Latent Topic Networks:
A Versatile Probabilistic Programming Framework for Topic Models

James Foulds JFOULDS@UCSC.EDU
Shachi H. Kumar SHACHIHKUMAR@SOE.UCSC.EDU
Lise Getoor GETOOR@SOE.UCSC.EDU
Department of Computer Science, University of California, Santa Cruz, CA 95064 USA

Abstract

Topic models have become increasingly prominent text-analytic machine learning tools for research in the social sciences and the humanities. In particular, custom topic models can be developed to answer specific research questions. The design of these models requires a nontrivial amount of effort and expertise, motivating general-purpose topic modeling frameworks. In this paper we introduce *latent topic networks*, a flexible class of richly structured topic models designed to facilitate applied research. Custom models can straightforwardly be developed in our framework with an intuitive first-order logical probabilistic programming language. Latent topic networks admit scalable training via a parallelizable EM algorithm which leverages ADMM in the M-step. We demonstrate the broad applicability of the models with case studies on modeling influence in citation networks, and U.S. Presidential State of the Union addresses.

1. Introduction

In the last decade topic models have become a core tool for the analysis of text corpora. Their usage has spread beyond machine learning to answer substantive questions in disciplines such as cognitive science (Griffiths et al., 2007), political science (Grimmer & Stewart, 2013; Lucas et al., 2015), and sociology (McFarland et al., 2013). The success of these models is due in part to their extensibility. The latent Dirichlet allocation (LDA) topic model of Blei et al. (2003) is frequently used as a foundational building block for constructing more sophisticated latent variable models (Grimmer, 2010; Gerrish & Blei, 2011; Nguyen et al., 2014). Topic models can also be extended to model additional structure such as correlations (Blei & Lafferty, 2007) or covariates (Mimno & McCallum, 2008), which are key for social science applications (Roberts et al., 2014).

However, the development of new topics models is a time-consuming and challenging process. As well as designing the model, a corresponding inference algorithm must be derived and implemented in order to fit the model to data, with the model/algorithm pair being carefully selected to make inference tractable and efficient. The model and algorithm then need to be evaluated, either relative to some extrinsic task, or with respect to an intrinsic measure of quality such as semantic coherence (Newman et al., 2010; Mimno et al., 2011) or predictive performance (Wallach et al., 2009; Foulds & Smyth, 2014), or by human judgment (Chang et al., 2009). Depending on the outcome of the evaluation, the entire process may need to be repeated iteratively until the desired level of performance is reached.

The effort required for the development process motivates general purpose modeling frameworks for building and inferring custom topic models. From a practical perspective, the challenge is to design a topic modeling framework which *(1)* is general enough to be widely applicable to many latent variable modeling applications, while *(2)* remaining scalable. We would also like the framework to be *(3)* easy to use by non-specialist domain scientists, which suggests a probabilistic programming approach.

Regarding *(1)*, Roberts et al. (2013; 2014) argue that a general-purpose topic modeling framework for applied social science needs to model covariates as well as dependencies/correlations between documents and topics, for which they propose the structural topic model (STM). The STM combines several previous models into a unified framework, and its application-focused design represents a substantial and useful step towards a general social science topic modeling toolkit. This model does not however provide the capability for including additional latent variables to build more sophisticated latent variable models. Another

Proceedings of the 32^{nd} *International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

Table 1. A comparison of general-purpose topic modeling frameworks.

	Correlations / Dependencies	Observed Covariates	Additional Latent Variables	Constraints	Probabilistic Programming
Systems for Encoding Domain Knowledge, Covariates and Correlations					
CTM (Blei & Lafferty, 2007)	✓	×	×	×	×
DMR (Mimno & McCallum, 2008)	×	✓	×	×	×
Dir. Forests (Andrzejewski et al., 2009)	×	×	×	✓	×
xLDA (Wahabzada et al., 2010)	✓	✓	✓	×	×
SAGE (Eisenstein et al., 2011)	×	✓	×	×	×
STM (Roberts et al., 2013; 2014)	✓	✓	×	×	×
Graphical Modeling and Probabilistic Programming Systems					
CTRF (Zhu & Xing, 2010)	Intractable[a]	✓	×	×	×
Fold.all (Andrzejewski et al., 2011)	✓	✓	×[b]	×	✓
Logic LDA (Mei et al., 2014)	×	✓	×	✓	✓
LTN (this paper)	✓	✓	✓	✓	✓

[a] Conditional topic random fields (CTRFs) are tractable when the graph structure is a chain, but are intractable for general graphs.
[b] Andrzejewski et al. (2011) mention the possibility of latent query variables as a possible future direction.

desirable property missing from the STM is the ability to encode domain knowledge using constraints, as in Andrzejewski et al. (2009). Other general-purpose topic modeling frameworks have been proposed, some of which use probabilistic programming systems (Andrzejewski et al., 2011) and/or represent problem-specific domain knowledge via posterior constraints (Mei et al., 2014), but none of these frameworks satisfy all of our desiderata (Table 1).

To address these limitations, this article introduces a flexible probabilistic programming framework for designing custom topic models. Using the framework, an analyst can specify models using a declarative first-order logical probabilistic programming language called *probabilistic soft logic* (Bach et al., 2015). The resulting models, which we refer to as *latent topic networks*, directly generalize LDA, but add prior structure, dependency relationships, and additional latent and observed variables, using a tractable class of graphical models called hinge-loss Markov random fields (HL-MRFs) (Bach et al., 2013). We show how to fit latent topic networks using an EM algorithm which is highly parallelizable without approximation, leveraging an alternating direction method of multipliers (ADMM) (Boyd et al., 2011) algorithm in the M-step. We demonstrate the system with several case studies, including a model for exploring influence between scientific articles, and for modeling State of the Union addresses.

