Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors

David Andrzejewski, Xiaojin Zhu, Mark Craven

University of Wisconsin–Madison

ICML 2009

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

New Year's Wishes

Goldberg et al 2009

- 89,574 New Year's wishes (NYC Times Square website)
- Example wishes:
  - Peace on earth
  - own a brewery
  - I hope I get into Univ. of Penn graduate school.
  - The safe return of my friends in Iraq
  - find a cure for cancer
  - To lose weight and get a boyfriend
  - I Hope Barack Obama Wins the Presidency
  - To win the lottery!

Topic Modeling of Wishes

Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .

- Use topic modeling to understand common wish themes
- Topic 13 mixes *college* and *illness* wish topics
- Want to **split** [go school into college] and [cancer free cure well]
- Resulting topics separate these words,

Topic Modeling of Wishes

Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .

- Use topic modeling to understand common wish themes
- Topic 13 mixes *college* and *illness* wish topics
- Want to **split** [go school into college] and [cancer free cure well]
- Resulting topics separate these words,

Topic Modeling of Wishes

Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .

- Use topic modeling to understand common wish themes
- Topic 13 mixes *college* and *illness* wish topics
- Want to **split** [go school into college] and [cancer free cure well]
- Resulting topics separate these words,

Topic Modeling of Wishes

Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .

Topic 13(a)	job go school great into good college . . . business graduate finish grades away law accepted . . .
Topic 13(b)	mom husband cancer hope free son well . . . full recovery surgery pray heaven pain aids . . .

- Use topic modeling to understand common wish themes
- Topic 13 mixes *college* and *illness* wish topics
- Want to **split** [go school into college] and [cancer free cure well]
- Resulting topics separate these words, as well as related words

Topic Modeling of Wishes

Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .

Topic 13(a)	job go school great into good college . . . business graduate finish grades away law accepted . . .
Topic 13(b)	mom husband cancer hope free son well . . . full recovery surgery pray heaven pain aids . . .

- Use topic modeling to understand common wish themes
- Topic 13 mixes *college* and *illness* wish topics
- Want to **split** [go school into college] and [cancer free cure well]
- Resulting topics separate these words, **as well as related words**

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate "school" from "cure")
  - Structured sources (encode Gene Ontology term "transcription factor activity")

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate “school” from “cure”)
  - Structured sources (encode Gene Ontology term “transcription factor activity”)

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate "school" from "cure")
  - Structured sources (encode Gene Ontology term "transcription factor activity")

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate “school” from “cure”)
  - Structured sources (encode Gene Ontology term “transcription factor activity”)

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate “school” from “cure”)
  - Structured sources (encode Gene Ontology term “transcription factor activity”)

Topic Modeling with Domain Knowledge

- Why domain knowledge?
  - Topics may not correspond to meaningful concepts
  - Topics may not align well with user modeling goals
- Possible sources of domain knowledge:
  - Human guidance (separate “school” from “cure”)
  - Structured sources (encode Gene Ontology term “transcription factor activity”)

Word Preferences within Topics

Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

- Need a suitable "language" for expressing our preferences
- Pairwise "primitives" $\rightarrow$ higher-level operations

Word Preferences within Topics

Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

- Need a suitable "language" for expressing our preferences
- Pairwise "primitives" $\rightarrow$ higher-level operations

Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school \mid t) \approx P(college \mid t)$

Word Preferences within Topics

Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

- Need a suitable "language" for expressing our preferences
- Pairwise "primitives" $\rightarrow$ higher-level operations

Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school \mid t) \approx P(college \mid t)$
Cannot-Link (school,cure)	no topic t has $P(school \mid t)$ and $P(cure \mid t)$ both high

Word Preferences within Topics

Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school\|t) \approx P(college\|t)$
Cannot-Link (school,cure)	no topic t has $P(school\|t)$ and $P(cure\|t)$ both high

split	[*go school into college*] vs [*cancer free cure well*] $\rightarrow$ Must-Link among words for each concept $\rightarrow$ Cannot-Link between words from different concepts

Word Preferences within Topics

Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school\|t) \approx P(college\|t)$
Cannot-Link (school,cure)	no topic t has $P(school\|t)$ and $P(cure\|t)$ both high

split	[*go school into college*] vs [*cancer free cure well*] $\rightarrow$ Must-Link among words for each concept $\rightarrow$ Cannot-Link between words from different concepts
merge	[*love marry together boyfriend*] in one topic [*married boyfriend engaged wedding*] in another $\rightarrow$ Must-Link among concept words

Word Preferences within Topics

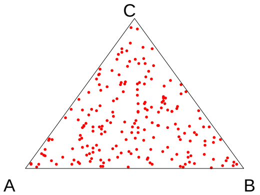
Inspired by constrained clustering (Basu, Davidson, & Wagstaff 2008)

Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school\|t) \approx P(college\|t)$
Cannot-Link (school,cure)	no topic t has $P(school\|t)$ and $P(cure\|t)$ both high

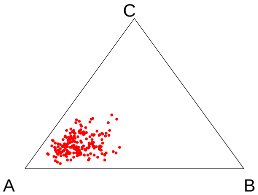
split	[*go school into college*] vs [*cancer free cure well*] $\rightarrow$ Must-Link among words for each concept $\rightarrow$ Cannot-Link between words from different concepts
merge	[*love marry together boyfriend*] in one topic [*married boyfriend engaged wedding*] in another $\rightarrow$ Must-Link among concept words
isolate	[*the year in 2008*] in many wish topics $\rightarrow$ Must-Link among words to be isolated $\rightarrow$ Cannot-Link vs other Top N words for each topic

Dirichlet Prior ("dice factory")

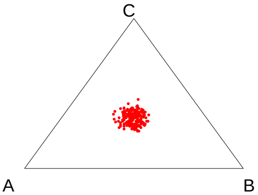
- $P(\phi|\beta)$ for K-dimensional multinomial parameter ϕ
- K-dimensional hyperparameter β ("pseudocounts")



