SPRITE: Generalizing Topic Models with Structured Priors

Michael J. Paul   Mark Dredze

Department of Computer Science and Human Language Technology Center of Excellence

Johns Hopkins University, Baltimore, MD USA

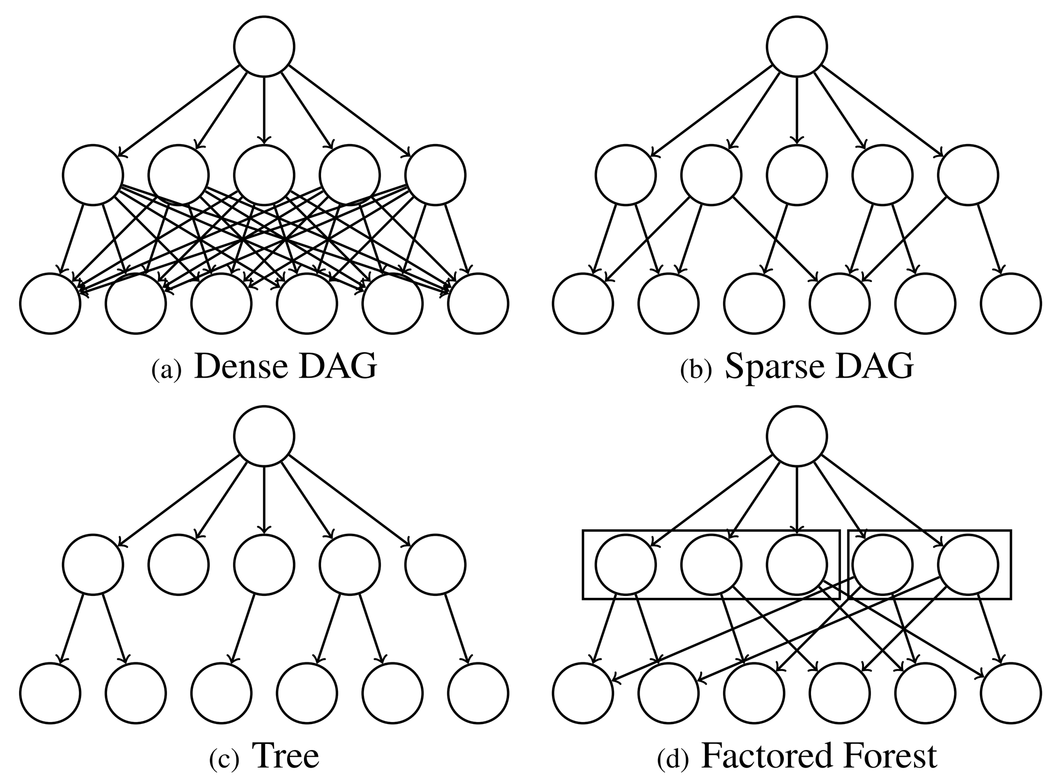
SPRITE: STRUCTURED-PRIOR TOPIC MODELS

SPRITE is based on LDA, but the Dirichlet priors are log-linear functions of underlying **components.** The components provide an additional level of latent structure that can model relations between topics.

1. Generate hyperparameters: $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}$
2. For each document m, generate parameters:
   (a) $\tilde{\theta}_{mt} = \exp(\sum_{c=1}^{C^{(\theta)}} \alpha_{mc}\,\delta_{ct}),\ 1 \le t \le T$
   (b) $\theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
3. For each topic t, generate parameters:
   (a) $\tilde{\phi}_{tv} = \exp(\sum_{c=1}^{C^{(\phi)}} \beta_{tc}\,\omega_{cv}),\ 1 \le v \le V$
   (b) $\phi_t \sim \text{Dirichlet}(\tilde{\phi}_t)$
4. For each token (n, m), generate data:
   (a) Topic (unobserved): $z_{mn} \sim \theta_m$
   (b) Word (observed): $w_{mn} \sim \phi_{z_{mn}}$

Hyperparameters	
ω_c	cth topic component (vector over words)
β_t	tth topic's component coefficients
δ_c	cth document component (vector over topics)
α_m	mth documents's component coefficients
Parameters	
ϕ_t	tth topic's distribution over words
θ_m	mth document's distribution over topics
Model size	
$C^{(\phi)}$	Number of topic components
$C^{(\theta)}$	Number of document components
T	Number of topics
Data size	
M	Number of documents
N_m	Number of tokens in mth document

TOPIC STRUCTURES



(a) Dense DAG (b) Sparse DAG (c) Tree (d) Factored Forest

Components can be combined in many different ways to form priors.