2. Latent Topic Networks

The proposed models extend latent Dirichlet allocation (LDA) topic models (Blei et al., 2003). LDA encodes the semantic themes of a text corpus with K "topics" $\phi^{(k)}$, each of which is a discrete (categorical) distribution over the M words in the dictionary. It associates each document d with a discrete distribution $\theta^{(d)}$ over the K topics. The model then posits the following generative process:

- For each document d, $1, \ldots, D$
  - For each word token i, $1, \ldots, N_d$
    - Draw a latent topic assignment, $z_i^{(d)} \sim \text{Discrete}(\theta^{(d)})$
    - Draw the word token, $\omega_i^{(d)} \sim \text{Discrete}(\phi^{(z_i^{(d)})})$.

In LDA, the parameters $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are given Dirichlet priors

$$\theta^{(d)} \sim \text{Dirichlet}(\alpha) \qquad \phi^{(k)} \sim \text{Dirichlet}(\beta)\,. \tag{1}$$

The independence assumptions implicit in the Dirichlet priors are what prevents LDA from capturing complex dependencies (Blei & Lafferty, 2007). The priors are therefore our point of attack in developing a rich, flexible class of topic models. In our proposed models, we replace the Dirichlet priors of Equation 1 with a tractable class of conditional random field (CRF) models over continuous random variables, known as *hinge-loss Markov random fields* (HL-MRFs) (Bach et al., 2013). While we are not the first to employ undirected graphical models to encode structure in topic models, cf. Zhu & Xing (2010) and Andrzejewski et al. (2011), HL-MRFs admit tractable MAP inference regardless of the graph structure of the graphical model, making it feasible to reason over complex user-specified dependencies. This is possible because HL-MRFs operate on *continuous* random variables and encode dependencies using potential functions that are *convex*, so MAP inference in these models is always a convex optimization problem. Specifically, hinge-loss MRFs define probability densities

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\Big(-\sum_{j=1}^{M} \lambda_j \psi_j(\mathbf{X},\mathbf{Y})\Big)$$
$$\psi_j(\mathbf{X},\mathbf{Y}) = [\max\{l_j(\mathbf{X},\mathbf{Y}), 0\}]^{\rho_j}\,, \tag{2}$$

where the entries of $\mathbf{Y}$ and $\mathbf{X}$ are continuous random variables in the range $[0,1]$, and the ψ_j are *hinge-loss potentials*, specified by a linear function $l_j^{(a)}$ and an exponent

$p_j^{(a)} \in \{1, 2\}$ which optionally squares the potential. HL-MRFs can further optionally include linear equality and inequality constraints on the support of the distribution.

We obtain our customizable topic models, which we refer to as *latent topic networks* (LTNs), by positing that the topic model parameters, as well as optional sets of user-specified observed "target" variables $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ and hidden variables $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$, are drawn via HL-MRFs,

$$P(\mathbf{\Phi}, \mathbf{Y}^{(1)}, \mathbf{H}^{(1)} | \mathbf{X}^{(1)}) \propto \exp\Big(- \sum_{j=1}^{M^{(1)}} \lambda_j^{(1)} \psi_j^{(1)}(\mathbf{\Phi}, \mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{H}^{(1)})\Big) \quad (3)$$

$$P(\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)} | \mathbf{X}^{(2)}) \propto \exp\Big(- \sum_{j=1}^{M^{(2)}} \lambda_j^{(2)} \psi_j^{(2)}(\mathbf{\Theta}, \mathbf{X}^{(2)}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)})\Big), \quad (4)$$

where $\psi_j^{(a)}$ are hinge-loss potentials and $\mathbf{X}^{(a)}$ are observed covariates that are conditioned on. Each topic $\phi^{(k)}$ and each distribution over topics $\theta^{(d)}$ is constrained to sum to one, as in LDA. If there are any topics $\phi^{(k)}$ or distributions over topics $\theta^{(d)}$ for which we do not want to specify prior structure, we give them Dirichlet priors as in Equation 1. For consistency between structured and unstructured variables, and for smoothing purposes, we also include "Dirichlet-like" unary potentials $\prod_k {\theta_k^{(d)}}^{\alpha-1}$ and $\prod_w {\phi_w^{(k)}}^{\beta-1}$ for the variables covered by the HL-MRF priors. The log posterior of the variables of interest is then

$$\begin{aligned}
&\log Pr(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)} | w, \beta, \alpha, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \lambda) \\
&= \sum_{d=1}^{D} \sum_{i=1}^{N_d} \log \Big(\sum_{k=1}^{K} Pr(w_i^{(d)}, z_i^{(d)} = k | \theta^{(d)}, \mathbf{\Phi}) \Big) \\
&\quad + \sum_{d=1}^{D} \sum_{k=1}^{K} (\alpha - 1) \log(\theta_k^{(d)}) + \sum_{w=1}^{W} \sum_{k=1}^{K} (\beta - 1) \log(\Phi_w^{(k)}) \\
&\quad - \sum_{j=1}^{M^{(1)}} \lambda_j^{(1)} \psi_j^{(1)}(\mathbf{\Phi}, \mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{H}^{(1)}) \\
&\quad - \sum_{j=1}^{M^{(2)}} \lambda_j^{(2)} \psi_j^{(2)}(\mathbf{\Theta}, \mathbf{X}^{(2)}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)}) + \text{ const.} \qquad (5)
\end{aligned}$$