$\beta = [1, 1, 1]$



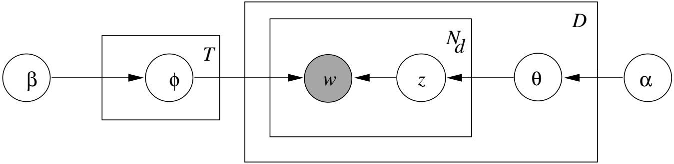
$\beta = [20, 5, 5]$



$\beta = [50, 50, 50]$

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic *t*

$\phi_t \sim$ Dirichlet(β)

For each doc *d*

$\theta_d \sim$ Dirichlet(α)

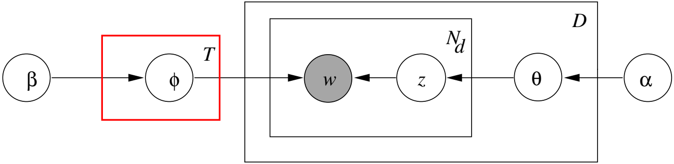
For each word *w*

$z \sim$ Multinomial(θ_d)

$w \sim$ Multinomial(ϕ_z)

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim$ Dirichlet(β)

For each doc d

$\theta_d \sim$ Dirichlet(α)

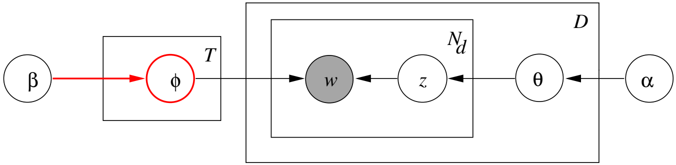
For each word w

$z \sim$ Multinomial(θ_d)

$w \sim$ Multinomial(ϕ_z)

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim$ Dirichlet(β)

For each doc d

$\theta_d \sim$ Dirichlet(α)

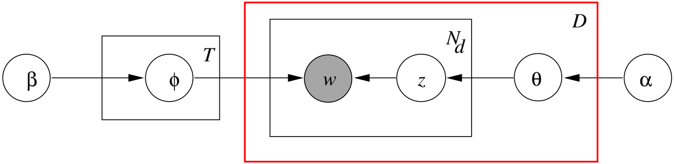
For each word w

$z \sim$ Multinomial(θ_d)

$w \sim$ Multinomial(ϕ_z)

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim \text{Dirichlet}(\beta)$

For each doc d

$\theta_d \sim \text{Dirichlet}(\alpha)$

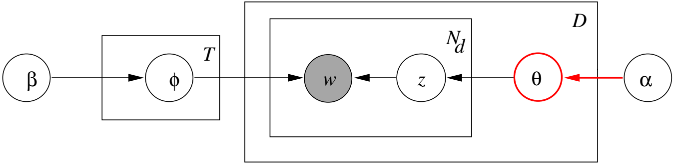
For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim \text{Dirichlet}(\beta)$

For each doc d

$\theta_d \sim \text{Dirichlet}(\alpha)$

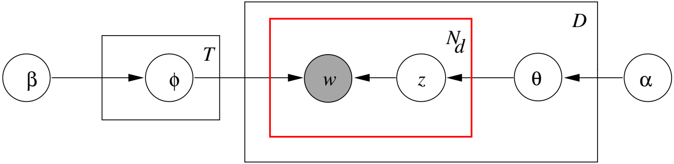
For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim \text{Dirichlet}(\beta)$

For each doc d

$\theta_d \sim \text{Dirichlet}(\alpha)$

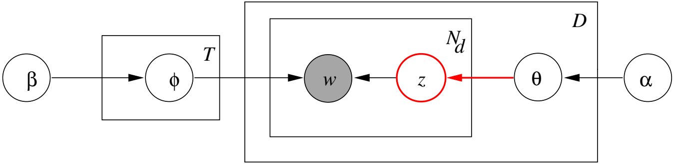
For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic t

$\phi_t \sim \text{Dirichlet}(\beta)$

For each doc d

$\theta_d \sim \text{Dirichlet}(\alpha)$

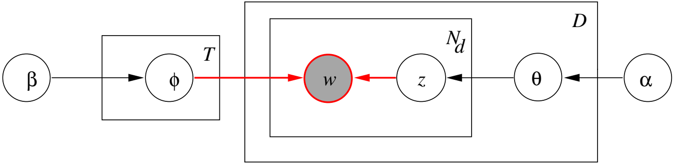
For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan 2003



For each topic *t*

$\phi_t \sim \text{Dirichlet}(\beta)$

For each doc *d*

$\theta_d \sim \text{Dirichlet}(\alpha)$

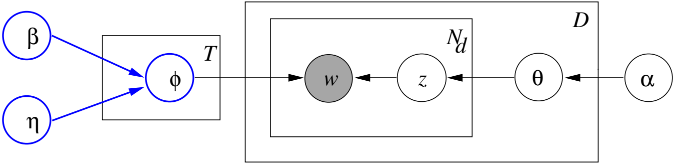
For each word *w*

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

LDA with Dirichlet Forest Prior

This work



For each topic t

$\cancel{\phi_t \sim \text{Dirichlet}(\beta)}$ $\phi_t \sim \text{DirichletForest}(\beta, \eta)$

For each doc d

$\theta_d \sim \text{Dirichlet}(\alpha)$

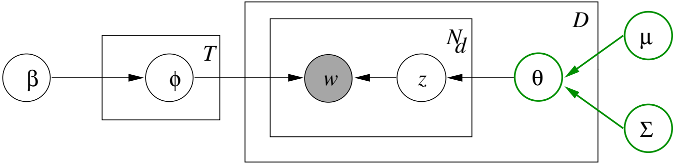
For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$

Related work: Correlated Topic Model (CTM)

