Topic Models Conditioned on Relations

Mirwaes Wahabzada, Zhao Xu, and Kristian Kersting

Knowledge Discovery Department
Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
firstname.lastname@iais.fraunhofer.de

Abstract. Latent Dirichlet allocation is a fully generative statistical language model that has been proven to be successful in capturing both the content and the topics of a corpus of documents. Recently, it was even shown that relations among documents such as hyper-links or citations allow one to share information between documents and in turn to improve topic generation. Although fully generative, in many situations we are actually not interested in predicting relations among documents. In this paper, we therefore present a Dirichlet-multinomial nonparametric regression topic model that includes a Gaussian process prior on joint document and topic distributions that is a function of document relations. On networks of scientific abstracts and of Wikipedia documents we show that this approach meets or exceeds the performance of several baseline topic models.

1 Introduction

One of the most fundamental problems in information retrieval is the extraction of meaningful, low-dimensional representations of data. In computer vision, where it is natural to represent images as vectors in a high-dimensional space, they represent e.g. visual words and have been used for face and object recognition or color classification. Social networks such as Flickr, Facebook and Myspace, allow for a diverse range of interactions amongst their members, resulting in temporal datasets relating users, media objects and actions. Here, low-dimensional representations may be used to identify and summarize social activities. If the data are words of documents, low-dimensional representations yield topic models representing each document as a mixture of a small number of topics and each word is attributable to one of the topics.

Topic models, originally explored by Deerwester *et al.* [9], Hofmann [13], Blei *et al.* [5], Griffiths and Steyvers [11], Buntine and Jakulin [6], and many others, have received a lot of attention due to their simplicity, usefulness in reducing the dimensionality of the data, and ability to produce interpretable and semantically coherent topics. They are typically fully generative, probabilistic models that uncover the underlying semantic structure of a document collection based on an hierarchical Bayesian analysis, see e.g. [3], and have been proven successful in a number of applications such as analyzing emails [19], classifying natural scenes [16], detecting topic evolutions of a document corpus [4,31], analyzing the human semantic memory [30], and many more.

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 402–417, 2010.

The starting point for our analysis here is an often perceived limitation of latent Dirichlet allocation (LDA), which is one of the most popular and commonly used topic models today [5]: *it fails to make use of relations among documents.* Nowadays, however, networks of documents such as citation networks of scientific papers, hyperlinked networks of web pages, and social networks of friends are becoming pervasive in machine learning applications. Consider a collection of hyperlinked webpages. LDA uses the word distribution within the body of each page only to determine a per-document topic distribution. More precisely, LDA models each document as a mixture over topics, where each vector of per-document mixture proportions is assumed to have been drawn from a Dirichlet distribution with the hyperparameters $\boldsymbol{\alpha}$ shared among all documents. Webpages, however, do not exist in isolation: there are links connecting them. Two pages having a common set of links are evidence for similarity between such pages. For instance, if W1 and W2 both link to W3, this is commonly considered to be evidence for W1 and W2 having similar topic distributions. In other words, relational knowledge can further reveal additional correlations between variables of interest such as topics. Therefore, it is not surprising that several fully generative relational topic models have been proposed recently, see e.g. [2,12,20,23,7], that take correlations among inter-related documents into account. They have been proven to be successful for modeling networks of documents and even for predicting relations among documents. However, adding additional complexity such as relations to a fully generative model generally results in a larger number of variables to sample and in turn in a more complicated sampling distribution. Thus, the flexibility of relational topic models comes at the cost of increasingly intractable inference. If we are actually not interested in predicting relations, this is an unnecessary complication.

Our main contribution is a novel relational topic model, called xLDA. At the expense of not being able to predict relations among documents anymore, we condition topic models on the metadata such as the relations among the documents (citations, hyperlinks, and so on) as well as attributes describing each document d (authors, year, venue, and so on) provided in the data. Specifically, xLDA is a Dirichlet-multinomial (nonparametric) regression topic model that includes a Gaussian process prior on joint document and topic distributions that is a function of document attributes and relations. That is, given metadata such as relations we generate a per-document $\boldsymbol{\alpha}_d$, the (hyper-)parameters of a Dirichlet distribution. Then, we model each document using LDA with the generated $\boldsymbol{\alpha}_d$. Intuitively, documents from the same authors, or published in the same conference, or being related by citations are stronger correlated than other documents. The more correlated two documents are, the more likely they have similar topics. Because all the relational information is accounted for in the document-specific Dirichlet hyperparameters $\boldsymbol{\alpha}_d$, the sampling phase of xLDA is no more complicated than a simple LDA sampler. In other words, we sacrifice flexibility for a relatively simple inference. Moreover, we can extend the basic xLDA model through topic meta-information that allows us to express or even to learn conditional independencies that cannot be explained well by the document

meta-information only. On networks of scientific abstracts and of Wikipedia documents we show that xLDA meets or exceeds the performance of several baseline topic models.

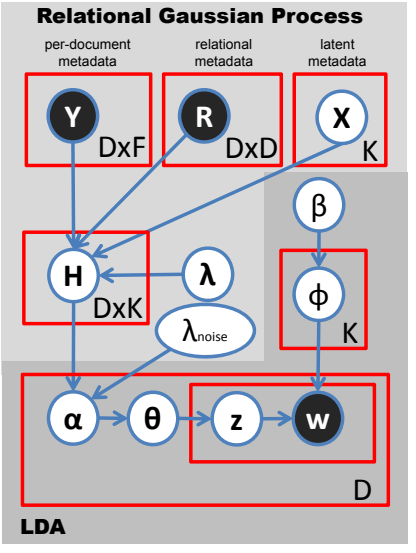
We proceed as follows. After touching upon further related work, we will introduce the xLDA model in Section 3. In Section 4, we then discuss its approximate inference and learning methods. Before concluding, we present our experimental evaluation.