3. Probabilistic Programming for Latent Topic Networks

Another advantage of hinge-loss MRFs is that they can be specified using a declarative probabilistic programming language called *probabilistic soft logic* (PSL) (Bach et al., 2015). A PSL program consists of a collection of weighted first-order logical rules, analogous to Markov logic (Richardson & Domingos, 2006) except that they operate on real-valued predicates with values between zero and one, such as the entries of $\mathbf{\Phi}$ and $\mathbf{\Theta}$, and any other latent and observed variables that are included in the model. The rules consist of convex relaxations of Boolean logical operators which are exact when applied to 0 and 1 values:

$$\begin{aligned}
A \,\&\, B &= \max(A + B - 1, 0) \\
A \vee B &= \min(A + B, 1) \\
\neg A &= 1 - A\,.
\end{aligned}$$

Rules are created by applying these operators recursively. Each grounding (instantiation) of a valid PSL rule corresponds to a hinge-loss potential function in the resulting conditional random field. Each hinge-loss potential reduces the probability of states according to its rules' *distance from satisfaction*, defined to be the negation of the value of the relaxed rule. For example, $A \Rightarrow B = \neg A \vee B$ has a distance to satisfaction which is a hinge function, $\max(A - B, 0)$. Thus, $A \Rightarrow B$ penalizes the probability of a state based on the extent to which $A > B$. This class of feature functions can be derived from several motivating formulations, including linear programming relaxations of MAX SAT, local consistency relaxations for discrete MRFs, and Lukasiewicz logic (Bach et al., 2015). Following Beltagy et al. (2014), we extend PSL with the averaging operator $\wedge$, an alternative linear approximation to logical conjunction which is useful when variables have small values, such as the entries of $\mathbf{\Phi}$ and $\mathbf{\Theta}$:

$$A_1 \wedge A_2 \wedge \ldots \wedge A_N = \frac{\sum_{i=1}^{N} A_i}{N}\,. \qquad (6)$$

Both $A \,\&\, B$ and $A \wedge B$ treat A and B additively, but map the result to $[0, 1]$ differently, by translating by -1 or dividing by 2, respectively. If $A = 1$, $A \,\&\, B = B$, and if $A = 0$, $A \,\&\, B = 0$, so & is useful as a "selection" operator when we want rules to only apply in certain cases, and when we want to reduce the number of groundings to be considered. On the other hand, $A \,\&\, B = 0$ if $(A + B)/2 \leq 0.5$, which will be the case when A and B are entries of $\mathbf{\Phi}$ or $\mathbf{\Theta}$, in which case $\wedge$ conjunctions will be more useful. In the context of topic modeling, some example rules include:

- **Correlated topics:** $(\text{correlated}(k, k') \,\&\, \theta_k^{(d)}) \Rightarrow \theta_{k'}^{(d)}$
- **Influence:** $(\text{influences}(d, d') \,\&\, \theta_k^{(d)}) \Rightarrow \theta_k^{(d')}$
- **Covariates:** $\text{covariate}(c, d) \Rightarrow \theta_k^{(d)}$
- **Time series modeling:**
  - $\theta_k^{(d,t)} \Rightarrow \theta_k^{(d,t+1)}, \qquad \theta_k^{(d,t+1)} \Rightarrow \theta_k^{(d,t)}$
  - $\phi_w^{(k,t)} \Rightarrow \phi_w^{(k,t+1)}, \qquad \phi_w^{(k,t+1)} \Rightarrow \phi_w^{(k,t)}$
- **Must-link relationships:**
  - $(\text{must-link}(w, w') \,\&\, \phi_w^{(k)}) \Rightarrow \phi_{w'}^{(k)}\,.$

4. Training via EM

We train the model by maximum a posteriori (MAP) estimation, optimizing Equation 5 with respect to $\mathbf{\Theta}$, $\mathbf{\Phi}$, $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$.[1] This equation cannot be optimized directly due to the sum inside the logarithm, which ironically arises from the LDA portion of the model. Instead, we use an EM algorithm to optimize Equation 5 by iteratively optimizing a lower bound arising from Jensen's inequality,

$$R(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)}; \Theta^{(t)}, \Phi^{(t)}, \mathbf{H}^{(1,t)}, \mathbf{H}^{(2,t)}) \leq$$
$$\log P(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)} | w, \beta, \alpha, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \lambda)$$

where $R(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)}; \Theta^{(t)}, \Phi^{(t)}, \mathbf{H}^{(1,t)}, \mathbf{H}^{(2,t)}) =$

$$\begin{aligned}
&- \sum_{j=1}^{M^{(1)}} \lambda_j^{(1)} \psi_j^{(1)} (\mathbf{\Phi}, \mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{H}^{(1)}) \\
&- \sum_{j=1}^{M^{(2)}} \lambda_j^{(2)} \psi_j^{(2)} (\mathbf{\Theta}, \mathbf{X}^{(2)}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)}) \\
&+ \sum_{wk} (\sum_{id: w_i^{(d)} = w} \gamma_{idk} + \beta - 1) \log \phi_w^{(k)} \\
&+ \sum_{dk} (\sum_i \gamma_{idk} + \alpha - 1) \log \theta_k^{(d)} \\
&- \sum_{idk} \bar{\gamma}_{idk} \log \bar{\gamma}_{idk} + \text{ const} \qquad (7)
\end{aligned}$$

is the expected complete data log-likelihood with respect to the distribution $Pr(\mathbf{Z} | \Theta^{(t)}, \Phi^{(t)})$ over latent topic assignments given the parameters at the previous iteration, plus terms arising from the prior and entropy terms, and where $\gamma_{idk} \triangleq Pr(z_i^{(d)} = k | \mathbf{\Theta}^{(t)}, \mathbf{\Phi}^{(t)}, w_i^{(d)})$ are E-step "responsibilities," which encode the distribution over the latent variables based on the previous parameter values. The algorithm consists of an E-step and an M-step, which are iterated until convergence. Both the E and M-steps can be parallelized without approximation.

