Birds of a Feather Linked Together: A Discriminative Topic Model using Link-based Priors

Weiwei Yang
Computer Science
University of Maryland
College Park, MD
wwyang@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber@colorado.edu

Philip Resnik
Linguistics and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Abstract

A wide range of applications, from social media to scientific literature analysis, involve graphs in which documents are connected by links. We introduce a topic model for link prediction based on the intuition that linked documents will tend to have similar topic distributions, integrating a max-margin learning criterion and lexical term weights in the loss function. We validate our approach on the tweets from 2,000 Sina Weibo users and evaluate our model's reconstruction of the social network.

1 Introduction

Many application areas for text analysis involve documents connected by links of one or more types—for example, analysis of scientific papers (citations, co-authorship), Web pages (hyperlinks), legislation (co-sponsorship, citations), and social media (followers, mentions, etc.). In this paper we work within the widely used framework of topic modeling (Blei et al., 2003, LDA) to develop a model that is simple and intuitive, but which identifies high quality topics while also accurately predicting link structure.

Our work here is inspired by the phenomenon of *homophily*, the tendency of people to associate with others who are like themselves (McPherson et al., 2001). As manifested in social networks, the intuition is that people who are associated with one another are likely to discuss similar topics, and vice versa. The new topic model we propose therefore takes association links into account so that a document's topic distribution is influenced by the topic distributions of its neighbors. Specifically, we propose a joint model that uses link structure to define clusters (cliques) of documents and, following the intuition that documents in the same cluster are likely to have similar topic distributions, assigns each cluster its own separate Dirichlet prior over the cluster's topic distribution. This use of priors is consistent with previous work that has shown document-topic priors to be useful in encoding various types of prior knowledge and improving topic modeling performance (Mimno and McCallum, 2008). We then use distributed representations to "seed" the topic representations before getting down to modeling the documents. Our joint objective function uses a discriminative, max-margin approach (Zhu et al., 2012; Zhu et al., 2014) to both model the contents of documents and produce good predictions of links; in addition, it improves prediction by including lexical terms in the decision function (Nguyen et al., 2013).

Our baseline for comparison is the Relational Topic Model (Chang and Blei, 2010, henceforth RTM), which jointly captures topics and binary link indicators in a style similar to supervised LDA (McAuliffe and Blei, 2008, sLDA), instead of modeling links alone, e.g., as in the Latent Multi-group Membership Graph model (Kim and Leskovec, 2012, LMMG). We also compare our approach with Daumé III (2009), who uses document links to create a Markov random topic field (MRTF). Daumé does not, however, look at link prediction, as his upstream model (Mimno and McCallum, 2008) only generates documents *conditioned on* links. In contrast, our downstream model allows the prediction of links, like RTM.

Our model's primary contribution is in its novel combination of a straightforward joint modeling approach, max-margin learning, and exploitation of lexical information in both topic seeding and regression, yielding a simple but effective model for topic-informed discriminative link prediction. Like other topic models which treat binary values "probabilistically", our model can convert binary link indicators into non-zero weights, with potential application to improving models like Volkova

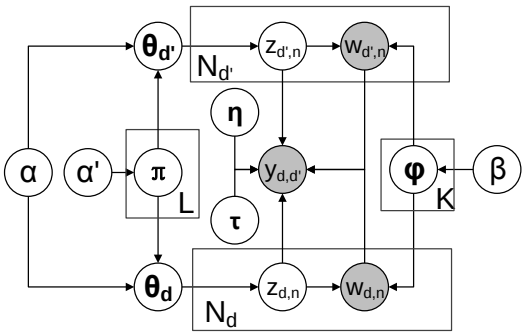

Figure 1: A graphical model of our model for two documents. The contribution of our model is the use of document clusters (π), the use of words (w) in the prediction of document links (y), and a max-margin objective.

et al. (2014), who use neighbor relationships to improve prediction of user-level attributes.

Our corpus is collected from Sina Weibo with three types of links between documents. We first conduct a reality check of our model against LDA and MRTF and then perform link prediction tasks. We demonstrate improvements in link prediction as measured by predictive link rank and provide both qualitative and quantitative perspectives on the improvements achieved by the model.