Blei and Lafferty 2006



For each topic t

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each doc d

$$\cancel{\theta_d \sim \text{Dirichlet}(\alpha)} \; \theta_d \sim \text{LogisticNormal}(\mu, \Sigma)$$

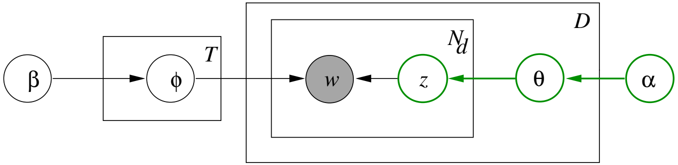
For each word w

$$z \sim \text{Multinomial}(\theta_d)$$

$$w \sim \text{Multinomial}(\phi_z)$$

Related work: Pachinko Allocation Model (PAM)

Li and McCallum 2006



For each topic *t*

$\phi_t \sim$ Dirichlet(β)

For each doc *d*

$\cancel{\theta_d \sim \text{Dirichlet}(\alpha)}$ Pachinko$(\theta_d) \sim$ Dirichlet-DAG(α)

For each word *w*

$\cancel{z \sim \text{Multinomial}(\theta_d)}$ $z \sim$ Pachinko(θ_d)

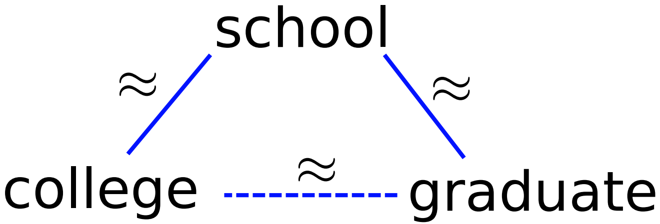
$w \sim$ Multinomial(ϕ_z)

Must-Link (*college*,*school*)

- $\forall t$, we want $P(college|t) \approx P(school|t)$
- Must-Link is **transitive**
- Cannot be encoded by a single Dirichlet

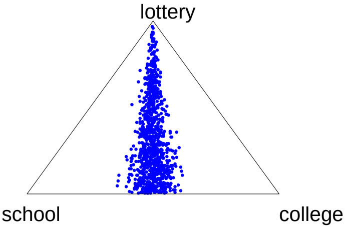
Must-Link (*college*,*school*)

- $\forall t$, we want $P(college|t) \approx P(school|t)$
- Must-Link is **transitive**
- Cannot be encoded by a single Dirichlet

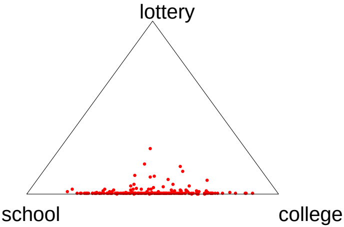


Must-Link (*college*,*school*)

- $\forall t$, we want $P(college|t) \approx P(school|t)$
- Must-Link is **transitive**
- Cannot be encoded by a single Dirichlet



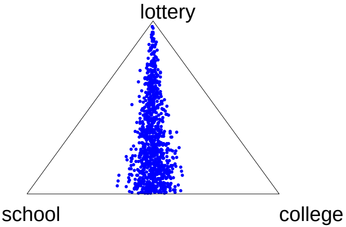
Goal



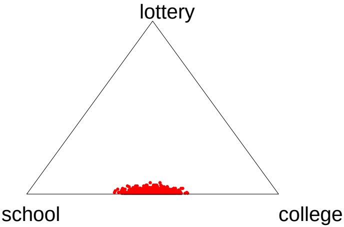
$\beta = [5, 5, 0.1]$

Must-Link (*college*,*school*)

- $\forall t$, we want $P(college|t) \approx P(school|t)$
- Must-Link is **transitive**
- Cannot be encoded by a single Dirichlet



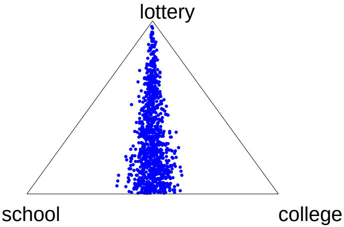
Goal



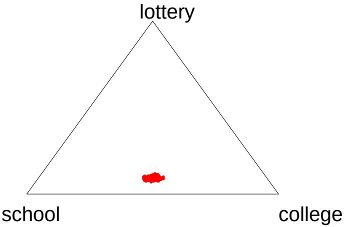
$\beta = [50, 50, 1]$

Must-Link (*college*,*school*)

- $\forall t$, we want $P(college|t) \approx P(school|t)$
- Must-Link is **transitive**
- Cannot be encoded by a single Dirichlet



Goal

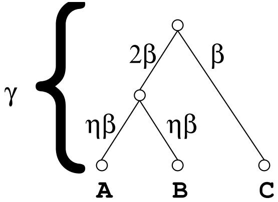


$\beta = [500, 500, 100]$

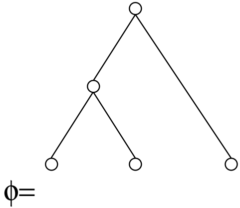
Dirichlet Tree ("dice factory 2.0")

Dennis III 1991, Minka 1999

- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node $s \rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out ("collapse") ϕ



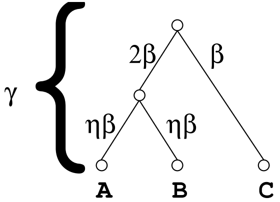
$(\beta = 1, \eta = 50)$



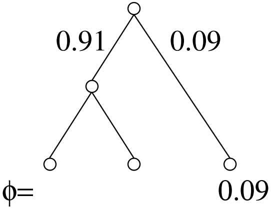
Dirichlet Tree ("dice factory 2.0")

Dennis III 1991, Minka 1999

- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node $s \rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out ("collapse") ϕ