4.1. E-step

To perform the E-step of the EM algorithm at iteration t, we find $R(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{H}; \Theta^{(t)}, \Phi^{(t)}, \mathbf{H}^{(t)})$ by computing the E-step responsibilities γ_{idk} in Equation 7,

$$\begin{aligned}
\gamma_{idk} &\propto P(w_i^{(d)} | z_i^{(d)} = k, \mathbf{\Theta}^{(t)}, \mathbf{\Phi}^{(t)}) P(z_i^{(d)} = k | \mathbf{\Theta}^{(t)}, \mathbf{\Phi}^{(t)}) \\
&= \phi_{w_i^{(d)}}^{(k,t)} \theta_k^{(d,t)} \, . \qquad (8)
\end{aligned}$$

4.2. M-step

The M-step update optimizes $R(\mathbf{\Theta}, \mathbf{\Phi}, \mathbf{H}; \Theta^{(t)}, \Phi^{(t)}, \mathbf{H}^{(t)})$ with respect to $\mathbf{\Phi}$, $\mathbf{\Theta}$, and $\mathbf{H}$. The negative of this function is convex in these variables so it is frequently feasible to solve this exactly, although a generalized EM algorithm that simply improves this function is sufficient for convergence. We first perform the M-step for any $\mathbf{\Theta}$ and $\mathbf{\Phi}$ parameters not involved in the hinge-loss MRF, for which the update is identical to the M-step of the EM algorithm for LDA. For these parameters, by adding Lagrange terms $-\sum_k \eta_k^{\Phi} (\sum_w \phi_w^{(k)} - 1)$ and $-\sum_d \eta_d^{\Theta} (\sum_k \theta_k^{(d)} - 1)$ to constrain the parameter vectors to sum to one, taking derivatives and setting to zero, we obtain the updates

$$\begin{aligned}
\phi_w^{(k)} &:\propto \sum_{id: w_i^{(d)} = w} \bar{\gamma}_{idk} + \beta - 1 \\
\theta_k^{(d)} &:\propto \sum_i \bar{\gamma}_{idk} + \alpha - 1 \, . \qquad (9)
\end{aligned}$$

Fixing these updated non-hinge-loss parameters, we optimize Equation 7 jointly over the remaining parameters. The problem decomposes over $\mathbf{\Theta}$ and $\mathbf{\Phi}$, which may be optimized separately. We minimize the negative of each of these two subproblems $-R(\mathbf{\Phi}, \mathbf{H}^{(1)}; \mathbf{\Phi}^{(t)})$ and $-R(\mathbf{\Theta}, \mathbf{H}^{(2)}; \mathbf{\Theta}^{(t)})$ using a consensus-optimization algorithm based on the alternating direction method of multipliers (ADMM) (Boyd et al., 2011), building upon Bach et al. (2013)'s ADMM algorithm for MAP inference in HL-MRFs. For each potential function and constraint, Bach et al. (2013)'s algorithm creates a local copy of its variables. With the constraint that the local copies are equal to the original variables, this problem is equivalent to the original one. The equality constraints are then relaxed using the method of Lagrange multipliers, splitting the problem into independent subproblems that may be solved in parallel. The algorithm proceeds by repeatedly solving the independent subproblems and updating original variables, known as *consensus variables*, to be the average of the local copies. It is guaranteed to find the global optimal of the objective function. For more information on the algorithm including pseudocode see Bach et al. (2013), and see Boyd et al. (2011) for more information on ADMM.

We follow the ADMM algorithm of Bach et al. (2013), but extend it to include the objective function terms in Equation 7 of the form $n_i \log \Psi_i$, where n_i is a sum of γ variables and prior terms, and Ψ_i is an entry of either $\mathbf{\Phi}$ or $\mathbf{\Theta}$. Note that the entropy terms and normalization constant of Equation 7 are not relevant to the M-step optimization over $\mathbf{\Theta}$, $\mathbf{\Phi}$, and $\mathbf{H}$. Our consensus ADMM algorithm creates local copies x_i of each parameter Ψ_i, and adds Lagrange terms η_i corresponding to the relaxed constraint that these copies are equal to the consensus variables. In each iteration of the resulting ADMM algorithm, in addition to the

[1] Hidden variables such as $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ are often treated as "nuisance" variables and marginalized out when performing EM. In this case, we interpret them as parameters to maximize over, in part for computational reasons, as we can straightforwardly maximize over them but cannot easily marginalize over them. Furthermore, in many social science applications we would like to report the values of the latent variables, and these point estimates are easier to interpret than variational distributions.

steps of Bach et al. the x_i and η_i are updated. The consensus ADMM update (Boyd et al., 2011) sets each x_i to

$$\arg\min_{x'_i} \left(-n_i \log x'_i + \eta_i (x'_i - \Psi_i) + \frac{\rho}{2} (x'_i - \Psi_i)^2 \right) , \quad (10)$$

where ρ is an ADMM step-size parameter. To minimize Equation 10, we take the derivative and set it to zero,

$$0 = \frac{d\text{Local Eqn}}{dx'_i} = \rho x'^2_i + x'_i(\eta_i - \rho \Psi_i) - n_i \,. \quad (11)$$

Equation 11 is a quadratic which can be solved in closed form using the quadratic formula. It has two solutions, however one of them will be negative and can be discarded. We update x_i to the positive solution. Finally, the Lagrange parameters η_i are updated using the consensus ADMM update for dual parameters (Boyd et al., 2011),

$$\eta_i \leftarrow \eta_i + \rho (x_i - \Psi_i) \,. \quad (12)$$

In our experiments, we found that it is highly beneficial to warm-start the ADMM variables at their values from the previous EM iteration, which results in decreasing time to convergence per M-step as the EM algorithm proceeds.