2 Related Work

Network data is currently receiving a lot of attention. Several latent variable models that decompose the network according to hidden patterns of connections between its nodes have been proposed, see e.g. [33,15,1]. Indeed quite powerful, these models mainly account for the structure of the network, essentially ignoring the observed attributes of the nodes. Relational matrix factorization approaches such as [28,17,14] are not tailored towards discovering topics.

Recently, several relational topic models have been proposed that also take the observed attributes of nodes, i.e., documents into account [2,12,20,23,7]. They are all fully generative models and due to the additional relations modeled have a more complicated sampling distribution. If we are not interested in predicting relations, this is an unnecessary complication and conditioning on the relations is an attractive alternative.

The idea of conditioning topic models on metadata is not new. Several models have been proposed in which a hidden variable selects one of several topic models conditioned on some metadata. For instance, Rosen-Zvi *et al.*'s author-topic models [26] generates words by first selecting an author uniformly from an observed author list and then selecting a topic from a distribution over topics that is specific to that author. Mimno and McCallum [21] extend this author-topic model to the author-persona topic model that associates multiple topical mixtures with each individual author. McCallum *et al.* [18] employ the "conditioning" idea to model authors and recipients of email, and Dietz *et al.* [10] use it for inferring the influence of individual references on citing papers. Recently, Mimno and McCallum [22] introduced the Dirichlet-multinomial regression topic model. It includes a log-linear prior on document-topic distributions that is a function of observed features of the document, such as author, publication venue, references, and dates. An investigation of this model was the seed that grew into the current paper. It is important, however, to distinguish xLDA from Mimno and McCallum's model. Whereas Mimno and McCallum proposed to model relational information such as citations as per-document attributes of a log-normal prior with diagonal covariance, xLDA employs Silva *et al.*'s [27] *directed mixed graph* Gaussian process framework to incorporate relational information into the topic model. A directed mixed graph model propagates training data information through other training points. Reconsider our webpage domain where each page may have links to several other pages. A chain of intermediated pages between two pages W1 and W2 is likely to be more informative if we know the



Symbol	Description
$y_d \in Y$	observed attribute vector of a document d
$r_d \in R$	observed relation vector of a document d
$x_k \in X$	latent metainformation vector
$\eta_{dk} \in H$	noise-free topic k concentration for document d
λ, λ_{noise}	hyperparameters of covariance
β	prior belief on the distribution over the vocabulary
α_d	prior belief on topic proportions, $\alpha_d = \exp(\eta_d + \epsilon_{noise}) = \exp(\tau_d)$
$\phi_k \in \Phi$	preference of a topic k over the vocabulary with $\sum_n \phi_{k,n} = 1$
θ_d	topic proportions of a document
$W_{d,n} \in W$	n'th word in the document d
$Z_{d,n} \in Z$	topic assignment of a word $W_{d,n}$
D	number of documents
K	number of topics

Fig. 1. The xLDA topic model and the notation used in the paper. Unlike all previous models, the hyperparameters $\boldsymbol{\alpha}$ of the Dirichlet distribution over topics are a function of observed document features Y, relations R, and hidden topic features X, and is therefore specific to each distinct combination of document feature values and relations among the documents.

$\boldsymbol{\alpha}$ values of the pages in this chain. In contrast, a relational probabilistic model such as a Markov logic network would — without additional modeling effort — ignore all training pages in this chain besides the endpoints due to the Markov assumption, see [27] for more details. In addition, most state-of-the art probabilistic relational models focus on discrete quantities and not continuous ones such as $\boldsymbol{\alpha}$.

3 Modeling the Influence of Document Relations with Dirichlet-multinomial Regression

Nowadays, networks of D many documents, such as citation networks of scientific papers, hyperlinked networks of web pages, and social networks of friends, are becoming pervasive in machine learning applications. For each document d, we observe N_d words, each of which is an element in a V-term Vocabulary. To capture the per-document meta-information, let y_d be a vector containing F many features that encode metadata values for the document d. For example, if the observed features are indicators for the type of venue the paper was published in, then y_d would include a 1 in the positions for venue the document d has published in, and a 0 otherwise. The network among the documents (citations,

hyperlinks, friends relationships, and so on), form a graph and are captured by the adjacency matrix R. For instance, if documents d_i cites d_j, there is a 1 in position R_{ij}.

The topic model we propose, called xLDA, is a model of data composed of documents, which are collections of words, and relations among them. It is graphically depicted in Fig. 1 and consists essentially of two phases: (1) a relational Gaussian process (GP) phase, and (2) a document-specific LDA phase. In the relational GP phase, given the per-document metadata Y, the relations R among the documents and some optional meta-information for topics X, we generate the per-document α_d Dirichlet hyperparameters for each document. To do so, we use Silva *et al.*'s [27] *directed mixed graph* Gaussian process framework as it propagates training α_d through other training documents' α_d (as discussed in the related work section). Then, in the LDA phase, we run standard LDA using the generated α_d for each document d. We assume that each latent $\alpha_{d,k}$ is a function value of document metadata y_d, document relations r_d and topic metadata x_k. The (optional) topic-metainformation x_k allows one to accommodate for correlations not well explained by document metainformation only. All the function values are drawn from a GP prior with mean zero and covariance function $c((y_d, r_d, x_k), (y_{d'}, r_{d'}, x_{k'}))$. Then, the topic proportion θ_d of document d is a sample of $\mathrm{Dir}(\alpha_d)$. That is, we use a distinct Dirichlet distribution $\mathrm{Dir}(\alpha_d)$ with hyperparameters α_d as predicted by the Gaussian process. However, we have to be a little bit more careful: we have to ensure that the α_ds are positive. We do so by predicting a noisy $\tau_d \in \tau$, the logarithms $\tau_d = \eta_d + \epsilon_{noise} = \log(\alpha_d)$ of the concentration parameters α_d.

