Twitter-Network Topic Model

A Full Bayesian Treatment for Social Network and Text Modeling

Kar Wai Lim (ANU & NICTA)
Changyou Chen (ANU & NICTA)
Wray Buntine (NICTA & ANU)

Contribution/Highlight

1. A fully Bayesian nonparametric topic model that models tweets very well.
2. A combination of the HPDP to model text, hashtags and authors, and the GP to jointly model authors and followers network.
3. Significantly outperform simpler nonparametric topic models.
4. Ablation study shows all components are significant.
5. Allows additional informative inference such as authors' interest, hashtags analysis.
6. Leads to further applications such as author recommendation, automatic topic labeling and hashtags suggestion.

Tweets

+

Followers Network

Author

(1 author per tweet)

+

Raw text
- Very short
- Unstructured
- Informal languages
- Grammatical errors
- Might contain spam

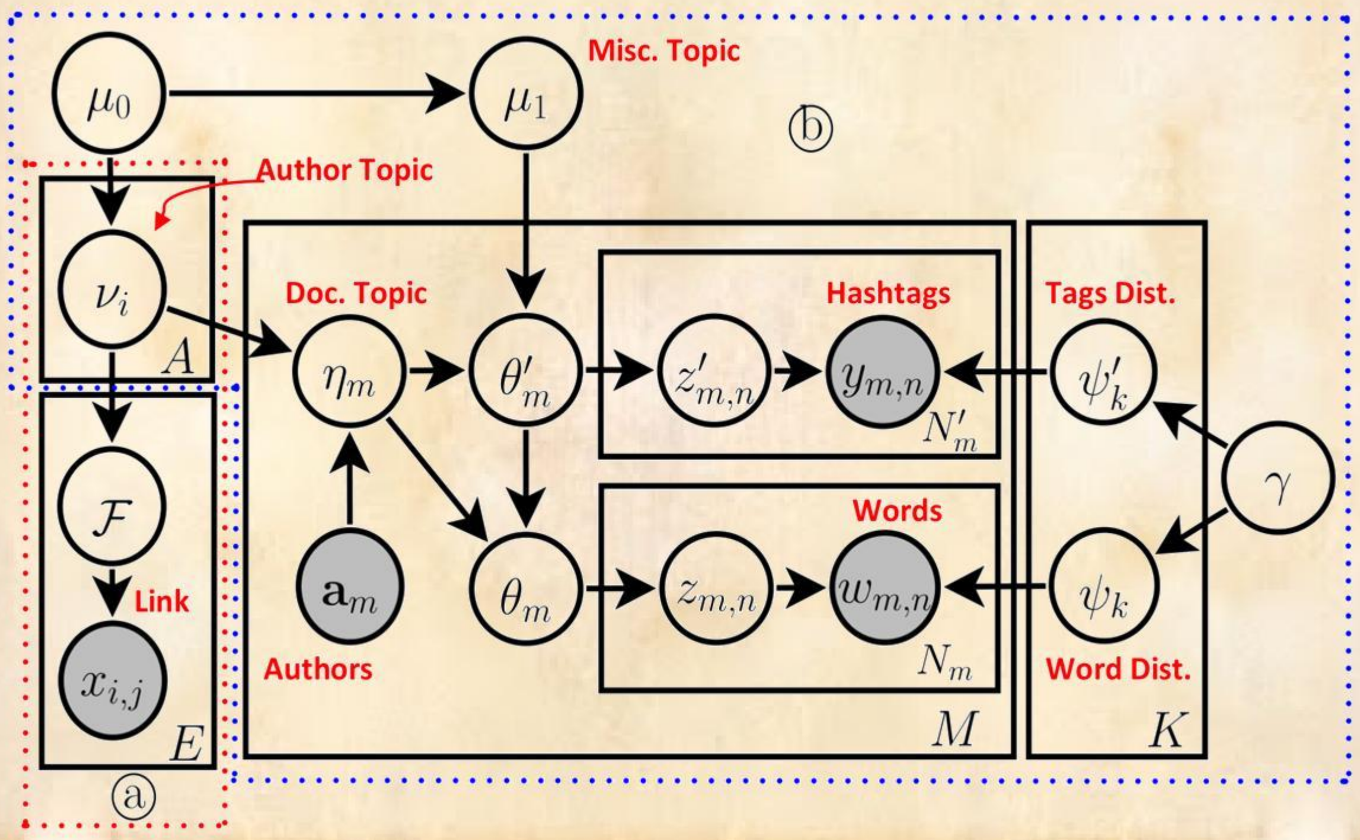
+

Hashtags
- Themes of tweets
- Noisy labels
- Not standardized
- Inappropriate usage

Hashtags Hijacking: when a hashtag is used for a different purpose than the one originally intended.

+

Graphical Model



Combining Text and Network

- **HPDP Topic Model (Region b)**
  - Jointly model text, hashtags and authorship.
  - A network of PDP nodes.
  - Explicitly model influence of hashtags to words.
  - Hashtags and words shared same tokens. (e.g. #happy is the same as happy)
- **GP Network Model (Region a)**
  - Jointly model the authors and the followers network with a GP random function model.

$$Q_{ij}|\nu_{1:A} \sim \mathcal{F}(\nu_i, \nu_j),$$
$$w_{ij}|Q_{ij} = \sigma(Q_{ij}),$$
$$x_{ij}|w_{ij} \sim \text{Bernoulli}(w_{ij})$$

  - Assume the followers network is bidirected.

Inference Algorithms

- **Collapsed Gibbs Sampling**
  - Jointly sample topics and table multiplicity for words and hashtags in the topic model.
  - Work generally with any Bayesian network of PDPs with no dynamic memory needed.
- **Metropolis Hastings Algorithm**
  - Jointly sample the author topic distribution and the followers network.
  - Use Elliptical Slice Sampler for the GP.
- **Hyperparameters Sampling**
  - Sample concentration parameters with the auxiliary variable sampler (Teh, 2006).