2 Discriminative Links from Topics

Figure 1 is a two-document segment of our model, which has the following generative process:

1. For each related-document cluster $l \in \{1, \ldots, L\}$
   Draw $\boldsymbol{\pi_l} \sim \text{Dir}(\alpha')$
2. For each topic $k \in \{1, \ldots, K\}$
   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
3. For each word $v \in \{1, \ldots, V\}$
   Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$
4. For each document $d \in \{1, \ldots, D\}$
   (a) Draw topic proportions $\boldsymbol{\theta_d} \sim \text{Dir}(\alpha \boldsymbol{\pi_{l_d}})$
   (b) For each word $t_{d,n}$ in document d
      i. Draw a topic assignment $z_{d,n} \sim \text{Mult}(\boldsymbol{\theta_d})$
      ii. Draw a word $t_{d,n} \sim \text{Mult}(\boldsymbol{\phi_{z_{d,n}}})$
5. For each linked pair of documents d and d'
   Draw binary link indicator
   $y_{d,d'} | \boldsymbol{z}_d, \boldsymbol{z}_{d'}, \boldsymbol{w}_d, \boldsymbol{w}_{d'} \sim \Psi(\cdot | \boldsymbol{z}_d, \boldsymbol{z}_{d'}, \boldsymbol{w}_d, \boldsymbol{w}_{d'}, \boldsymbol{\eta}, \boldsymbol{\tau})$

Step 1: Identifying birds of a feather. Prior to the generative process, given a training set of documents and document-to-document links, we begin by identifying small clusters or cliques using strongly connected components, which automatically determines the number of clusters from the link graph. Intuitively, documents in the same clique are likely to have similar topic distributions. Therefore, each of the L cliques l (the "birds of a feather" of our title) is assigned a separate Dirichlet prior $\boldsymbol{\pi_l}$ over K topics.

Step 2a: Using seed words to improve topic quality. To improve topic quality, we identify *seed words* for the K topics using distributed lexical representations: the key idea is to complement the more global information captured in LDA-style topics with representations based on local contextual information. We cluster the most frequent words' word2vec representations (Mikolov et al., 2013) into K word-clusters using the k-means algorithm, based on the training corpus.[1] We then enforce a one-to-one association between these discovered word clusters and the K topics. For any word token $w_{d,n}$ whose word type is in cluster k, the associated topic assignment $z_{d,n}$ can only be k. To choose topic k's seed words, within its word-cluster we compute each word $w_{k,i}$'s skip-gram transition probability sum $S_{k,i}$ to the other words as

$$S_{k,i} = \sum_{j=1, j \neq i}^{N_k} p(w_{k,j} \,|\, w_{k,i}), \qquad (1)$$

where N_k denotes the number of words in topic k.

We then select the three words with the highest sum of transition probabilities as the seed words for topic k. In the sampling process (Section 3), seed words are only assigned to their corresponding topics, similar to the use of hard constraints by Andrzejewski and Zhu (2009).

Steps 2b-3: Link regression parameters. Given two documents d and d', we want to predict whether they are linked by taking advantage of their topic patterns: the more similar two documents are, the more likely it is that they should be linked together. Like RTM, we will compute a regression in Step 5 using the topic distributions of d and d'; however, we follow Nguyen et al. (2013) by also including a document's word-level distribution as a regression input.[2] The regression value of document d and d' is

$$R_{d,d'} = \boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'}), \qquad (2)$$

where $\overline{\boldsymbol{z}}_d = \frac{1}{N_d} \sum_n z_{d,n}$, and $\overline{\boldsymbol{w}}_d = \frac{1}{N_d} \sum_n w_{d,n}$; $\circ$ denotes the Hadamard product; $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are the

[1] In the experiment, seed words must appear at least 1,000 times.

[2] Both approaches contrast with the links-only approach of Kim and Leskovec (2012).

weight vectors for topic-based and lexically-based predictions, respectively.

Step 4: Generating documents. Documents are generated as in LDA, where each document's topic distribution $\boldsymbol{\theta}$ is drawn from the cluster's topic prior (a parametric analog to the HDP of Teh et al. (2006)) and each word's topic assignment is drawn from the document's topic distribution (except for seed words, as described above).