The xLDA topic model integrates heterogeneous information, namely the per-document metadata and the relations among documents, into a single probabilistic framework. The dependencies from different sources are captured in a natural and elegant way. Moreover, it can directly be used in combination with various existing relational GPs to realize multiple relations, relation prediction using fully generative relational models [8,32] — of course to the expense of a more complicated GP inference step — or even transfer learning among topic models [34].

Let us now discuss the prior distribution and how to generate words and documents in more details.

Prior Distribution: For each document, we introduce a K-dimensional vector α_d where each value $\alpha_{d,k}$ denotes the preference of a document d on a topic k. It is a function of the document's metadata y_d, its relations r_d, and the (optional) latent topic meta-information x_k. Additionally, to meet the constraint on Dirichlet parameters, i.e. $\alpha_{d,k} > 0$, we assume $\alpha_{d,k} = \exp(\tau_{d,k})$, where $\tau_{d,k} = f(y_d, r_d, x_k)$. Now, we assume that an infinite number of latent function values $\{\tau_{1,1}, \tau_{1,2}, \ldots\}$ follows a GP prior with mean function $m(y_d, r_d, x_k)$ and covariance function $c((y_d, r_d, x_k), (y_{d'}, r_{d'}, x_{k'}))$. Consequently, any finite set of function values $\{\tau_{d,k} : d = 1 \ldots D; k = 1 \ldots K\}$ has a multivariate Gaussian distribution with mean and covariance matrix defined in terms of the mean and

covariance functions of the GP, see e.g. [25]. Without loss of generality, we assume zero mean so that the GP is completely specified by the covariance function only.

In other words, to define the GP prior it is enough to specify the covariance function $c((y_d, r_d, x_k), (y_{d'}, r_{d'}, x_{k'}))$. How does it look like in our case? The input features of $\tau_{d,k}$ include document-oriented information, namely y_d and r_d, and topic-oriented information, namely x_k. Therefore, we decompose the overall covariance as a product of two types of covariances, i.e., $c_d((y_d, r_d), (y_{d'}, r_{d'})) \times c_x(x_k, x_{k'})$. Then, we notice that the document covariance component c_d involves per-document metadata and relational information. Unfortunately, it is difficult — if not impossible — to represent both jointly using a single kernel only. Consequently, we borrow the underly assumption in Silva et al.'s relational GP model [27]: $c_d((y_d, r_d), (y_{d'}, r_{d'}))$ is further decomposed into a sum of two kernels $c_y(y_d, y_{d'}) + c_r(r_d, r_{d'})$. Putting everything together, the covariance function of the GP prior is defined as $[c_y(y_d, y_{d'}) + c_r(r_d, r_{d'})] \times c_x(x_k, x_{k'})$. The decomposition of the covariance matrix is based on the direct sum and tensor product of kernels [25].

For the per-document respectively per-topic covariance functions $c_y(y_d, y_{d'})$ respectively $c_x(x_k, x_{k'})$, we can select any Mercer kernel. A typical choice is the squared exponential covariance function with isotropic distance measure:

$$c_y(y_d, y_{d'}) = \kappa^2 \exp(-\frac{\rho^2}{2} \sum\nolimits_s^S (y_{d,s} - y_{d',s})^2), \tag{1}$$

where κ and ρ are parameters of the covariance function, and $y_{d,s}$ denotes the s-th dimension of the attribute vector y_d.

For the relation-wise covariance function $c_r(r_d, r_{d'})$, any graph kernel is a natural candidate [29,35,27]. Here, we used the p-steps random walk kernel:

$$(1 - \gamma^{-1}\triangle)^p = \left[(1 - \gamma^{-1})I + \gamma^{-1} G^{-1/2} W G^{-1/2}\right]^p \tag{2}$$

where γ (with $\gamma \geq 2$) and p are the two parameter of the graph kernel and $\triangle = I - G^{-1/2} W G^{-1/2}$. The matrix W denotes the adjacency matrix of a weighted, undirected graph, i.e., $W_{i,j}$ is taken to be the weight associated with the edge between i and j. G is a diagonal matrix with entries $g_{i,i} = \sum_j w_{i,j}$. Here, mutiple relations could be encoded by weighted sum of graph kernels, kernels over weighted graphs (also to incorporate link counts) [24] or multi-relational GP's [32].