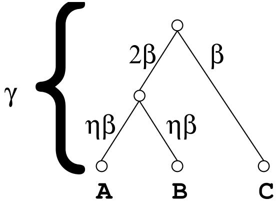
$(\beta = 1, \eta = 50)$



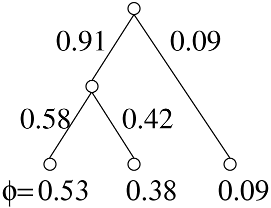
Dirichlet Tree (“dice factory 2.0”)

Dennis III 1991, Minka 1999

- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node $s \rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out (“collapse”) ϕ



$(\beta = 1, \eta = 50)$



Dirichlet Tree ("dice factory 2.0")

Dennis III 1991, Minka 1999

- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node $s \rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out ("collapse") ϕ

$$p(\phi|\gamma) =$$

$$\left(\prod_k^L {\phi^{(k)}}^{\gamma^{(k)}-1}\right)\left(\prod_s^I \frac{\Gamma\left(\sum_k^{C(s)} \gamma^{(k)}\right)}{\prod_k^{C(s)} \Gamma\left(\gamma^{(k)}\right)} \left(\sum_k^{L(s)} \phi^{(k)}\right)^{\Delta(s)}\right)$$

- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node s $\rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out ("collapse") ϕ

$$p(\phi|\gamma) =$$

$$\left(\prod_k^L {\phi^{(k)}}^{\gamma^{(k)}-1}\right)\left(\prod_s^I \frac{\Gamma\left(\sum_k^{C(s)}\gamma^{(k)}\right)}{\prod_k^{C(s)}\Gamma\left(\gamma^{(k)}\right)}\left(\sum_k^{L(s)}\phi^{(k)}\right)^{\Delta(s)}\right)$$

Dirichlet Tree ("dice factory 2.0")

Dennis III 1991, Minka 1999

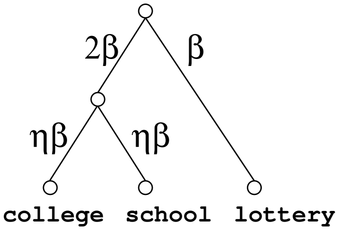
- Control variance of *subsets* of variables
  - Sample Dirichlet(γ) at parent, distribute mass to children
  - Mass reaching leaves are final multinomial parameters ϕ
  - $\Delta(s) = 0$ for all internal node $s \rightarrow$ standard Dirichlet (for our trees, true when $\eta = 1$)
  - Conjugate to multinomial, can integrate out ("collapse") ϕ

$$p(\mathbf{w}|\gamma) =$$

$$\prod_s^I \left(\frac{\Gamma\left(\sum_k^{C(s)} \gamma^{(k)}\right)}{\Gamma\left(\sum_k^{C(s)} \left(\gamma^{(k)} + n^{(k)}\right)\right)} \prod_k^{C(s)} \frac{\Gamma\left(\gamma^{(k)} + n^{(k)}\right)}{\Gamma(\gamma^{(k)})} \right)$$

Must-Link (*school*,*college*) via Dirichlet Tree

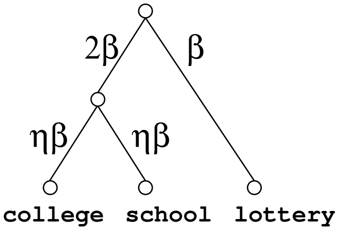
- Place (*school*,*college*) beneath internal node
- Large edge weights beneath this node (large η)

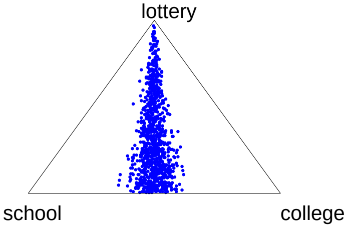


$(\beta = 1, \eta = 50)$

Must-Link (*school*,*college*) via Dirichlet Tree

- Place (*school*,*college*) beneath internal node
- Large edge weights beneath this node (large η)

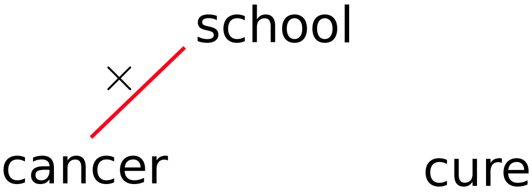




$(\beta = 1, \eta = 50)$

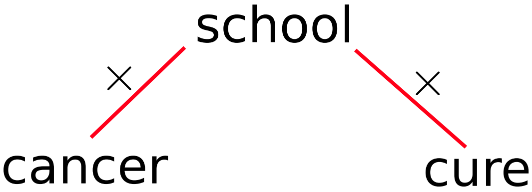
Cannot-Link (*school*,*cancer*)

- Do not want words to co-occur as high-probability for any topic
- No topic-word multinomial $\phi_t = P(w|t)$ should have:
  - High probability $P(school|t)$
  - High probability $P(cancer|t)$
- Cannot-Link is **non-transitive**
- Cannot be encoded by single Dirichlet/DirichletTree
- Will require **mixture** of Dirichlet Trees (Dirichlet Forest)



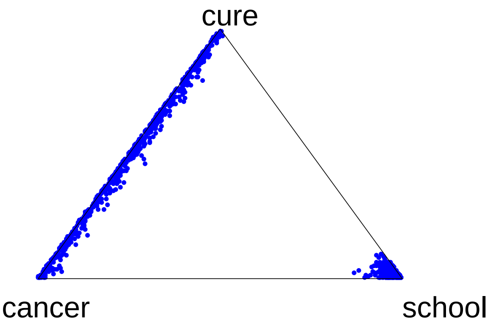
Cannot-Link (*school*,*cancer*)

- Do not want words to co-occur as high-probability for any topic
- No topic-word multinomial $\phi_t = P(w|t)$ should have:
  - High probability $P(school|t)$
  - High probability $P(cancer|t)$
- Cannot-Link is **non-transitive**
- Cannot be encoded by single Dirichlet/DirichletTree
- Will require **mixture** of Dirichlet Trees (Dirichlet Forest)



Cannot-Link (*school*,*cancer*)