Step 5: Generating links. Our model is a "downstream" supervised topic model, i.e., the prediction of the observable variable (here, document links) is informed by the documents' topic distributions, as in sLDA (Blei and McAuliffe, 2007). In contrast to Chang and Blei (2010), who use a sigmoid as their link prediction function Ψ, we instead use hinge loss: the probability Ψ that two documents d and d' are linked is

$$p(y_{d,d'} = 1 \mid \overline{\boldsymbol{z}}_d, \overline{\boldsymbol{z}}_{d'}, \overline{\boldsymbol{w}}_d, \overline{\boldsymbol{w}}_{d'}) = \exp(-2c\max(0, \zeta_{d,d'})),$$

where c is the regularization parameter. In the hinge loss function, $\zeta_{d,d'}$ is

$$\zeta_{d,d'} = 1 - y_{d,d'} R_{d,d'}. \quad (3)$$

3 Posterior Inference

Sampling Topics. Following Polson and Scott (2011), by introducing an auxiliary variable $\lambda_{d,d'}$, we derive the conditional probability of a topic assignment

$$p(z_{d,n} = k \mid \boldsymbol{z}_{-d,n}, \boldsymbol{w}_{-d,n}, w_{d,n} = v) \propto \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \times (N_{d,k}^{-d,n} + \alpha\pi_{l_d,k}^{-d,n}) \times \prod_{d'} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \quad (4)$$

where $N_{k,v}$ denotes the count of word v assigned to topic k; $N_{d,k}$ is the number of tokens in document d that are assigned to topic k.[3] Marginal counts are denoted by $\cdot$; $^{-d,n}$ denotes that the count excludes token n in document d; d' denotes the indexes of documents which are linked to document d; $\pi_{l_d,k}^{-d,n}$ is estimated based on the maximal path assumption (Wallach, 2008)

$$\pi_{l_d,k}^{-d,n} = \frac{\sum_{d' \in S(l_d)} N_{d',k}^{-d,n} + \alpha'}{\sum_{d' \in S(l_d)} N_{d',\cdot}^{-d,n} + K\alpha'}, \quad (5)$$

where $S(l_d)$ denotes the cluster which contains document d (Step 1 in the generative process).

[3]More details here and throughout this section appear in the supplementary materials.

Optimizing topic and lexical regression parameters. While topic regression parameters $\boldsymbol{\eta}$ and lexical regression parameters $\boldsymbol{\tau}$ can be sampled (Zhu et al., 2014), the associated covariance matrix is huge (approximately $12K \times 12K$ in our experiments). Instead, we optimize these parameters using L-BFGS.

Sampling auxiliary variables. The likelihood of auxiliary variables $\boldsymbol{\lambda}$ follows a generalized inverse Gaussian distribution $\mathrm{GIG}(\lambda_{d,d'}; \frac{1}{2}, 1, c^2\zeta_{d,d'}^2)$. Thus we sample $\lambda_{d,d'}^{-1}$ from a an inverse Gaussian distribution

$$p(\lambda_{d,d'}^{-1} \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\tau}) = \mathrm{IG}\left(\lambda_{d,d'}^{-1}; \frac{1}{c|\zeta_{d,d'}|}, 1\right). \quad (6)$$

4 Experimental Results

4.1 Dataset

We crawl data from Sina Weibo, the largest Chinese micro-blog platform. The dataset contains 2,000 randomly-selected verified users, each represented by a single document aggregating all the user's posts. We also crawl links between pairs of users when both are in our dataset. Links correspond to three types of interactions on Weibo: mentioning, retweeting and following.[4]

4.2 Perplexity Results

As an initial reality check, we first apply a simplified version of our model which only uses user interactions for topic modeling and does not predict links. This permits a direct comparison of our model's performance against LDA and Markov random topic fields (Daumé III, 2009, MRTF) by evaluating perplexity.

We set $\alpha = \alpha' = 15$ and run the models on 20 topics for all models in this and following sections. The results are the average values of five independent runs. Following Daumé, in each run, for each document, 80% of its tokens are randomly selected for training and the remaining 20% are for test. As the training corpus is generated randomly, seeding is not applied in this section. The results are given in Table 1, where I- denotes that the model incorporates user interactions.

The results confirm that our model outperforms both LDA and MRTF and that its use of user interactions holds promise.

[4]We use ICTCLAS (Zhang et al., 2003) for segmentation. After stopword and low-frequency word removal, the vocabulary includes 12,257 words, with ∼755 tokens per document and 5,404 links.