The overall covariance matrix Σ (a $DK \times DK$ matrix) computed with the covariance function $c((y_d, r_d, x_k), (y_{d'}, r_{d'}, x_{k'}))$ can be represented as $\Sigma = (\Sigma_Y + \Sigma_R) \otimes \Sigma_T$, where $\otimes$ denotes the Kronecker product between two matrices. The matrix Σ_Y is a $D \times D$ matrix that represents the per-document metadata covariances between documents. It is computed using (1). The matrix Σ_R is also a $D \times D$ matrix that represents the relation-wise covariances between documents' metadata. It is computed using (2). The sum $\Sigma_D = \Sigma_Y + \Sigma_R$ represents the document-oriented covariances. Finally, Σ_T is a $K \times K$ matrix that represents the covariances between (optional/latent) topic metadata. Every element (k, k')

of Σ_T is computed based on the latent attributes x_k and $x_{k'}$ of topics k and k' using (1). Together, this leads to the following prior distribution:

$$P(\tau|X,R) = \mathcal{N}(0,\Sigma) = \frac{1}{(2\pi)^{DK}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\tau^T \Sigma^{-1} \tau}{2}\right), \tag{3}$$

where τ denotes the logarithmic level of the $\boldsymbol{\alpha} = (\alpha_{1,1} \ldots \alpha_{d,k} \ldots \alpha_{D,K})$.

Generating Documents and Words: Given the prior $\mathcal{N}(0,\Sigma)$ and hyperparameters β, the generative process for documents and their words is as follows:

1. Draw $\tau \sim \mathcal{N}(0,\Sigma)$.
2. For each topic k, draw $\phi_k \sim \mathrm{Dir}(\beta)$.
3. For each document d,
   (a) Draw $\theta_d \sim \mathrm{Dir}(\alpha_d) = \mathrm{Dir}(\exp(\tau_d))$ with $\tau_d \in \tau$.
   (b) For each word n,
      – Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      – Draw $W_{d,n} \sim \mathrm{Mult}(\phi_{Z_{d,n}})$

The model therefore includes the following fixed parameters: the hyperparameters of the covariance matrix Σ; β, the Dirichlet prior on the topic-word distributions; and K, the number of topics.

4 Inference and Learning

With the xLDA model defined, we now turn to approximate posterior inference and parameter estimation. The main insight for both is that knowing $\boldsymbol{\alpha}$ d-separates the relational Gaussian process phase and the LDA phase.

Inference: We predict $\boldsymbol{\alpha}$ given the metadata and the relations and then run any LDA sampler.

Learning: Given $\boldsymbol{\alpha}$, (1) the GP phase of xLDA is no more complicated than a standard XGP, and (2) the sampling phase of xLDA is no more complicated than a simple LDA sampler. Thus, we can train xLDA using a stochastic EM sampling scheme. That is we alternate between sampling topic assignments from the current prior distribution conditioned on the observed words, features and relations, and numerically optimizing the parameters of the relational Gaussian process given the topic assignments. For that we need the gradients of the (log)likelihood for parts of the model that contain the GP prior respectively the topics Z.

The likelihood can be found to be $P(\tau, z|GP) = P(\tau|GP)P(z|\tau)$ with $\tau = \log(\alpha)$. Due to (3), the first term on the right-hand side is

$$P(\tau|GP) = \mathcal{N}(0,\Sigma) = \frac{1}{(2\pi)^{DK}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\tau^T \Sigma^{-1} \tau}{2}\right)$$

The second term can be written as (see [11] for details)

$$P(z|\tau) = \prod_d^D \frac{\Gamma(\sum_k^K \exp(\tau_{d,k}))}{\Gamma(\sum_k^K \exp(\tau_{d,k}) + n_{d,k})} \prod_k^K \frac{\Gamma(\exp(\tau_{d,k}) + n_{d,k})}{\Gamma(\exp(\tau_{d,k}))}$$

Consequently, the log likelihood is $LL = \log P(\tau|GP) + \log P(z|\tau) =$

$$= \underbrace{-\frac{1}{2}\tau^T \Sigma^{-1}\tau - \frac{1}{2}\log|\Sigma| - \frac{DK}{2}log(2\pi)}_{=:\triangle llgp}$$
$$+ \underbrace{\sum_d^D \left(\log \frac{\Gamma(\sum_k^K \exp(\tau_{d,k}))}{\Gamma(\sum_k^K \exp(\tau_{d,k}) + n_{d,k})} + \sum_k^K \log \frac{\Gamma(\exp(\tau_{d,k}) + n_{d,k})}{\Gamma(\exp(\tau_{d,k}))} \right)}_{=:\triangle lld}$$

The derivative of LL with respect to τ can be found to be:

$$\frac{\partial LL}{\partial \tau} = \frac{\partial \triangle llgp}{\partial \tau} + \frac{\partial \triangle lld}{\partial \tau} = -\tau^T \Sigma^{-1} + \frac{\partial \triangle lld}{\partial \exp(\tau)} * \frac{\partial \exp(\tau)}{\partial \tau}$$
$$= -\tau^T \Sigma^{-1} + \frac{\partial \triangle lld}{\partial \exp(\tau)} * \exp(\tau).$$

With respect to each element $\tau_{d,k} \in \tau$, we can find $(\partial \triangle lld)/(\partial \exp(\tau_{d,k})) =$

$$= \frac{\partial}{\partial \exp(\tau_{d,k})} \sum_d^D \left(\log \frac{\Gamma(\sum_k^K \exp(\tau_{d,k}))}{\Gamma(\sum_k^K \exp(\tau_{d,k}) + n_{d,k})} + \sum_k^K \log \frac{\Gamma(\exp(\tau_{d,k}) + n_{d,k})}{\Gamma(\exp(\tau_{d,k}))} \right)$$
$$= \Psi(\sum_k^K \exp(\tau_{d,k})) - \Psi(\sum_k^K (\exp(\tau_{d,k}) + n_{d,k})) + \Psi(\exp(\tau_{d,k}) + n_{d,k}) - \Psi(\exp(\tau_{d,k}))$$

where $\Psi(\cdot)$ is the logarithmic derivative of the Gamma function. This completes the partial derivative of LL w.r.t $\tau_{d,k} = \log(\alpha_{d,k})$. In other words, we can numerically optimize the $\alpha_{d,k}$ respectively $\tau_{d,k}$ values given topic assignments.