Automatic Topic Labeling

Table 2: Labeling Topics with Hashtags

	Top hashtags/words
T0	**#finance #money #economy** finance money bank marketwatch stocks china group
T1	**#politics #iranelection #tcot** politics iran iranelection tcot tlot topprog obama
T2	**#music #folk #pop** music folk monster head pop free indie album gratuit

- Hashtags can be good labels for topics.
- Previously unseen hashtags are candidates.
- Empirically, 90% of the proposed hashtags are good candidates as the topic labels.

Inference on Authors' Interest

- Summary of topics by different authors, where the topics are obvious from the Twitter ID.

Table 3: Inference on Authors' Interest

Twitter ID	Top topics represented by hashtags
finance_yard	#finance #money #realestate
ultimate_music	#music #ultimatemusiclist #mp3
seriouslytech	#technology #web #tech
seriouspolitics	#politics #postrank #news
pr_science	#science #news #postrank

Author Recommendation

- Recommend authors based on authors' topic distributions using the GP network model.
- Our proposed similarity kernel function is much better than the original kernel function.

Table 4: Cosine Similarity of Author Recommendation

Recommended	1st	2nd	3rd
Original	$0.00_{\pm 0.00}$	$0.05_{\pm 0.00}$	$0.06_{\pm 0.09}$
TN	$\mathbf{0.78}_{\pm 0.05}$	$\mathbf{0.57}_{\pm 0.10}$	$\mathbf{0.55}_{\pm 0.17}$
Not-recommended	1st	2nd	3rd
Original	$0.36_{\pm 0.05}$	$0.33_{\pm 0.05}$	$0.14_{\pm 0.07}$
TN	$\mathbf{0.17}_{\pm 0.03}$	$\mathbf{0.09}_{\pm 0.05}$	$\mathbf{0.10}_{\pm 0.08}$

Comparison and Ablation Study

- TN topic model significantly outperform HDP-LDA and a nonparametric Author-topic model.
- Ablation study shows that all components are significant.

Table 1: Test Perplexity & Network Likelihood

	Perplexity	Network
HDP-LDA	$358.1_{\pm 6.7}$	N/A
ATM	$302.9_{\pm 8.1}$	N/A
Random Function	N/A	$-294.6_{\pm 5.9}$
No Author	$243.8_{\pm 3.4}$	N/A
No Hashtag	$307.5_{\pm 8.3}$	$-269.2_{\pm 9.5}$
No μ_1 node	$221.3_{\pm 3.9}$	$-271.2_{\pm 5.2}$
No Word-tag link	$217.6_{\pm 6.3}$	$-275.0_{\pm 10.1}$
No Power-law	$222.5_{\pm 3.1}$	$-280.8_{\pm 15.4}$
No Network	$218.4_{\pm 4.0}$	N/A
TN Topic Model	$\mathbf{208.4}_{\pm 3.2}$	$\mathbf{-266.0}_{\pm 6.9}$

* Perplexity is calculated with left to right algorithm rather than document completion (Wallach et al., 2009).

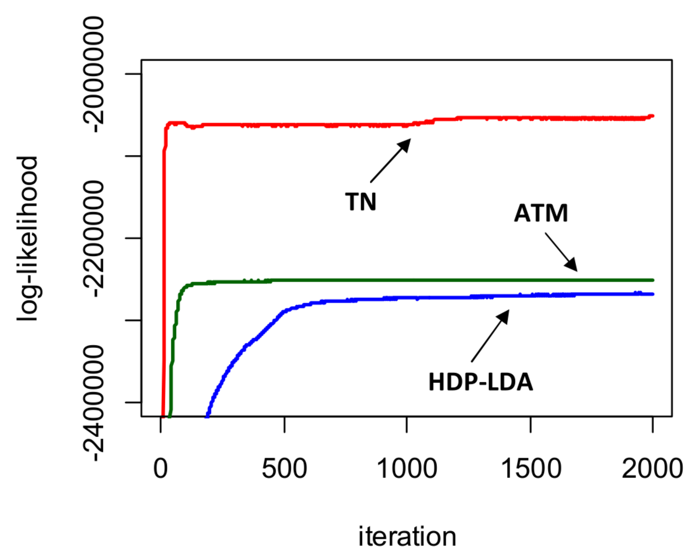


Figure 4: Training Log-likelihood vs. Iterations

Clustering and Topic Coherence

- TN topic model outperform state-of-the-art tweets pooling techniques (multiple tweets combined into a single document).
- Better performance in clustering measure (Purity and NMI) and topic coherence (PMI).

Table 5: Clustering and Topic Coherence Results

Methods	Purity			NMI Score			PMI score		
	Generic	Specific	Events	Generic	Specific	Events	Generic	Specific	Events
No pooling	0.49	0.64	0.69	0.28	0.22	0.39	−1.27	0.47	0.47
Author	0.54	0.62	0.60	0.24	0.17	0.41	0.21	0.79	0.51
Hourly	0.45	0.61	0.61	0.07	0.09	0.32	−1.31	0.87	0.22
Burstwise	0.42	0.60	0.64	0.18	0.16	0.33	0.48	0.74	0.58
Hashtag	0.54	**0.68**	0.71	0.28	0.23	0.42	0.78	**1.43**	1.07
TN	**0.66**	**0.68**	**0.79**	**0.43**	**0.31**	**0.52**	**0.79**	0.81	**1.66**

You can find the paper, poster and the supplementary material at the authors' websites.
Scanning the QR code on the right leads to the author's website.