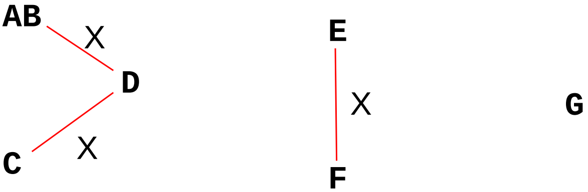
- Do not want words to co-occur as high-probability for any topic
- No topic-word multinomial $\phi_t = P(w|t)$ should have:
  - High probability $P(school|t)$
  - High probability $P(cancer|t)$
- Cannot-Link is **non-transitive**
- Cannot be encoded by single Dirichlet/DirichletTree
- Will require **mixture** of Dirichlet Trees (Dirichlet Forest)



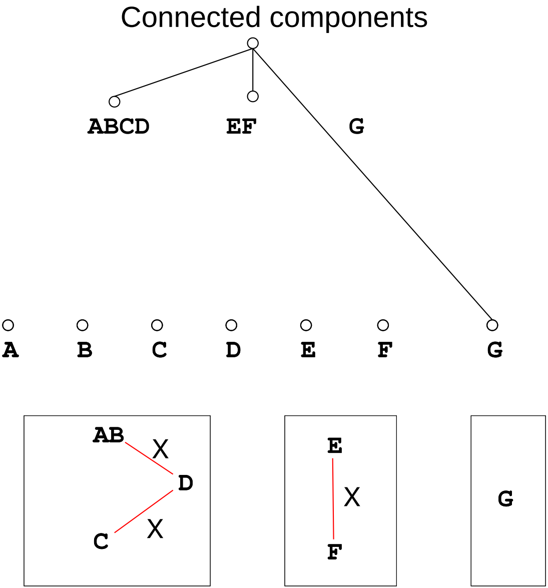
Sampling a Tree from the Forest

Vocabulary	$[A, B, C, D, E, F, G]$
Must-Links	(A, B)
Cannot-Links	(A, D), (C, D), (E, F)