The partial derivatives of the GP with respect to (hyper)parameters are essentially the same as for standard Gaussian processes; they only appear in $\triangle llgp$, which is the standard data log-likelihood of GPs. Only due to the use of the Kronecker product, the derivatives look slightly different than the standard ones. Let us exemplify this for the (optional/latent) topic metadata; the other ones can be found in a similar fashion. We note that $\frac{\partial LL}{\partial x} = \frac{\partial \triangle llgp}{\partial x} + \frac{\partial \triangle lld}{\partial x}$. First,

$$\frac{\partial \triangle llgp}{\partial x} = -\frac{1}{2}\tau^T \frac{\partial \Sigma^{-1}}{\partial x}\tau - \frac{1}{2}tr(\Sigma^{-1}\frac{\partial \Sigma}{\partial x})$$
$$= \frac{1}{2}\tau^T \Sigma^{-1}\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\tau - \frac{1}{2}tr(\Sigma^{-1}\frac{\partial \Sigma}{\partial x})$$
$$= \frac{1}{2}\tau^T \Sigma^{-1}(I_D \otimes (\frac{\partial \Sigma_T}{\partial x}\Sigma_T^{-1}))\tau - \frac{D}{2}tr(\Sigma_T^{-1}\frac{\partial \Sigma_T}{\partial x})$$

with $\Sigma = \Sigma_D \otimes \Sigma_T$ where $\otimes$ denotes the Kronecker product. Then, we can find

$$\frac{\partial \triangle lld}{\partial x} = \underbrace{\frac{\partial \triangle lld}{\partial \exp(\tau)} * \frac{\partial \exp(\tau)}{\partial \tau}}_{\text{as above}} * \frac{\partial \tau}{\partial x},$$

where we have $\frac{\partial \tau}{\partial x} = \frac{\partial}{\partial x}\Sigma^*(\Sigma + \lambda^2_{noise} I)^{-1}\tau =$

$$= \Sigma^* \frac{\partial \Sigma^{-1}}{\partial x}\tau = -\Sigma^*\Sigma^{-1}\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\tau = -I(I_D \otimes (\frac{\partial \Sigma_T}{\partial x}\Sigma_T^{-1}))\tau.$$

Σ^* is the covariance of the training input. In our current setting, it coincides with Σ. For sparse extensions, however, it might be different. Now, we have all gradients together required to implement the stochastic EM approach.

5 Experimental Evaluation

Our intention here is to explore the relationship between the latent space computed by xLDA and the underlying link structure. More precisely, we investigated the following question:

(Q) Does the latent space computed by xLDA capture the underlying link structure better than LDA respectively xLDA without relational information?

To do so, we implemented LDA and xLDA in Python and C/C++. We used a standard conjugate-gradient optimizer and a collapsed Gibbs sampling-based LDA trainer. We also compared xLDA with Chang and Blei's recent relational topic model (RTM) [7]. All experiments ran on a standard Intel(R) Core(TM)2 Duo CPU with 3 GHz and 4GB main memory. All LDA and xLDA models were initialized with Dirichlet hyperparameters set to 5. The parameters of the p-steps graph kernel were set to $\gamma = 2.0$ and $p = 3.0$. (While we omit a full sensitivity study here, we observed that the performance of the models was similar for $p = 1, 2, \ldots, 8$).

Description of the Datasets: For the experiments, we used two datasets: a small dataset[1] of Wikipedia web pages used by Gruber *et al.* [12] and the Cora dataset[2] (abstracts with citations) used by Chang and Blei [7].

The Wikipedia dataset is a collection of 105 web pages with in total 89349 words and 790 links between the pages. Gruber *et al.* downloaded the web pages from Wikipedia by crawling within the Wikipedia domain, starting from the NIPS Wikipedia page. The vocabulary consists of 2247 words. The Cora dataset is a collection of 2410 abstracts from the Cora computer science research paper search engine, with in total 126394 words and 4356 links between documents that cite each other. The vocabulary consists of 2961 words. Directed links were

[1] `http://www.cs.huji.ac.il/~amitg/lthm.html`
[2] `http://cran.r-project.org/web/packages/lda/`

converted to undirected links, and documents with no links were removed. For Wikipedia, we also excluded the links of the "Machine Learning" Wikipedia page as it essentially linked to all pages. Furthermore, we turned the link structure into the co-citation link structure. That is if two documents link to another common document, we added an undirected link between these two documents.

Experimental Protocol: Due to the transductive nature of our datasets, we considered how well the models predict the remaining words of a document after observing a portion of it. Specifically, we observe p words from a document and are interested in which model provides a better predictive distribution of the remaining words $P(w|w_1, w_2, \ldots, w_p)$. To compare these distributions, we use perplexity, which can be thought of as the effective number of equally likely words according to the model:

$$Perp(\Theta) = \left(\prod_{d=1}^{D} \prod_{i=p+1}^{N_d} P(w_i|w_1, w_2, \ldots, w_p) \right)^{-1/(\sum_{d=1}^{D}(N_d - p))}$$