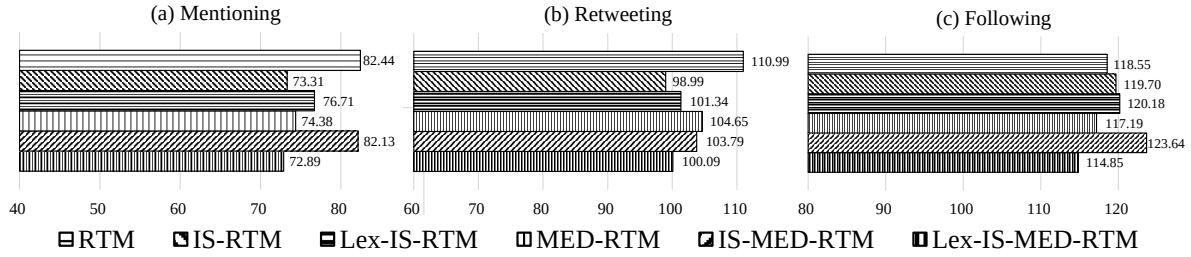

Figure 2: Lex-IS-MED-RTM, combining all three extensions, performs the best on predicting mentioning and following links, although IS-RTM achieves a close value on mentioning links and even a slightly better value on retweeting links. User interactions (denoted by "I") sometimes bring down the performance, as cluster priors are not applied in this intrinsic evaluation.

Link	Model	Perplexity
–	LDA	2605.06
Mentioning	MRTF	2582.08
	I-LDA	2522.58
Retweeting	MRTF	2588.30
	I-LDA	2519.27
Following	MRTF	2587.26
	I-LDA	2530.67

Table 1: Our simplified model I-LDA achieves lower perplexities than both LDA and MRTF, by incorporating different cliques extracted from three types of user interactions.

4.3 Link Prediction Results

In this section, we apply our model on link prediction tasks and evaluate by predictive link rank (PLR). A document's PLR is the average rank, among all documents, of the documents to which it actually links. This means that lower values of PLR are better.

Figure 2 breaks out the 5-fold cross validation results and the distinct extensions of RTM.[5] The results support the value in combining all three extensions using Lex-IS-MED-RTM, although for mentioning and retweeting, Lex-IS-MED-RTM and IS-RTM are quite close.

Applying user interactions does not always produce improvements. This is because in our intrinsic evaluation, we assume that the links on the test set are not observable and cluster priors are not applied. However, according to the training performance (extrinsic evaluations which we are still in progress), user interactions do benefit link prediction performance when links are partially available, e.g., suggesting more links based on observed links. In contrast, hinge loss and lexical term weights do not depend on metadata availability and generally produce improvements in link prediction performance.

4.4 Illustrative Example

We illustrate model behavior qualitatively by looking at two test set users, designated A and B. User A is a reporter who runs "We Media" on his account, sending news items to followers, and B is a consultant with a wide range of interests. Their tweets reveal that both are interested in social news—a topic emphasizing words like *society*, *country*, *government*, *laws*, *leaders*, *political party*, *news*, etc. Both often retweet news related to unfairness in society and local government scandals (*government*, *police*, *leaders*, *party*, *policy*, *chief secretary*). For example, User A retweeted a report that a person about to be executed was unable to take a photo with his family before his execution, writing *I feel heartbroken*. User B retweeted news that a mayor was fired and investigated because of a bribe; in his retweet, he expresses his dissatisfaction with what the mayor did when he was in power. In addition, User A follows new technology (*smart phone*, *Apple*, *Samsung*, *software*, *hardware*, etc.) and B is interested in food (*snacks*, *noodles*, *wine*, *fish*, etc.).

As ground truth, there is a *mentioning* link from A to B; Table 2 shows this link's PLR in the *mentioning* models, which generally improves with model sophistication. The mentioning tweet is a news item that is consistent with the model's

[5] IS- denotes that the model incorporates user interactions and seed words, Lex- means that lexical terms were included in the link probability function (Equation 3), and MED- denotes max-margin learning (Zhu et al., 2014; Zhu et al., 2012). Each type of link is applied separately; e.g., in Figure 2(a) results are based *only* on mentioning links, ignoring retweeting and following links.