4.3. Weight Learning

In some cases we may be able to select the first-order rule weights λ based on domain knowledge, as in Andrzejewski et al. (2011). If this information is not available the weights must be learned from data. Weight learning in Markov random fields is in general a challenging problem, as even a gradient ascent update for the log-likelihood requires the computation of an intractable expectation, arising from the normalization constant that was dropped in Equation 7. For weight learning, we therefore extend the EM algorithm to update λ by optimizing a pseudo-likelihood approximation to the posterior in the M-step. The relevant portion of the posterior for Θ is $P(\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)} | \mathbf{X}^{(2)}, \alpha)$, as given in Equation 4 but with the "Dirichlet-like" potentials included, which result in terms $\sum_{dk} (\alpha - 1) \log(\theta^{(d)}_k)$ being added inside the exp. We can define the pseudo-likelihood,

$$P^*(\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)} | \mathbf{X}^{(2)}, \alpha) = \prod_{V \in \{\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)}\}} P(V | B(V))$$

where $B(V)$ is the Markov blanket of V. We perform gradient descent on the pseudo log-likelihood, via

$$\frac{d}{d\lambda^{(2)}_q} \log P^*(\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)} | \mathbf{X}^{(2)}, \alpha) \quad (13)$$
$$= \sum_{V \in \{\mathbf{\Theta}, \mathbf{Y}^{(2)}, \mathbf{H}^{(2)}\}} \left(E_{P(V|B(V))} [\psi^{(2)}_q(\cdot)] - \psi^{(2)}_q(\cdot) \right) .$$

We group the variables in each $\theta^{(d)}$ together as a single variable V in the pseudo-likelihood, due to the simplex constraint. The expectations in Equation 13 are estimated via importance sampling, with a uniform proposal for the $\mathbf{Y}$ and $\mathbf{H}$ variables, and a Dirichlet proposal for the $\theta^{(d)}$'s, using the Dirichlet with concentration parameter α implied by the Dirichlet-like potentials. An analogous procedure is applied to learn the weights $\lambda^{(1)}$ of the MRF for Φ.

5. Experiments

We explore the generality and effectiveness of the latent topic network framework by employing it for two applications: influence modeling in citation networks of scientific articles, and time-series modeling of United States Presidential State of the Union addresses.[2] These applications were chosen to illustrate the potential benefits of the framework for modeling in the social sciences.

5.1. Exploring Scientific Influence in Citation Networks

In our careers as scientists we often will be introduced to fields of study that we are not yet familiar with, and it would be beneficial to have automatic tools that can help us to quickly orient ourselves in the literature. For example, we may wish to find the articles that were influential on the work that followed them. Inspired by the topical influence regression (TIR) model of Foulds & Smyth (2013), we construct a latent topic network encoding the hypothesis that influential articles "coerce" the articles that cite them into having similar distributions over topics. This application demonstrates the ability of latent topic networks to encode networks of dependencies between documents, and to reason over latent variables jointly with the topic model.

The PSL program in Table 2 defines an LTN that infers latent real-valued node-wise and edge-wise citation influence variables influential(A) and influences(A, B) in addition to the topic model parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$, given an observed binary predicate cites(A, B) which encodes the citation graph. The model posits that influential articles are more likely to influence the articles that cite them, and vice-versa. Articles are encouraged to have similar topics to the articles that influence them, depending on the degree of influence exerted. Finally, articles whose topics overlap heavily with the topics of the articles that cite them are more likely to have influenced those articles to a greater degree. In the PSL rules, conjunctions with the *cites* predicate restrict all influence relationships to the citation graph. We conditioned on the citation network (the *cites* predicate), and performed inference jointly over $\mathbf{\Theta}$, $\mathbf{\Phi}$, and the *influences* and *influential* predicates.

[2] Our code will be available as part of the PSL software at http://psl.cs.umd.edu/.

Table 2. PSL rules for a latent topic network designed to model citation influence.

Document-level and edge-level influence			*Influence relationships on citation edges*		
cites(A, B) & influential(B)	$\Rightarrow$	influences(B, A)	cites(A, B) & (influences$(B, A) \wedge \theta_k^{(B)}$)	$\Rightarrow$	$\theta_k^{(A)}$
cites(A, B) & influences(B, A)	$\Rightarrow$	influential(B)	cites(A, B) & $(\theta_k^{(A)} \wedge \theta_k^{(B)})$	$\Rightarrow$	influences(B, A)
$\neg$influential(A)					

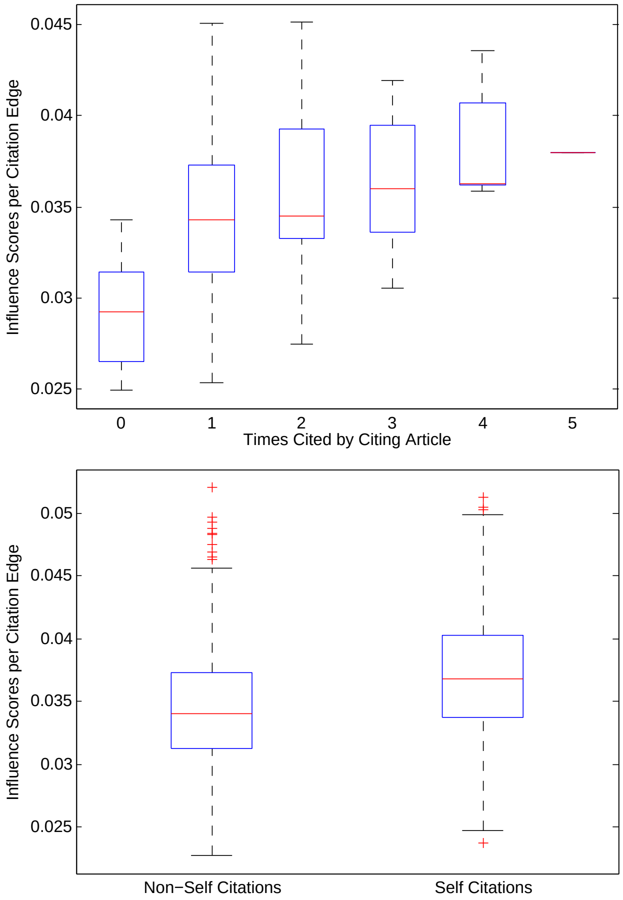

Figure 1. Evaluating the citation influence LTN model on the NIPS corpus. **Top:** Inferred influence scores per edge versus number of times cited by the citing article. **Bottom:** Inferred influence scores for self and non-self citation edges.