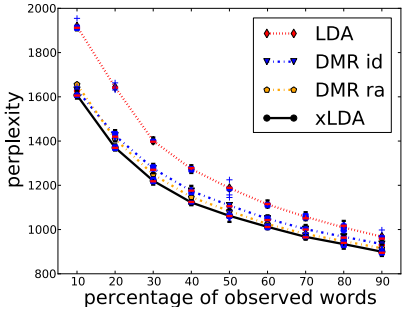
where Θ denotes the model (hyper-)parameters. Specifically, for each dataset, we created $p\%$ / $(100-p)\%$ train / test splits of the words per document for $p = 10, 20, 30, \ldots, 90$. We trained the models on each training set and evaluated the perplexity on the corresponding test set. We compared **LDA**, xLDA without relational information, which is identical to DMR with identity matrix (**DMR id**) and to LDA with hyperparameter optimization, DMR with relations as document attributes (**DMR ra**, this is essentially Mimno and McCallum's original DMR model [22] but now using a GP and treating the relations as per-document attributes), and xLDA using the co-citing information (**xLDA**). Both xLDAs estimated no topic metadata; its correlation matrix was set to the identity matrix. Each experiment was repeated 5 times, each time using a different random order of the words per document.

As noted by Gruber *et al.* [12], relational and non-relational LDA models can very well be of comparable quality in terms of perplexity on a dataset. The assignment of topics to documents, however, can be quite different. To measure this effect, we also report the Hellinger distances among related documents, i.e., documents are co-linked. Consider two documents d_i and d_j

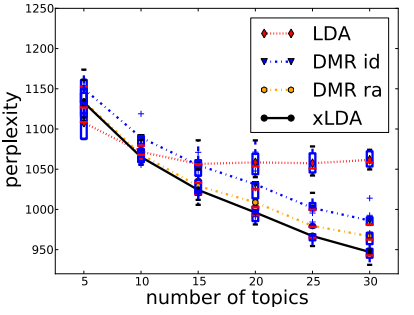
$$dist(d_i, d_j) = \sum\nolimits_k \left(\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}} \right)^2 .$$

If a model captures the link structure well, we expect the Hellinger distance smaller between co-linked documents.

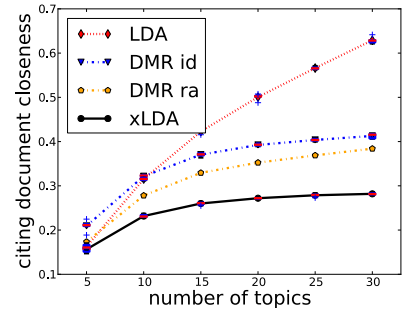
Additionally, although we are not interested in link prediction per se, we followed Chang and Blei [7] and evaluated the predictive link-likelihood of our models by first fitting the LDA models to the documents (on the full dataset) and then fitting a logistic regression model to the observed links, with input given by the Hadamard (element-wise) product of the latent class distributions of each pair of documents. That is, we first perform unsupervised dimensionality reduction, and then regression to understand the relationship between the latent space and underlying link structure. Here, we additionally compare to **RTM** [7].



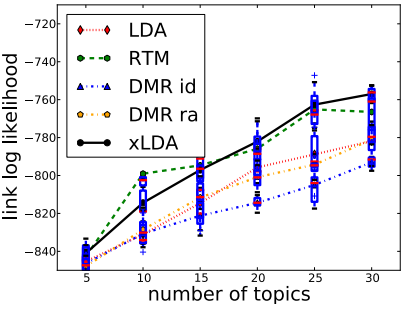
(a) Perplexity (the lower, the better) for different percentages of observed words per document.



(b) Perplexity using 70% / 30% training/test splits for different numbers of topics.



(c) Hellinger distance (the lower, the better) of linked documents for different number of topics.

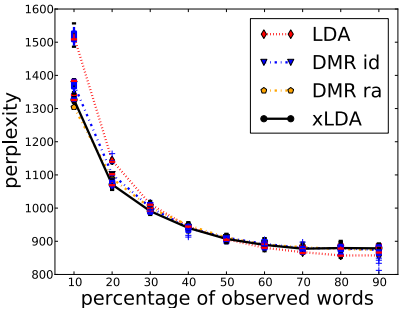


(d) Average link log likelihood (the higher, the better) for different number of topics.

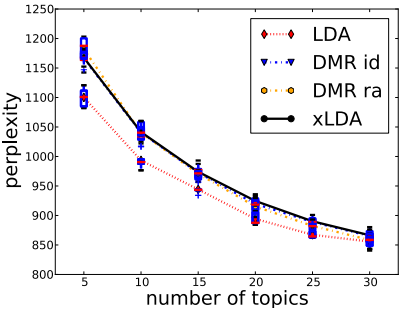
Fig. 2. Results on the Cora: Perplexity, Hellinger distance, and average link log-likelihood for **LDA**, **DMR id**, **DMR ra**, and **xLDA** using co-citation. For the average link log-likelihood, we also compare to **RTM**. (Best viewed in color.)

Finally, we investigated the benefit of latent topic meta-information. We ran xLDA estimating latent topic metadata (**xLDAmtic**) on the Wikipedia dataset with and without co-citation relations assuming 10 topics. We show the estimated covariance matrices and qualitatively compare the correlations found with the topics found.

