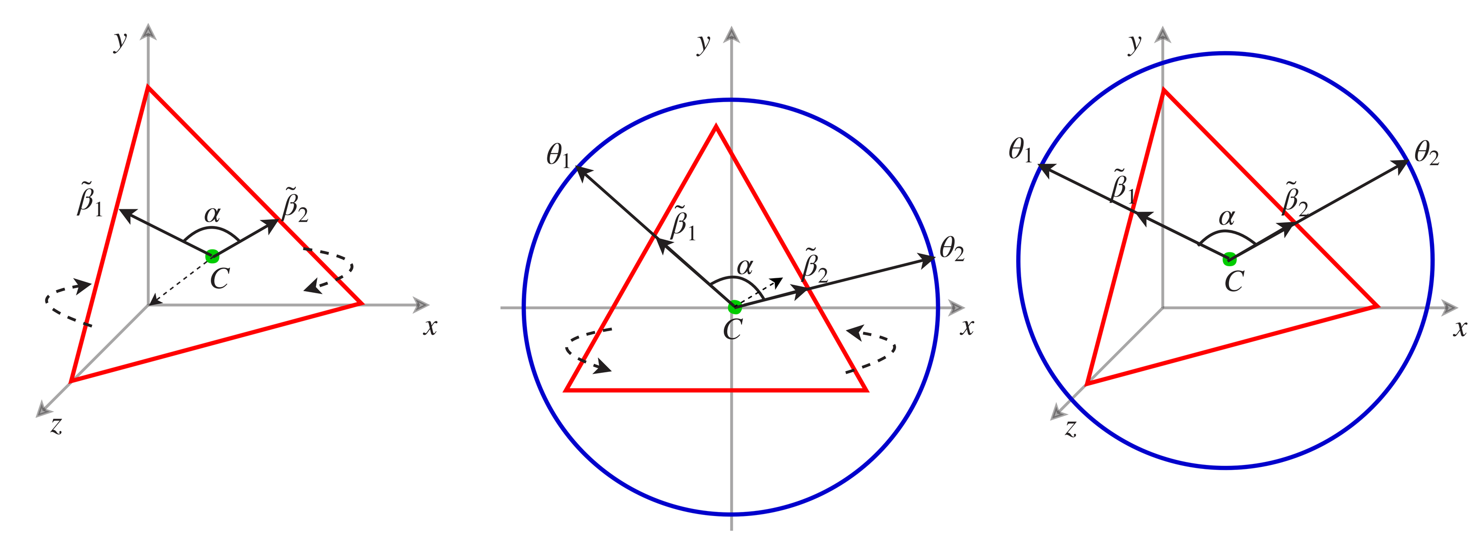
[1] Mikhail Yurochkin, [2] Zhiwei Fan,[3] Aritra Guha,[2] Paraschos Koutris, [3] XuanLong Nguyen
[1] IBM Research, [2] Department of Computer Science, University of Wisconsin-Madison, [3] Department of Statistics, University of Michigan

OVERVIEW

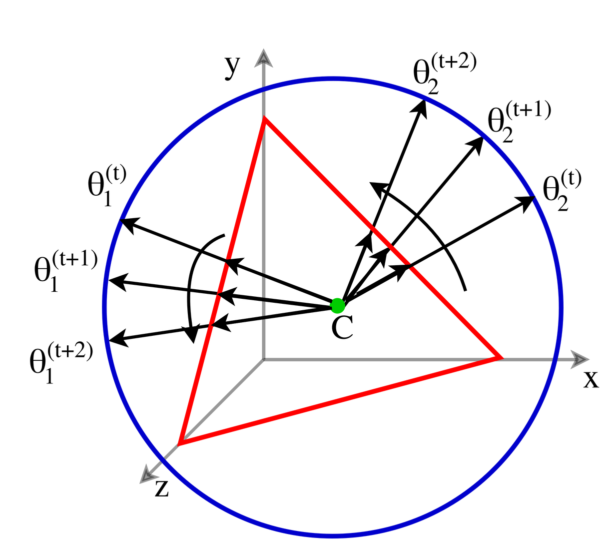
- series of Bayesian nonparametric models in increasing levels of complexity :
  - simpler model: topic polytope evolving over time
  - full model: temporal dynamics of topic polytope collection from multiple corpora
- scalable approximate inference algorithms suitable for online and distributed settings via the use of one-pass MAP estimates
- The Dynamic Topic Models (DTM) [Blei and Lafferty, 2006]:
  - lack of scalability
  - inefficient joint modeling at each time point and topic evolution over time
- solution: decoupling the two phases of inference.