Model		RTM	IS-RTM	Lex-IS-RTM	MED-RTM	IS-MED-RTM	Lex-IS-MED-RTM
PLR of the Link		24	10	9	74	18	26
Social News	**User A**	0.018	0.021	0.034	0.016	0.027	0.030
Topic Proportion	**User B**	0.309	0.413	0.408	0.318	0.355	0.392

Table 2: Data for Illustrative Example

Model		RTM	IS-RTM	Lex-IS-RTM	MED-RTM	IS-MED-RTM	Lex-IS-MED-RTM
Topic PMI		1.186	1.224	1.216	1.214	1.294	1.229
Average Regression Values	**Linked Pairs**	0.2403	0.3692	0.4031	0.7220	0.6321	0.7668
	All Pairs	0.06636	0.07729	0.08020	0.2482	0.2041	0.2428
	Ratio	3.621	4.777	5.026	2.909	3.097	3.158
	SD/Avg	0.9415	1.2081	1.2671	0.6364	0.7254	0.7353

Table 3: Values for Quantitative Analysis

characterization of the users' interests (particularly social news and technology): a Samsung Galaxy S4 exploded and caused a fire while charging. Consistent with intuition, the prevalence of the social news topic also generally increases as the models grow more sophisticated.[6]

4.5 Quantitative Analysis

Topic Quality. Automatic coherence detection (Lau et al., 2014) is an alternative to manual evaluations of topic quality (Chang et al., 2009). In each topic, the top n words' average pointwise mutual information (PMI)—based on a reference corpus—serves as a measure of topic coherence.[7]

Topic quality improves with user interactions and max-margin learning (Table 3). PMI drops when lexical terms are added to the link probability function, however. This is consistent with the role of lexical terms in the model; their purpose is to improve link prediction performance, not improve topic quality.

Average Regression Value. One way to assess the quality of link prediction is to compare the scores of (ground-truth) linked documents to documents in general. In Table 3, the Average Regression Values show this comparison as a ratio. The higher the ratio, the more linked document pairs differ from unlinked pairs, which means that linked documents are easier to distinguish. This ratio improves as RTM extensions are added, indicating better link modeling quality.

In the SD/Avg row of Table 3, we also compute a ratio of standard deviations to mean values. Ratios given by the models with hinge loss are lower than those not using hinge loss. This means that the regression values given by the models with hinge loss are more concentrated around the average value, suggesting that these models can better identify linked pairs, even though the ratio of linked pairs' average regression value to all pairs' average value is lower.

5 Conclusions and Future Work

We introduce a new topic model that takes advantage of document links, incorporating link information straightforwardly by deriving clusters from the link graph and assigning each cluster a separate Dirichlet prior. We also take advantage of locally-derived distributed representations to "seed" the model's latent topics in an informed way, and we integrate max-margin prediction and lexical regression to improve link prediction quality. Our quantitative results show improvements in predictive link rank, and our qualitative and quantitative analysis illustrate that the model's behavior is intuitively plausible.

In future work, we plan to engage in further model analysis and comparison, to explore alterations to model structure, e.g. introducing hierarchical topic models, to use other clustering methods to obtain priors, and to explore the value of predicted links for downstream tasks such as friend recommendation (Pennacchiotti and Gurumurthy, 2011) and inference of user attributes (Volkova et al., 2014).

[6] Numerically its proportion is consistently lower for User A, whose interests are more diverse.

[7] We set $n = 20$ and use a reference corpus of 1,143,525 news items from Sogou Lab, comprising items from June to July 2012, `http://www.sogou.com/labs/dl/ca.html`. Each averages ∼347 tokens, using the same segmentation scheme as the experimental corpus.

Acknowledgements

We thank Hal Daumé III for providing his code. This work was supported in part by NSF award 1211153. Boyd-Graber is supported by NSF Grants CCF-1409287, IIS-1320538, and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet allocation with topic-in-set knowledge. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, pages 124–150.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Hal Daumé III. 2009. Markov random topic fields. In *Proceedings of the Association for Computational Linguistics*.

Myunghwan Kim and Jure Leskovec. 2012. Latent multi-group membership graph model. In *Proceedings of the International Conference of Machine Learning*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

David M. Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*.

Viet-An Nguyen, Jordan L. Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*.

Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *Proceedings of World Wide Web Conference*.

Nicholas G. Polson and Steven L. Scott. 2011. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the Association for Computational Linguistics*.

Hanna M. Wallach. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2012. MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(1):2237–2278.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2014. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, 15(1).