We trained a latent topic network with 50 topics on a corpus of 1740 articles from the NIPS conference.[3] EM was run for 250 iterations, with weight learning performed every 20 iterations starting from iteration 50. Training took roughly five hours on a quad-core 2.4Ghz laptop, using 8 threads via hyper-threading, with the majority of the time spent in weight learning. Ground truth citation influence information was not available, so we instead validated the model using metadata as a proxy for ground truth, following the experimental setup of Foulds & Smyth (2013). The metadata validation results are overall similar to those reported by Foulds and Smyth for their purpose-built TIRE model, which was designed specifically to model influence in citation graphs. However, our model was constructed with just 5 lines of PSL code in our *general-purpose* framework.

Figure 1 (top) shows boxplots of inferred citation influence scores categorized by the number of times that the cited article is mentioned in the text of the citing article, using the pairs of in-text citation counts from 106 NIPS articles that were extracted by Foulds and Smyth. Our results follow the trend of the special-purpose TIRE model, with the influence scores increasing on average with the number of repeated in-text citations. For each pair of references per document, the most influential references according to the model were cited 168 times in the text overall, while the least influential references were cited 131 times. This is comparable to the purpose-built TIRE model, for which the most influential references were cited 171 times overall, and the least influential were cited 128 times. There were 45 articles for which the citation counts were not tied. Of these articles, the most influential references had the higher citation count 31 times, comparable to 33 times for TIRE. A sign test with $\alpha = 0.05$ rejects the null hypothesis that the median difference in citation counts between the most and least-influential is zero, with p-value $= 0.016$.

Self-citations are also likely to be informative for citation influence, as we would expect same-author citations to have higher influence on the citing article on average. Figure 1 (bottom) compares the inferred influence scores of the LTN model for citations with an author in common between the citing and cited articles, and those with no authors in common. A two-sample t-test with $\alpha = 0.05$ rejects the null hypothesis that the means of self-citation and non-self citation edges' influence scores are equal.

5.2. Modeling State of the Union Addresses

The Presidents of the United States of America have presented a State of the Union message to Congress annually, with a few exceptions, since 1790. We constructed an LTN to explore the extent to which these addresses depict the true underlying state of the Union, or are biased by political ideology. The model posits time-evolving latent variables for the "state" of the Union and the bias of each political party. This application explores the ability of our framework to perform dynamic time-series modeling, and to reason over distributional latent variables.

[3] The corpus is due to Gregor Heinrich. It is available at http://www.arbylon.net/resources.html.

Table 3. PSL rules for the State of the Union time series model. The SOTU and party bias variables are constrained to sum to one.

The latent state of the Union (SOTU) distribution varies smoothly per year and influences $\mathbf{\Theta}$.		
SOTU$(Y1, k)$ & precedes$(Y1, Y2)$	$\Rightarrow$	SOTU$(Y2, k)$
SOTU$(Y2, k)$ & precedes$(Y1, Y2)$	$\Rightarrow$	SOTU$(Y1, k)$
SOTU(Y, k)	$\Rightarrow$	$\theta_k^{(Y)}$
The latent party bias distributions vary smoothly per decade and influence $\mathbf{\Theta}$ *when the party has a President.*		
RepublicanTheta$(DEC1, k)$ & precedesDecade$(DEC1, DEC2)$	$\Rightarrow$	RepublicanTheta$(DEC2, k)$
RepublicanTheta$(DEC2, k)$ & precedesDecade$(DEC1, DEC2)$	$\Rightarrow$	RepublicanTheta$(DEC1, k)$
RepublicanTheta(DEC, k) & inDecade(Y, DEC) & RepublicanPresident(Y)	$\Rightarrow$	$\theta_k^{(Y)}$
(Similar rules for the other parties. . .)		

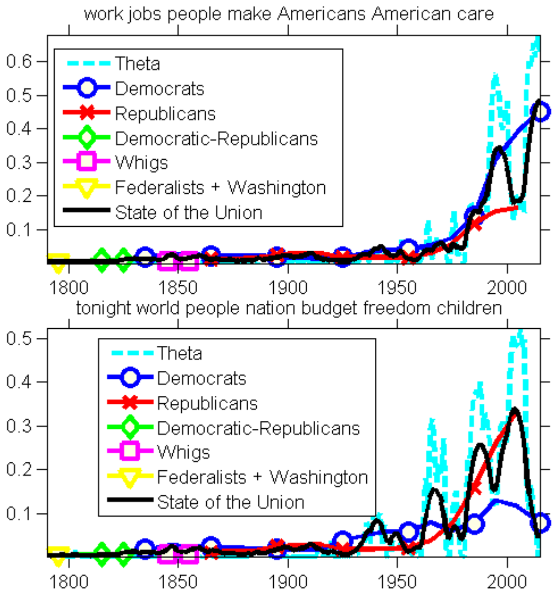

Figure 2. Modeling the latent state of the Union from Presidential State of the Union addresses. The plots show topic proportions and latent variable proportions for topics that have become associated with the two major U.S. political parties.