DYNAMICS FOR SINGLE TOPIC POLYTOPE

Isometric embedding of topic polytope on sphere



Available metadata: $\{v_k^{(t)}\}_{t,k}$, topic estimates at each time t



$$
\begin{aligned}
Q &= \sum q_i \delta_{\theta_i} | \gamma_0, H \sim \mathrm{BI}\\
\theta_i &:= \{\theta_i^{(t)}\}_{t=1}^{T} \sim H\\
\theta_i^{(t)} | \theta_i^{(t-1)} &\sim \mathrm{vMF}(\theta_i^{(t-1)}, \tau_0) \text{ for}\\
\theta_i^{(0)} &\sim \mathrm{vMF}(\cdot, 0) - \text{uniforn}\\
\mathcal{T}^{(t)} &:= \sum_i b_i^{(t)} \delta_{\theta_i^{(t)}}, b_i^{(t)} | q_i \sim \mathrm{Bern}(q_i),\\
v_k^{(t)} | \mathcal{T}^{(t)} &\sim \mathrm{vMF}(\mathcal{T}_k^{(t)}, \tau_1) \text{ for } k = 1, \ldots, K^{(t)}
\end{aligned}
$$

STREAMING DYNAMIC MATCHING PROBLEM

- assign topics, $v_k^{(t)}$ at stage t to previously discovered topics, $\theta_i^{(t-1)}$ or attribute to new topics.
- Cost function for assigning topics (L_t number of topics at stage t, $m_i^(t)$ topic occupancy upto stage t)

STREAMING DYNAMIC MATCHING PROBLEM

- Objective function:

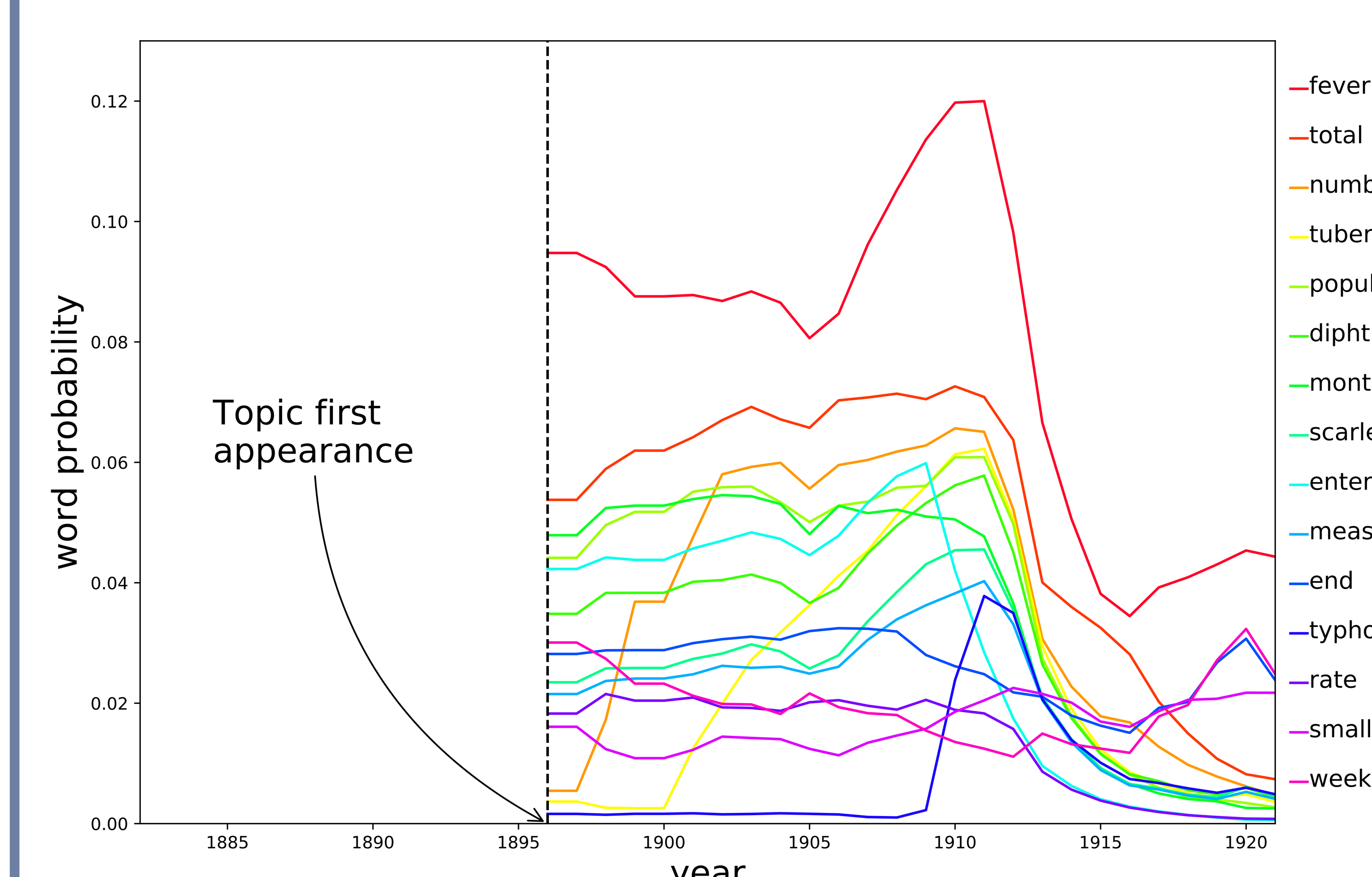
$$
C_{ik}^{(t)} = \begin{cases} \|\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2 - \tau_0 + \log \frac{m_i^{(t-1)}}{t - m_i^{(t-1)}}, & \text{if } i \text{ is a previous topic} \\ \tau_1 + \log \frac{\gamma_0}{t} - \log(i - L_{t-1}), & i \text{ new topic} \end{cases}
$$