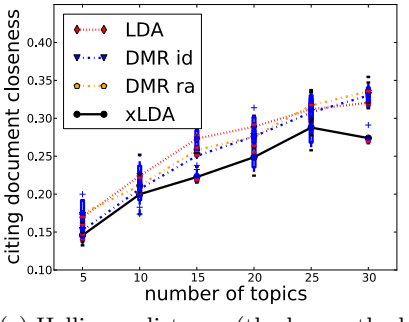
Results: The perplexity results on Cora, Fig. 2(a), clearly show that **xLDA** can significantly be less uncertain about the remaining words than **LDA** and **DMR** (K = 25). The reason is that after seeing a few words in one topic, **xLDA** uses the link structure to infer that words in a related topic may also be probable. In contrast, **LDA** cannot predict the remaining words as well until a large portion of the document has been observed so that all of its topics are represented. Only when a very small number of words have been observed, the difference starts to vanish. This performance gain was also very stable when varying the number of



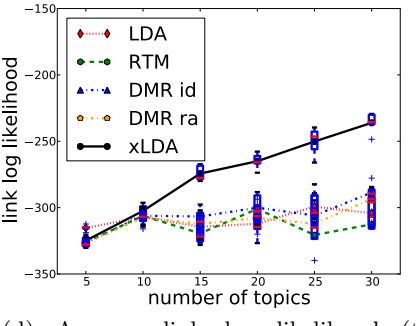
(a) Perplexity (the lower, the better) for different percentages of observed words per document.



(b) Perplexity using 70% / 30% training/test splits for different numbers of topics.



(c) Hellinger distance (the lower, the better) of linked documents for different number of topics.



(d) Average link log likelihood (the higher, the better) for different number of topics.

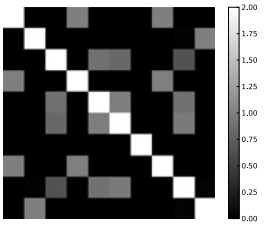
Fig. 3. Results on the Wikipedia: Perplexity, Hellinger distance, and average link log-likelihood for **LDA**, **DMR id**, **DMR ra**, and **xLDA** using co-citation. For the average link log-likelihood, we also compare to **RTM**. (Best viewed in color.)

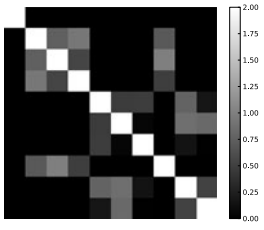
topics as shown in Fig. 2(b). For larger numbers of topics, **LDA** starts to break down compared to **DMR** and **xLDA**. This effect can be broken when optimizing the Dirichlet hyperparameters for each document separately as essentially done by **DMR id**. Again, however, **xLDA** can make use of the link structure to infer that words in a related topic may also be probable. That **xLDA** captures the link structure better is best seen when considering the Hellinger distances between co-linked documents as shown in Fig. 2(c). Most surprisingly, however, in predicting links based on the topics proportions only, **xLDA**'s performance is even comparable with **RTM**'s, a recent fully-generative model.

The perplexity results on Wikipedia, Fig. 3(a), show a similar result. When a small number of words have been observed, there is less uncertainty about the remaining words under **DMR** and **xLDA** than under **LDA** (K = 25). Given that this dataset is much smaller, we can better observe that **LDA** cannot

Table 1. The 5 nearest pages for two example Wikipedia pages ("Acadamic Conference", "Bayesian network") according to Hellinger distances (shown next to the page names) learned by **xLDA**, **DMR id**, and **LDA**. Bold pages denote that there is co-citation link between the two pages, italic ones that is a directed link.

Wikipage: Academic conference					
Proceedings	0.0839	*Proceedings*	0.0974	*Proceedings*	0.1634
NIPS	0.1109	MLMTA	0.1865	MLMTA	0.1641
Neural Information Processing Systems	0.1192	Morgan Kaufmann	0.1982	***NIPS***	0.1853
MLMTA	0.1323	Taxonomy	0.2145	***Neural Information Processing Systems***	0.1948
Morgan Kaufmann	0.1397	***NIPS***	0.2318	Inductive transfer	0.2172
Wikipage: Bayesian network					
Bayes net	0.0015	**Bayes net**	0.005	**Bayes net**	0.0022
Markov network	0.0892	***Markov network***	0.0991	***Graphical model***	0.1522
Graphical model	0.0932	Random forest	0.1449	***Loopy belief propagation***	0.1628
Bayesian statistics	0.1153	Minimum message length	0.1467	*Variational Bayes*	0.1922
Conditional probability	0.1351	***Graphical model***	0.1498	***Markov network***	0.1963
xLDA		**DMR id**		**LDA**	





No Relations		**Co-Citations**
theory problems logic called	**Topic 0**	theory problems logic called
information network time use	**Topic 1**	information network time use
brain systems vision processing	**Topic 2**	brain human systems processing
learning fixes import skins	**Topic 3**	learning fixes import skins
science amp intelligence press	**Topic 4**	science amp intelligence press
data neural search networks	**Topic 5**	data neural search networks
city retrieved colorado canada	**Topic 6**	city retrieved vancouver colorado
probability example recognition new	**Topic 7**	probability new example recognition
function algorithm model models	**Topic 8**	function algorithm model method
psychology used study field image	**Topic 9**	used models analysis psychology

Fig. 4. Correlations among latent topic metadata found by xLDA on the Wikipedia dataset

predict the remaining words as well until a large portion of the document has been observed so that all of its topics are represented. Zooming in, we found the **LDA** topics on this dataset were of comparable quality, even slightly better, cf. Fig. 3(b). The assignments of topics to documents, however, are very different. **xLDA**'s Hellinger distances between co-linked documents, as shown in Fig. 3(c), is significantly lower for larger number of topics. Table 1 additionally shows for two Wiki pages the 5 nearest Wiki pages. As one can see, **xLDA** gets more related pages closer together. Again kind of surprising, in predicting links based on the topics proportions only, **xLDA**'s performance is even comparable with **RTM**'s performance.