CONSTRAINTS

Different structures are induced by placing **constraints** on the values of α,β, such as indicator vector constraints:

$$\beta_{tc} \in \{0,1\}, \forall c \quad \sum_c \beta_{tc} = 1$$

We can **relax** these constraints for easier optimization, allowing real values but using a sparsity-inducing prior:

$$\beta_{tc} \in (0,1); \quad \beta_t \sim \text{Dirichlet}(<1)$$

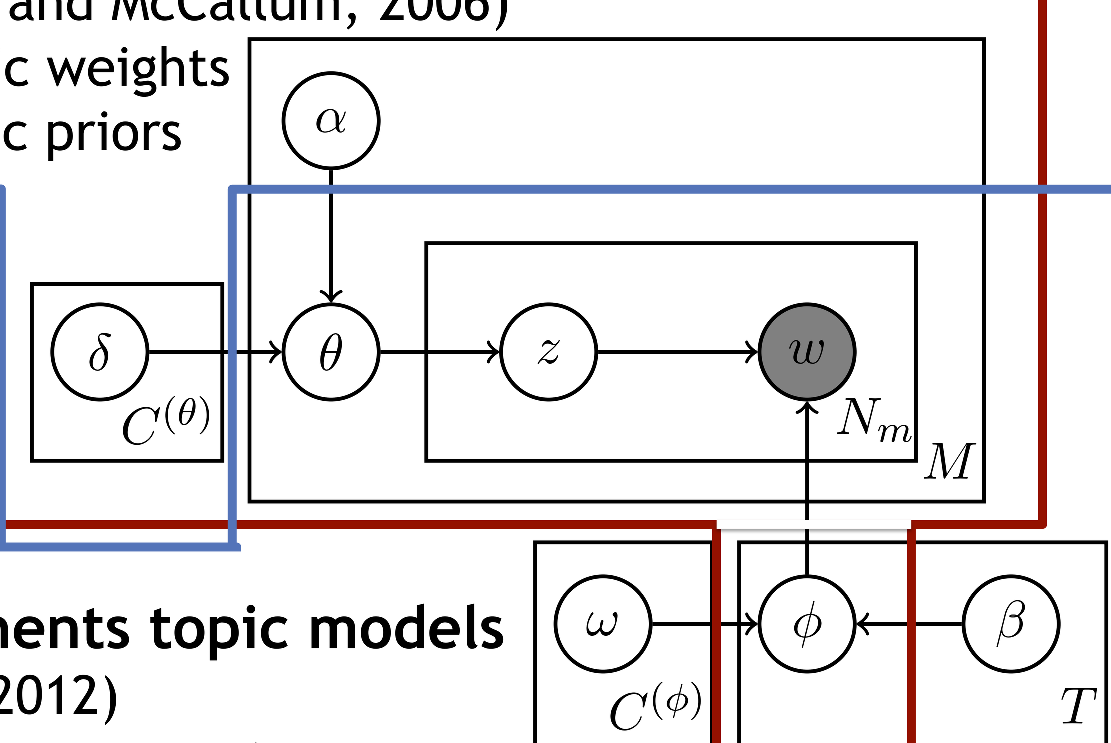
We can tighten these constraints during optimization using **annealing**.

RELATED MODELS

Dirichlet-multinomial regression (Mimno and McCallum, 2007)
- α are feature weights (supervision)
- δ are regression coefficients

Pachinko allocation (Li and McCallum, 2006)
- α behave like supertopic weights
- δ behave like supertopic priors