- Solution:

$$
\theta_i^{(t)} = \begin{cases} \frac{\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}}{\|\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2}, & \text{if new topic } k \text{ is assigned to previously discovered topic } i \\ v_k^{(t)}, & \text{if topic } k \text{ is a new topic} \\ \theta_i^{(t-1)} & \text{if topic is dormant at } t \end{cases}
$$

EXPERIMENTAL RESULTS

- The Early Journal Content dataset years $1665 - 1922$, aggregated to single timepoint for SDM
- $400k$ scientific articles, vocabulary 4516 words.



SDM *Epidemics*: evolution of top 15 words

	Perplexity	Time	Topics	Cores
SDM	**1179**	22min	125	1
DM	1361	5min	125	20
SDDM	1241	**2.3min**	103	20
DTM	1194	56hours	100	1
SVB	1840	3hours	100	20
CoSAC	1191	51min	132	1

Modeling topics of EJC

EXPERIMENTAL RESULTS CONTINUED

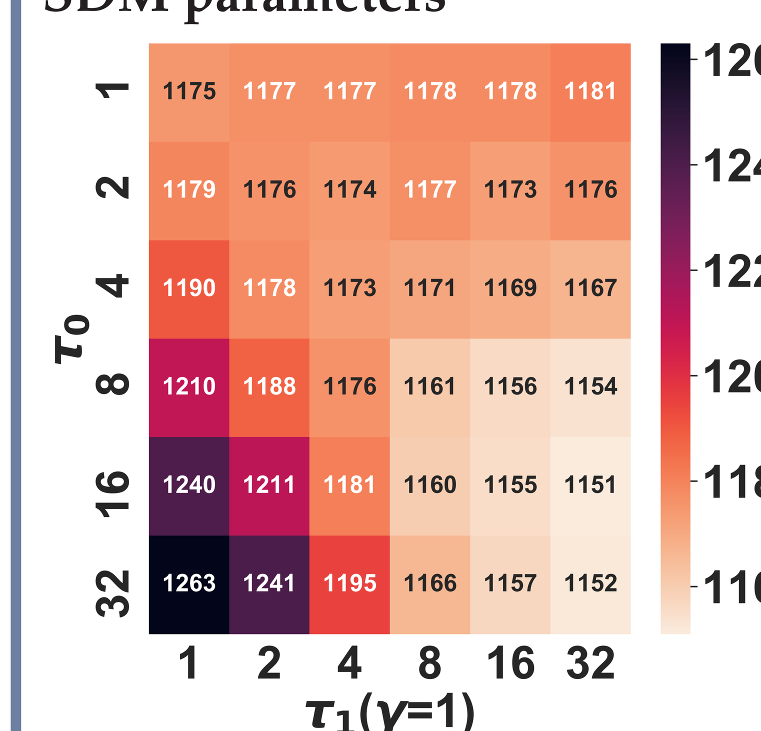
	Perplexity	Time	Topics	Cores
SDM	1254	2.4hours	182	1
DM	1260	**15min**	182	20
SDDM	**1201**	20min	238	20
DTM	NA	>72hours	100	1
SVB	1219	29.5hours	100	20
CoSAC	1227	4.4hours	173	1