Finally, Figure 4 shows the latent topic metdata correlations estimated by **xLDA** with and without co-citations on the Wikipedia dataset. Without link information, Topic 6 is unrelated to any other topic. This is not surprising as it is about cities (Denver and Vancouver, the current and previous venues of the NIPS conference). When we make uses of the link structure, however, it gets correlated to the meta information of topic 4, science and intelligence. Also the

metadata of topics 1,2,3 of the NIPS conference get more correlated. Note that we used only one-dimensional topic metadata x. To model richer correlations, one should move to higher dimensions.

To summarize, our experimental results clearly affirmatively answer our question **(Q)**: the latent space computed by xLDA captures the underlying link structure better than LDA respectively xLDA without relational information.

6 Conclusions

The xLDA model is a new topic model of networks of documents. It can be used to analyze linked corpora such as citation networks, linked web pages, and social networks with user profiles. We have demonstrated qualitatively and quantitatively that the xLDA model provides an effective and useful mechanism for analyzing and using such data. It significantly improves on non-relational topic models, integrating both node-specific information and link structure to give better predictions.

The xLDA model provides a useful complement to fully generative relational topic models such as hyper-linked LDA [12] and the RTM [7], which can make predictions on relations. More importantly, it opens the door to statistical relational reasoning and learning techniques in general. It is a very attractive avenue for future work to explore this connection and to build knowledge rich topic models using probabilistic relational models such as Markov logic network or ProbLog.

Acknowledgements. The authors would like to thank Jonathan Chang and David Blei as well as Amit Gruber, Michal Rosen-Zvi and Yair Weiss for making their document networks publically available. This work was partly supported by the Fraunhofer ATTRACT fellowship STREAM and by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project.

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. Journal of Machine Learning Research 9, 1981–2014 (2008)
2. Bhattacharya, I., Getoor, L.: A latent dirichlet model for unsupervised entity resolution. In: Proceeding of SIAM Conference on Data Mining, SDM (2006)
3. Blei, D., Lafferty, J.: Topic models. In: Srivastava, A., Sahami, M. (eds.) Text Mining: Theory and Applications. Taylor & Francis, Abington (2009)
4. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120. ACM, New York (2006)
5. Blei, D.M., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
6. Buntine, W., Jakulin, A.: Applying discrete pca in data analysis. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 59–66 (2004)
7. Chang, J., Blei, D.: Relational topic models for document networks. In: Proceeding of the International Conference on Artificial Intelligence and Statistics, AISTATS (2009)

8. Chu, W., Sindhwani, V., Ghahramani, Z., Keerthi, S.: Relational learning with gaussian processes. In: Neural Information Processing Systems (2006)
9. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
10. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influence. In: Proceeding of the International Conference on Machine Learning, ICML (2007)
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS 101(suppl. 1), 5228–5235 (2004)
12. Gruber, A., Rosen-Zvi, M., Weiss, Y.: Latent topic models for hypertext. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI (2008)
13. Hofmann, T.: Probabilistic latent semantic indexing. Research and Development in Information Retrieval, 50–57 (1999)
14. Tenenbaum, J., Sutskever, I., Salakhutdinov, R.: Modelling relational data using bayesian clustered tensor factorization. Neural Information Processing Systems (2009)
15. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: Proc. 21st AAAI (2006)
16. Li, F.-F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceeding of IEEE CVPR (2005)
17. Li, W., Yeung, D., Zhang, Z.: Probabilistic relational pca. In: Neural Information Processing Systems (2009)
18. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceeding of the International Joint Conference on Artificial Intelligence, IJCAI (2005)
19. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of International Joint Conference on Artificial Intelligence (2005)
20. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: Proceeding of the 17th International Conference on World Wide Web (2008)
21. Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD (2007)
22. Mimno, D., McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI (2008)
23. Nallapati, R., Cohen, W.: Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In: Proceedings of the International Conference on Weblogs and Social Media, ICWSM (2008)
24. Neumann, M., Kersting, K., Xu, Z., Schulz, D.: Stacked gaussian process learning. In: Kargupta, W.W.H. (ed.) Proceedings of the 9th IEEE International Conference on Data Mining (ICDM-09), Miami, FL, USA (December 6-9, 2009)
25. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
26. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceeding of UAI (2004)
27. Silva, R., Chu, W., Ghahramani, Z.: Hidden common cause relations in relational learning. In: Neural Information Processing Systems (2007)
28. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: Proc. 14th Intl. Conf. on Knowledge Discovery and Data Mining (2008)

29. Smola, A.J., Kondor, I.R.: Kernels and regularization on graphs. In: Annual Conference on Computational Learning Theory (2003)
30. Steyvers, M., Griffiths, T.L., Dennis, S.: Probabilistic inference in human semantic memory. Trends in Cognitive Science 10, 327–334 (2006)
31. Wang, C., Blei, D.M., Heckerman, D.: Continuous time dynamic topic models. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (2008)
32. Xu, Z., Kersting, K., Tresp, V.: Multi-relational learning with gaussian processes. In: Boutilier, C. (ed.) Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-09 (2009)
33. Xu, Z., Tresp, V., Yu, K., Kriegel, H.-P.: Infinite hidden relational models. In: Proc. 22nd UAI (2006)
34. Yu, K., Chu, W.: Gaussian process models for link analysis and transfer learning. In: Neural Information Processing Systems (2007)
35. Zhu, X., Kandola, J., Lafferty, J., Ghahramani, Z.: Graph kernels by spectral transforms. In: Chapelle, O., Schoelkopf, B., Zien, A. (eds.) Semi-Supervised Learning. MIT Press, Cambridge (2005)