Factorial LDA (Paul and Dredze, 2012)
- β is transpose of δ

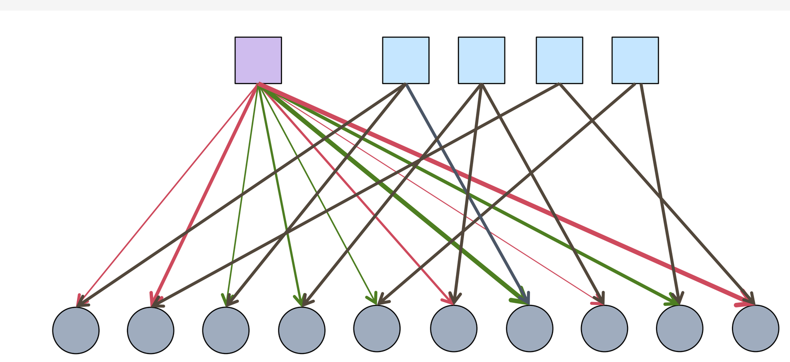


Shared components topic models (Gormley et al., 2012)
- ω behave like components
- β are binary
- $\phi = \tilde{\phi}$

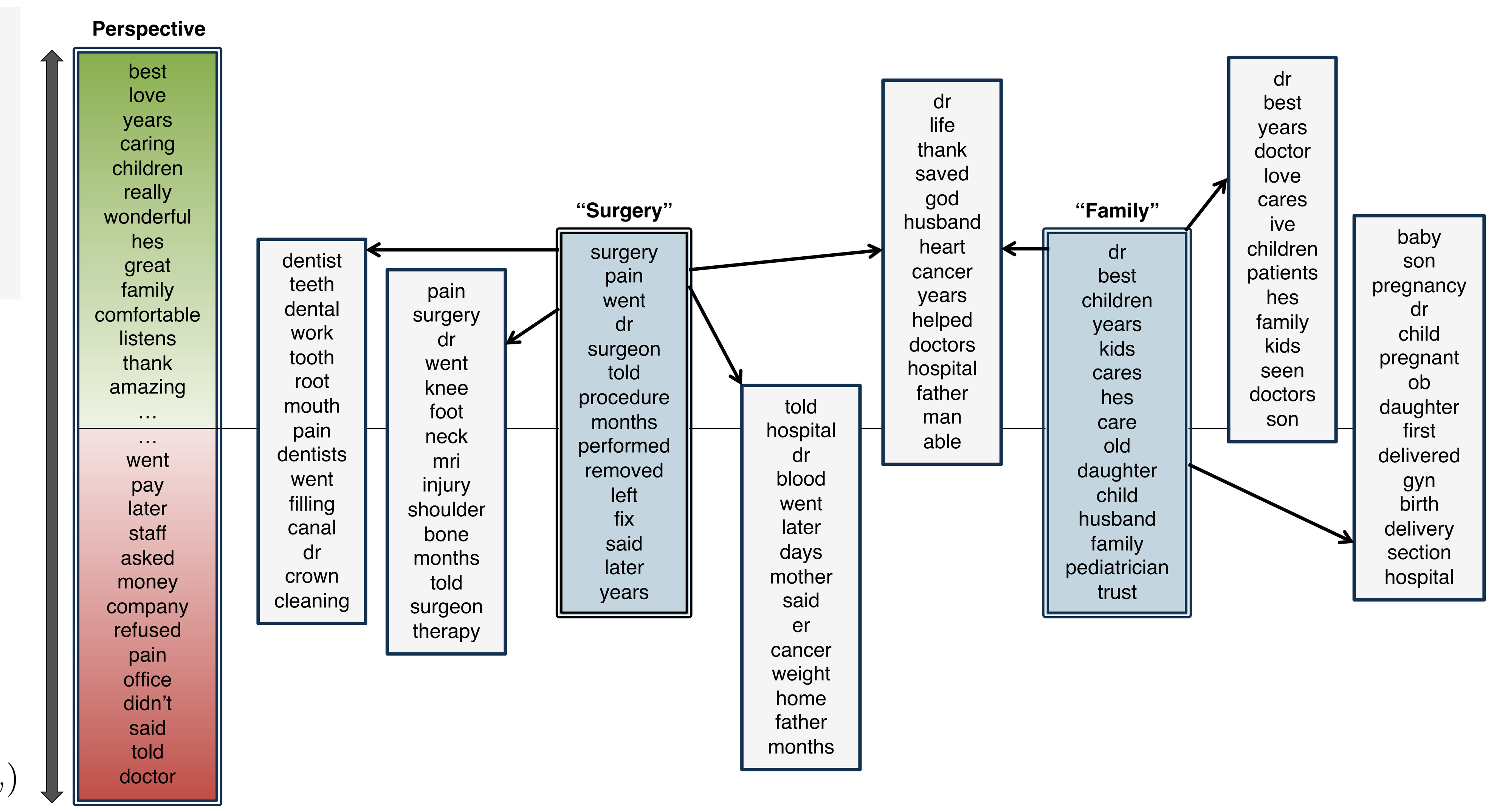
Sparse additive generative models (Eisenstein et al., 2011)
- ω are sparse
- β are pre-defined
- $\phi = \tilde{\phi}$