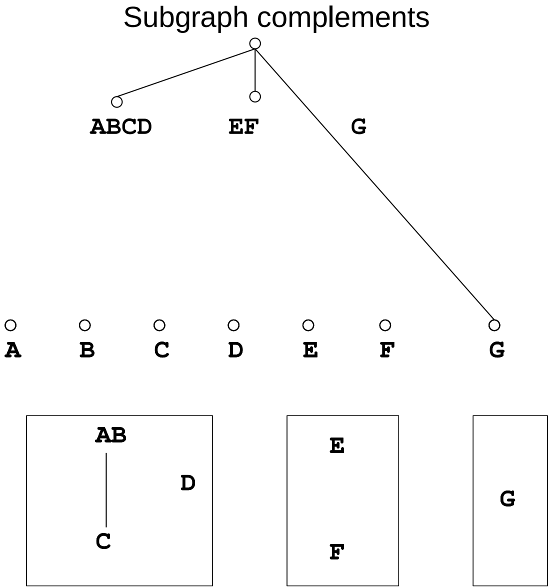
Cannot-Link-graph



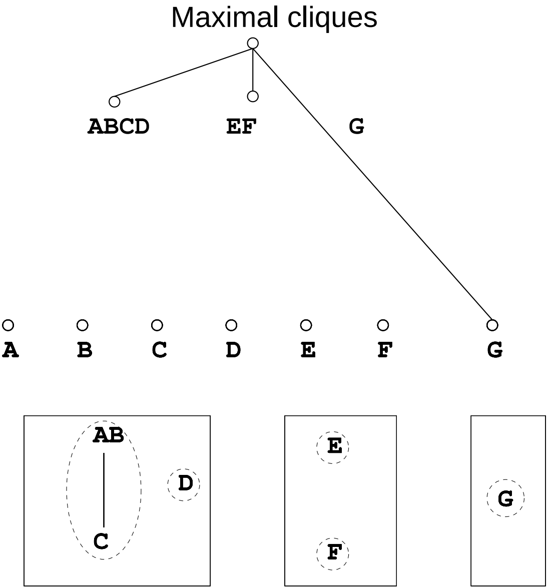
Sampling a Tree from the Forest



Sampling a Tree from the Forest

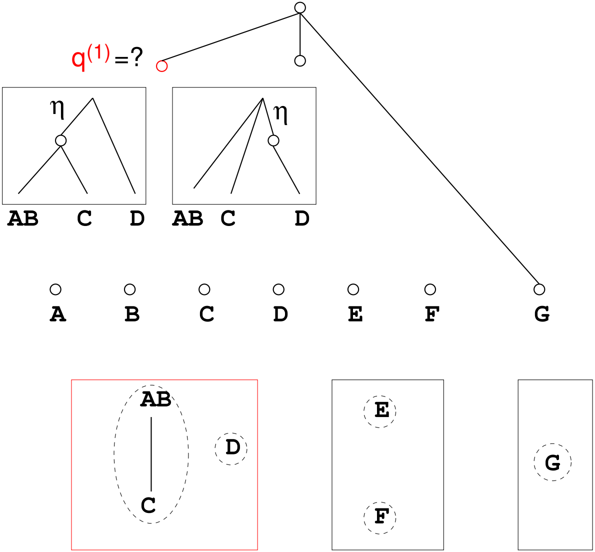


Sampling a Tree from the Forest

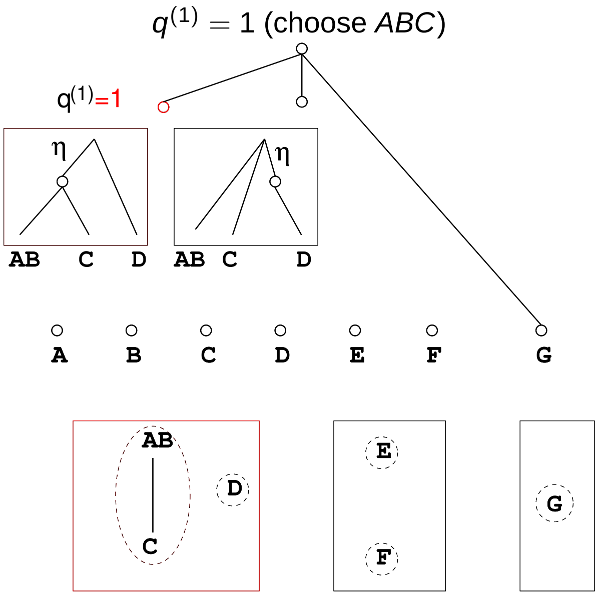


Sampling a Tree from the Forest

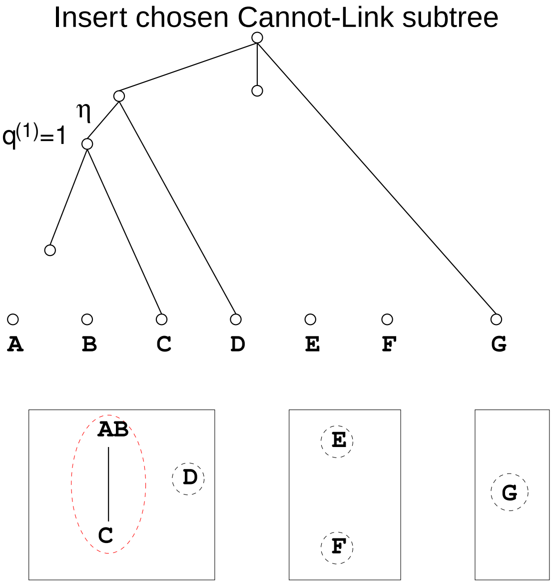
Sample $q^{(1)}$ for first connected component



Sampling a Tree from the Forest

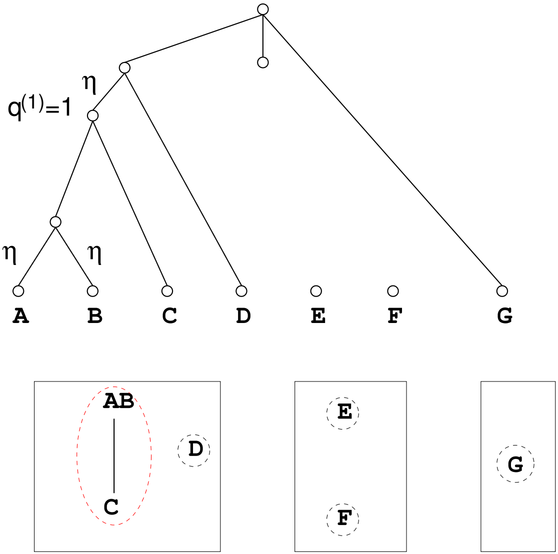


Sampling a Tree from the Forest



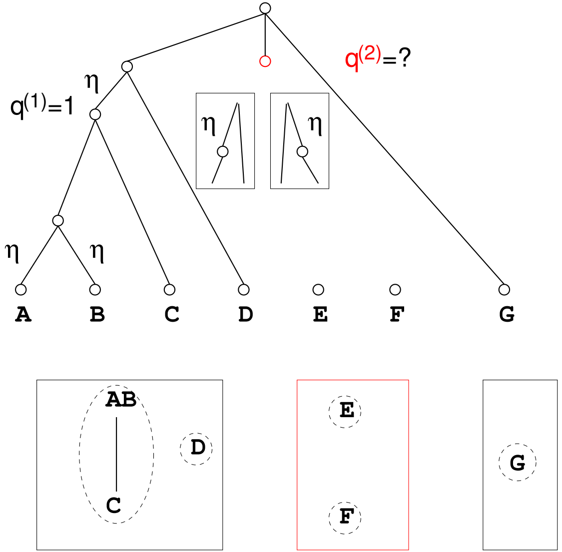
Sampling a Tree from the Forest

Put (A, B) under Must-Link subtree

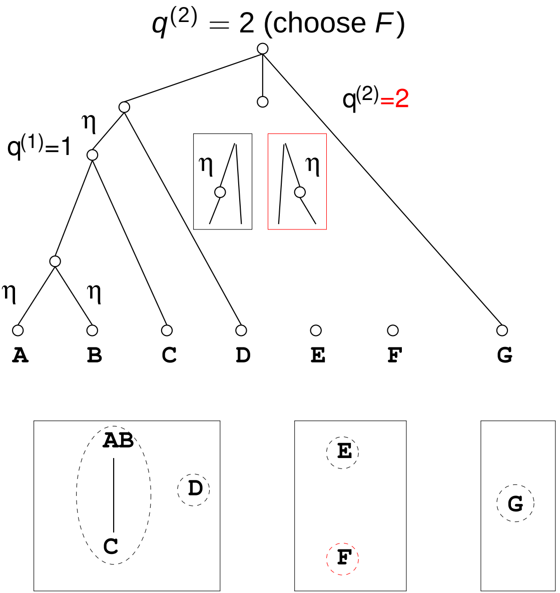


Sampling a Tree from the Forest

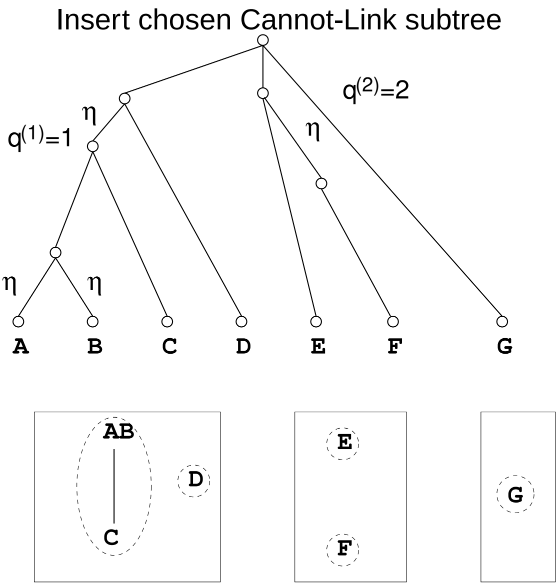
Sample $q^{(2)}$ for second connected component