Our LTN model in Table 3 represents the state of the Union at year Y by a latent distribution over the topics SOTU$(Y, :)$, where SOTU(Y, k) is the proportion that topic k is relevant to Congress in year Y. Each SOTU distribution is encouraged to be similar to its adjacent distributions, representing the assumption that the true state of the Union changes slowly over time. The distribution over topics $\theta^{(Y)}$ for the address at year Y is modeled as a noisy estimate of the true underlying state of the Union, and so the PSL rules encourage it to be similar to SOTU$(Y, :)$. It is also encouraged to be similar to a latent bias vector for the party of the President, representing the ideological bias of that party. These bias vectors are given similar time-series dynamics to SOTU, however they may only change once per decade.

We trained the model with 20 topics on the 225 addresses from 1790 to 2015, performing 500 EM iterations with weight learning every 20 iterations. Figure 2 shows a time-series plot of the topic proportions and latent variable proportions of two topics that have increasingly dominated recent addresses. The model infers that since around 1960, the two major United States political parties have become more associated with a different one of these two topics. According to the model, the Democrats are increasingly associated with a "welfare-state" topic focused on words such as "jobs" and "[health] care," while the Republicans have become increasingly associated with a "conservative" topic including words such as "budget" and "freedom." Another two interesting topics are plotted Figure 3. A "war" topic identifies the two World Wars and the Vietnam War, finding that Democrats focused on this topic more than Republicans before World War II. The reverse was true in the period beginning roughly after the Vietnam War. In each of these plots, the SOTU distribution varies more smoothly over time than the document-topic distribution Θ. Time-series plots of all topics are in the supplementary material.

We also evaluated the model's predictive performance at document completion and fully held-out document prediction tasks. Each document was shuffled and split into training and testing portions, with 50% of the words assigned to each portion. For the document completion task, the topic parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$ recovered on the training set were used to predict the test portions of the documents. This prediction task takes into account the time-series modeling performed by the LTN model on $\mathbf{\Theta}$. For the fully held-out prediction task, we treated the test portions of the addresses as completely unseen documents, and estimated the likelihood of each test document d marginalizing over $\theta^{(d)}$ using (Wallach et al., 2009)'s annealed importance sampling method with 2000 temperatures. The LTN outperformed LDA, trained via collapsed Gibbs sampling with similar settings, in terms of perplexity for both prediction tasks (Table 4). We also compared to another time-series model, the dynamic topic model (DTM) of Blei & Lafferty (2006), trained via variational inference. This model allows the topics to change over time, instead of the thetas. We allowed the topics to vary once per decade. To make a fair comparison, we interpret the variational posterior mean as a point estimate and evaluate it using the same procedure as for the LDA-based models. We found that the DTM exhibits poor performance on this data set, with likelihoods that were in one case worse than LDA. We hypothesize that the DTM is over-parameterized for this setting, and would perform better when given more documents.

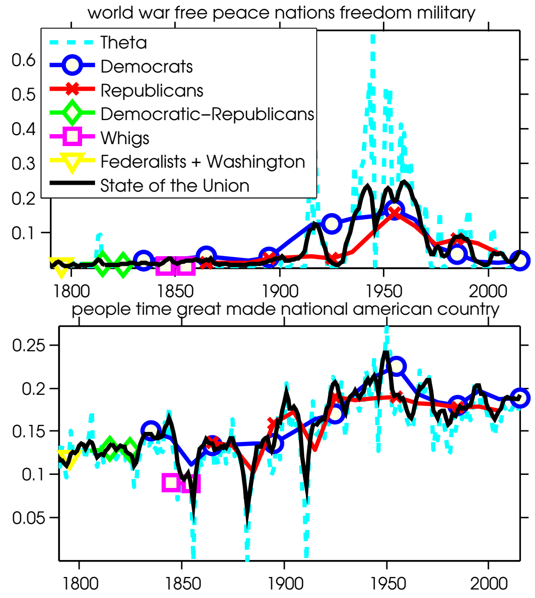

Figure 3. Topic proportions for the *war* and *nationalism* topics versus time.

Table 4. Predictive performance on the State of the Union data.

	Document Completion Perplexity	Fully Held-Out Perplexity
LTN	2.33×10^3	2.43×10^3
LDA	2.36×10^3	2.59×10^3
DTM	2.43×10^3	2.55×10^3

6. Related Work

A number of topic modeling frameworks have been developed which can encode domain knowledge. Topic models have been developed to make use of observed covariates (Mimno & McCallum, 2008) and relational information (Wahabzada et al., 2010), or to otherwise incorporate background knowledge (Andrzejewski et al., 2009) or seeding information (Jagarlamudi et al., 2012), though they do not facilitate the development of more sophisticated latent variable models (Table 1, top). Perhaps the most general of this class of models is the structural topic model (STM) of Roberts et al. (2013; 2014), which combines the ideas of DMR (Mimno & McCallum, 2008), the correlated topic model (Blei & Lafferty, 2007) and SAGE (Eisenstein et al., 2011).

More general probabilistic programming and graphical modeling systems for topic models have been proposed, but these approaches are limited in scalability when used in a general-purpose setting (Table 1, bottom). Conditional topic random fields (Zhu & Xing, 2010) connect topic models with CRF models. These models are scalable when the CRF is restricted to be a chain, but not in the general case. Logic-LDA (Mei et al., 2014) allows a modeler to specify constraints on the posterior distribution using logical rules, and does not facilitate the modeling of extra latent variables. Fully general languages such as Infer.net (Minka et al., 2014), Church (Goodman et al., 2008) and Stan (Stan Development Team, 2014) can potentially be applied to topic modeling, though methods that leverage the unique structure of topic models are likely to be more efficient.