EXAMPLE: MODELING PERSPECTIVE AND TOPIC HIERARCHIES

Each topic's prior comes from two components:
- **Perspective** (positive or negative)
- **Hierarchy** (each topic chooses one parent component)

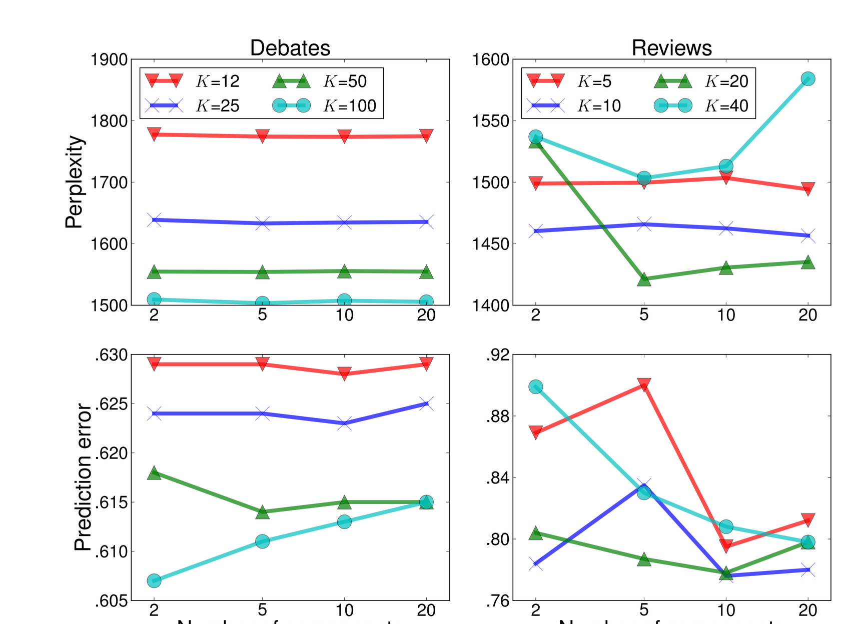


- $\mathbf{b}_k \sim \text{Dirichlet}(\rho < 1)$ (soft indicator)
- $\alpha^{(P)}$ is given as input (perspective value)
- $\delta_k^{(P)} = \beta_k^{(P)}$
- $\tilde{\phi}_{kv} = \exp(\omega_v^{(B)} + \beta_k^{(P)}\omega_v^{(P)} + \sum_c b_{kc}\hat{\beta}_{kc}\omega_{cv})$
- $\tilde{\theta}_{mk} = \exp(\delta_k^{(B)} + \alpha_m^{(P)}\delta_k^{(P)} + \sum_c b_{kc}\alpha_{mc}\hat{\delta}_{ck})$



		Debates			Reviews	
Model	Perplexity	Prediction error	Coherence	Perplexity	Prediction error	Coherence
Full model	†**1555.5** ± 2.3	†0.615 ± 0.001	-342.8 ± 0.9	†1421.3 ± 8.4	†**0.787** ± 0.006	-512.7 ± 1.6
Hierarchy only	†1561.8 ± 1.4	0.620 ± 0.002	-342.6 ± 1.1	†1457.2 ± 6.9	†0.804 ± 0.007	-509.1 ± 1.9
Perspective only	†1567.3 ± 2.3	†**0.613** ± 0.002	-342.1 ± 1.2	†**1413.7** ± 2.2	†0.800 ± 0.002	-512.0 ± 1.7
LDA	1579.6 ± 1.5	0.620 ± 0.001	-342.6 ± 0.6	1507.9 ± 2.4	0.846 ± 0.002	-501.4 ± 1.2

Sprite improves over several baselines (a sample is shown). Models with more structure are generally more predictive than those with less. Structured priors for topics, but not documents, improve coherence.



Different numbers of topics (K) and components

Code available: `http://cs.jhu.edu/~mpaul`