Sampling a Tree from the Forest



Sampling a Tree from the Forest



LDA with Dirichlet Forest Prior

For each topic $t = 1 \ldots T$

For each Cannot-Link-graph connected component $r = 1 \ldots R$

Sample $q_t^{(r)} \propto$ clique sizes

$\phi_t \sim \text{DirichletTree}(\mathbf{q}_t, \beta, \eta)$

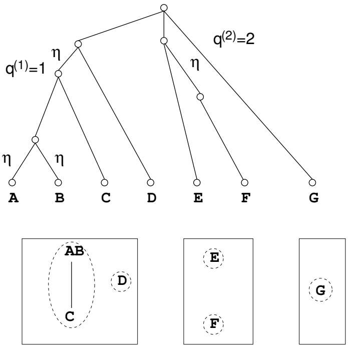
For each doc $d = 1 \ldots D$

$\theta_d \sim \text{Dirichlet}(\alpha)$

For each word w

$z \sim \text{Multinomial}(\theta_d)$

$w \sim \text{Multinomial}(\phi_z)$



Collapsed Gibbs Sampling of $(\mathbf{z}, \mathbf{q})$

Complete Gibbs sample: $z_1 \ldots z_N, q_1^{(1)} \ldots q_1^{(R)}, \ldots, q_T^{(1)} \ldots q_T^{(R)}$
Sample z_i for each word position i in corpus

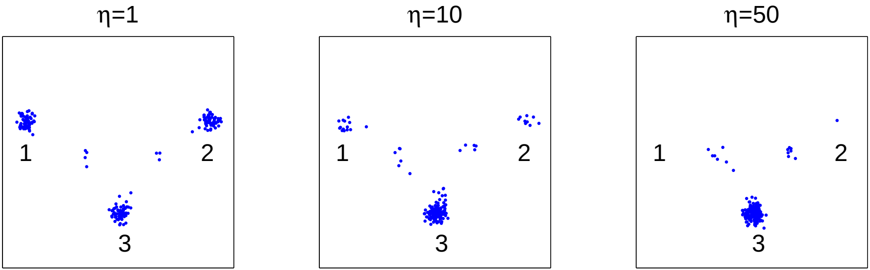
$$p(z_i = v | \mathbf{z}_{-i}, \mathbf{q}_{1:T}, \mathbf{w}) \propto (n_{-i,v}^{(d)} + \alpha) \prod_{s}^{I_v(\uparrow i)} \frac{\gamma_v^{(C_v(s\downarrow i))} + n_{-i,v}^{(C_v(s\downarrow i))}}{\sum_k^{C_v(s)} \left(\gamma_v^{(k)} + n_{-i,v}^{(k)} \right)}$$

Sample $q_j^{(r)}$ for each topic j and component r

$$p(q_j^{(r)} = q' | \mathbf{z}, \mathbf{q}_{-j}, \mathbf{q}_j^{(-r)}, \mathbf{w}) \propto$$
$$\left(\sum_k^{M_{rq'}} \beta_k \right) \prod_s^{I_{j,r=q'}} \left(\frac{\Gamma\left(\sum_k^{C_j(s)} \gamma_j^{(k)} \right)}{\Gamma\left(\sum_k^{C_j(s)} (\gamma_j^{(k)} + n_j^{(k)}) \right)} \prod_k^{C_j(s)} \frac{\Gamma(\gamma_j^{(k)} + n_j^{(k)})}{\Gamma(\gamma_j^{(k)})} \right)$$

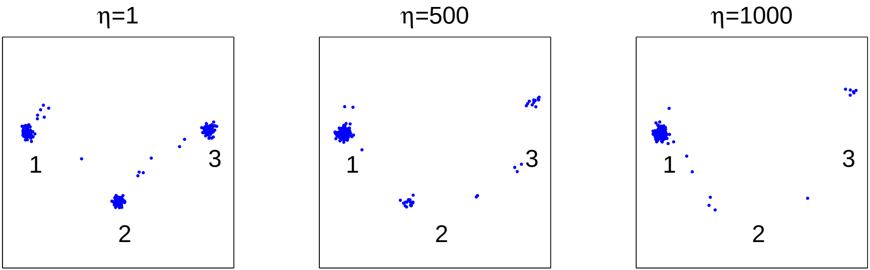
Synthetic Data - Must-Link (B,C)

- Prior knowledge: B and C should be in the same topic
- Corpus: ABAB, CDCD, EEEE, ABAB, CDCD, EEEE
- Standard LDA topics $[\phi_1, \phi_2]$ do *not* put (B, C) together
  1. $[\phi_1 = AB, \phi_2 = CDE]$
  2. $[\phi_1 = ABE, \phi_2 = CD]$
  3. $[\phi_1 = ABCD, \phi_2 = E]$
- As η increases, Must-Link (B,C) $\rightarrow [\phi_1 = ABCD, \phi_2 = E]$



Synthetic Data - **isolate**(B)

- Prior knowledge: B should be **isolated** from [A,C]
- Corpus: ABC, ABC, ABC, ABC
- Standard LDA topics $[\phi_1, \phi_2]$ do *not* isolate B
  1. $[\phi_1 = AC, \phi_2 = B]$
  2. $[\phi_1 = A \ \phi_2 = BC]$
  3. $[\phi_1 = AB, \phi_2 = C]$
- As η increases, Cannot-Link (A,B)+Cannot-Link (B,C) $\rightarrow [\phi_1 = AC, \phi_2 = B]$