Modeling Wikipedia articles

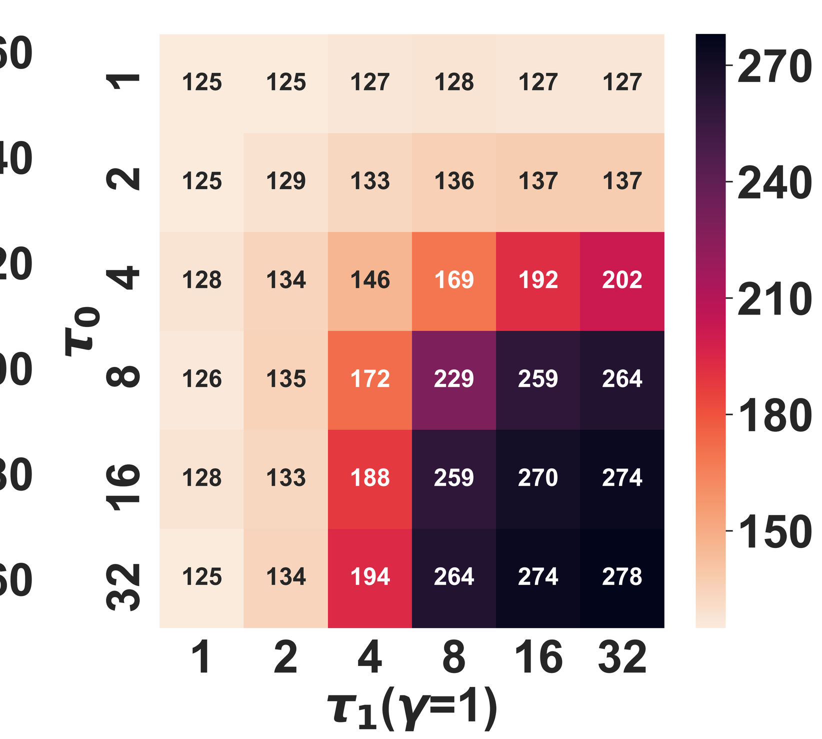
CHOICE OF PARAMETERS

- τ_0 controls rate of topic dynamics of the SDM, where smaller values imply higher dynamics rate
- Parameter τ_1 controls variance of local topics, when this variance is small, the model will tend to identify local topics as new global topics more often
- γ_0 affects the probability of new topic discovery

SDM parameters



(a)EJC perplexity



(b)# of topics in EJC

STREAMING DYNAMIC & DISTRIBUTED

- Global topic estimates $\theta_i^{(t)} = \frac{\tau_1 \sum_{j,k} B_{jik}^{(t)} v_{jk}^{(t)} + \tau_0 \theta_i^{(t-1)}}{\|\tau_1 \sum_{j,k} B_{jik}^{(t)} v_{jk}^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2}$.
- $B^{(t)}$ is assignment matrix at time t.
- At t, use CoSAC for noisy topic estimates of each group in parallel.

MORE MODEL FUSION

- Statistical model aggregation via parameter matching (NeurIPS 2019)
- Bayesian Nonparametric Federated Learning of Neural Networks (ICML 2019)