Fold.all (Andrzejewski et al., 2011) is an important precursor to this work which uses Markov logic networks (Richardson & Domingos, 2006) to introduce domain knowledge and dependencies between the topic assignments z for each word. The size of these models is $O(N^U)$, where N is the number of words in the corpus and U is the largest number of universally quantified variables in a rule. Instead of modeling structure in the z's, LTNs shift the dependencies up a level in the hierarchy, using PSL to encode dependencies between document parameters θ, between topic parameters ϕ, and other latent variables. This results in many fewer groundings, which can be further restricted to a network of interest such as a citation graph. While MAP inference for an MLN is NP-hard, HL-MRF inference for the inner loop of LTN training is a convex optimization problem which can be solved efficiently using ADMM. The document-level modeling of LTNs may also often be more applicable than the word-level modeling of Fold.all, as we are more likely to have metadata for documents than for particular word indices in the corpus.

7. Conclusions

We have introduced latent topic networks (LTNs), a flexible topic modeling framework designed specifically to enable applied social science research. The framework allows the development of custom latent variable topic models using a probabilistic programming language with an intuitive logical syntax. We demonstrated the usefulness of the framework for several application domains. In our ongoing research, we plan to use latent topic networks to answer substantive questions in social science, the humanities, and cognitive science. One use-case of particular interest is the incorporation of ontological information and semantic graphs to improve the interpretability of the topics.

There are many possible extensions that we plan to explore. To simultaneously model word, document, and topic dependencies, it is likely possible to combine Fold.all and LTNs. Topic models differ from typical PSL models in that the key variables are constrained to the simplex, which motivates the development of new language primitives. Following Schiegg et al. (2012), another direction is to extend LTNs to specify domain knowledge and structure for nonparametric Bayesian models. We also anticipate that many algorithmic developments for LDA can be adapted to latent topic networks, including variational Bayesian methods, stochastic algorithms, and sampling techniques.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and suggestions. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337, and by the National Science Foundation under grant number IIS1218488. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, the National Science Foundation, or the U.S. Government.

References

Andrzejewski, D., Zhu, X., and Craven, M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 25–32. Omnipress, 2009.

Andrzejewski, D., Zhu, X., Craven, M., and Recht, B. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1171–1177, 2011.

Bach, S., Huang, B., London, B., and Getoor, L. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 32–41, 2013.

Bach, S. H., Broecheler, M., Huang, B., and Getoor, L. Hinge-loss Markov random fields and probabilistic soft logic. arXiv:1505.04406 [cs.LG], 2015.

Beltagy, I., Erk, K., and Mooney, R. Probabilistic soft logic for semantic textual similarity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1210–1219, 2014.

Blei, D.M. and Lafferty, J.D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 113–120, 2006.

Blei, D.M. and Lafferty, J.D. A correlated topic model of science. *The Annals of Applied Statistics*, pp. 17–35, 2007.

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

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D.M. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS) 22*, pp. 288–296, 2009.

Eisenstein, J., Ahmed, A., and Xing, E.P. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1041–1048, 2011.

Foulds, J. R. and Smyth, P. Modeling scientific impact with topical influence regression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 113–123, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Foulds, J. R. and Smyth, P. Annealing paths for the evaluation of topic models. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 220–229, 2014.

Gerrish, S. and Blei, D.M. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 489–496, 2011.

Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., and Tenenbaum, J. Church: a language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 220–229, Corvallis, Oregon, 2008. AUAI Press.

Griffiths, T.L., Steyvers, M., and Tenenbaum, J.B. Topics in semantic representation. *Psychological Review*, 114 (2):211–244, 2007.

Grimmer, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Grimmer, J. and Stewart, B.M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

Jagarlamudi, J., Daumé III, H., and Udupa, R. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 204–213. Association for Computational Linguistics, 2012.

Lucas, C., Nielsen, R., Roberts, M.E., Stewart, B.M., Storer, A., and Tingley, D. Computer assisted text analysis for comparative politics. *Political Analysis*, 23:254–277, 2015.

McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C. D, and Jurafsky, D. Differentiating language usage through topic models. *Poetics*, 41(6):607–625, 2013.

Mei, S., Zhu, J., and Zhu, X. Robust RegBayes: Selectively incorporating first-order logic domain knowledge into Bayesian models. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, pp. 253–261, 2014.

Mimno, D. and McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–418, 2008.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 262–272. Association for Computational Linguistics, 2011.

Minka, T., Winn, J.M., Guiver, J.P., Webster, S., Zaykov, Y., Yangel, B., Spengler, A., and Bronskill, J. Infer.NET 2.6, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Newman, D., Lau, J.H., Grieser, K., and Baldwin, T. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pp. 100–108. Association for Computational Linguistics, 2010.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Cai, D.A., Midberry, J.E., and Wang, Y. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421, 2014.

Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Roberts, M.E., Stewart, B.M., Tingley, D., and Airoldi, E.M. The structural topic model and applied social science. In *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., and Rand, D. Structural topic models for open ended survey responses. *American Journal of Political Science*, 58:1064–1082, 2014.

Schiegg, M., Neumann, M., and Kersting, K. Markov logic mixtures of Gaussian processes: Towards machines reading regression data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1002–1011, 2012.

Stan Development Team. Stan: A c++ library for probability and sampling, version 2.2, 2014. URL http://mc-stan.org/.

Wahabzada, M., Xu, Z., and Kersting, K. Topic models conditioned on relations. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 402–417. Springer, 2010.

Wallach, H.M., Murray, I., Salakhutdinov, R., and Mimno, D. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 1105–1112. ACM, 2009.

Zhu, J. and Xing, E.P. Conditional topic random fields. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 1239–1246, 2010.