Original Wish Topics

Topic	Top words sorted by $\phi = p(\text{word}\mid\text{topic})$
0	love i you me and will forever that with hope
1	and health for happiness family good my friends
2	year new happy a this have and everyone years
3	that is it you we be t are as not s will can
4	my to get job a for school husband s that into
5	to more of be and no money stop live people
6	to our the home for of from end safe all come
7	to my be i find want with love life meet man
8	a and healthy my for happy to be have baby
9	a 2008 in for better be to great job president
10	i wish that would for could will my lose can
11	peace and for love all on world earth happiness
12	may god in all your the you s of bless 2008
13	the in to of world best win 2008 go lottery
14	me a com this please at you call 4 if 2 www

Original Wish Topics

Topic	Top words sorted by $\phi = p(\text{word}\mid\text{topic})$
0	love i you me and will forever that with hope
1	and health for happiness family good my friends
2	year new happy a this have and everyone years
3	that is it you we be t are as not s will can
4	my to get job a for school husband s that into
5	to more of be and no money stop live people
6	to our the home for of from end safe all come
7	to my be i find want with love life meet man
8	a and healthy my for happy to be have baby
9	a 2008 in for better be to great job president
10	i wish that would for could will my lose can
11	peace and for love all on world earth happiness
12	may god in all your the you s of bless 2008
13	the in to of world best win 2008 go lottery
14	me a com this please at you call 4 if 2 www

isolate([to and for] . . .)

50 stopwords vs Top 50 in existing topics

Topic	Top words sorted by $\phi = p(\text{word}\|\text{topic})$
0	love forever marry happy together mom back
1	health happiness good family friends prosperity
2	life best live happy long great time ever wonderful
3	out not up do as so what work don was like
4	go school cancer into well free cure college
5	no people stop less day every each take children
6	home safe end troops iraq bring war husband house
7	love peace true happiness hope joy everyone dreams
8	happy healthy family baby safe prosperous everyone
9	better job hope president paul great ron than person
10	make money lose weight meet finally by lots hope married
Isolate	and to for a the year in new all my 2008
12	god bless jesus loved know everyone love who loves
13	peace world earth win lottery around save
14	com call if 4 2 www u visit 1 3 email yahoo
Isolate	i to wish my for and a be that the in

isolate([to and for] . . .)

50 stopwords vs Top 50 in existing topics

Topic	Top words sorted by $\phi = p(\text{word}\mid\text{topic})$
0	love forever marry happy together mom back
1	health happiness good family friends prosperity
2	life best live happy long great time ever wonderful
3	out not up do as so what work don was like
MIXED	go school cancer into well free cure college
5	no people stop less day every each take children
6	home safe end troops iraq bring war husband house
7	love peace true happiness hope joy everyone dreams
8	happy healthy family baby safe prosperous everyone
9	better job hope president paul great ron than person
10	make money lose weight meet finally by lots hope married
`Isolate`	and to for a the year in new all my 2008
12	god bless jesus loved know everyone love who loves
13	peace world earth win lottery around save
14	com call if 4 2 www u visit 1 3 email yahoo
`Isolate`	i to wish my for and a be that the in

split([cancer free cure well],[go school into college])

0	love forever happy together marry fall
1	health happiness family good friends
2	life happy best live love long time
3	as not do so what like much don was
4	out make money house up work grow able
5	people no stop less day every each take
6	home safe end troops iraq bring war husband
7	love peace happiness true everyone joy
8	happy healthy family baby safe prosperous
9	better president hope paul ron than person
10	lose meet man hope boyfriend weight finally
Isolate	and to for a the year in new all my 2008
12	god bless jesus loved everyone know loves
13	peace world earth win lottery around save
14	com call if 4 www 2 u visit 1 email yahoo 3
Isolate	i to wish my for and a be that the in me get
Split	job go school great into good college
Split	mom husband cancer hope free son well

split([cancer free cure well],[go school into college])

LOVE	love forever happy together marry fall
1	health happiness family good friends
2	life happy best live love long time
3	as not do so what like much don was
4	out make money house up work grow able
5	people no stop less day every each take
6	home safe end troops iraq bring war husband
7	love peace happiness true everyone joy
8	happy healthy family baby safe prosperous
9	better president hope paul ron than person
LOVE	lose meet man hope boyfriend weight finally
Isolate	and to for a the year in new all my 2008
12	god bless jesus loved everyone know loves
13	peace world earth win lottery around save
14	com call if 4 www 2 u visit 1 email yahoo 3
Isolate	i to wish my for and a be that the in me get
Split	job go school great into good college
Split	mom husband cancer hope free son well

merge([love . . . marry. . .],[meet . . . married. . .])
(10 words total)

Topic	Top words sorted by $\phi = p(\text{word}\mid\text{topic})$
Merge	love lose weight together forever marry meet
success	health happiness family good friends prosperity
life	life happy best live time long wishes ever years
-	as do not what someone so like don much he
money	out make money up house work able pay own lots
people	no people stop less day every each other another
iraq	home safe end troops iraq bring war return
joy	love true peace happiness dreams joy everyone
family	happy healthy family baby safe prosperous
vote	better hope president paul ron than person bush
Isolate	and to for a the year in new all my
god	god bless jesus everyone loved know heart christ
peace	peace world earth win lottery around save
spam	com call if u 4 www 2 3 visit 1
Isolate	i to wish my for and a be that the
Split	job go great school into good college hope move
Split	mom hope cancer free husband son well dad cure

Conclusions/Acknowledgments

- Conclusions
  - DF prior expresses pairwise preferences among words
  - Can efficiently sample from DF-LDA posterior
  - Topics obey preferences, capture structure
- Future work
  - Hierarchical domain knowledge
  - Quantify benefits on tasks
  - Other application domains
- Code
  - http://www.cs.wisc.edu/~andrzeje/research/df_lda.html
- Funding

  - Wisconsin Alumni Research Foundation (WARF)
  - NIH/NLM grants T15 LM07359 and R01 LM07050
  - ICML student travel scholarship