Adding Domain Knowledge to Latent Topic Models

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

2011
Acknowledgments

David M. Andrzejewski
did most work; now postdoc at Livermore National Laboratory

Mark Craven    Ben Liblit    Ben Recht
University of Wisconsin-Madison

The Wisconsin Alumni Research Foundation, NSF CAREER Award IIS-0953219,
AFOSR FA9550-09-1-0313, and NIH/NLM R01 LM07050
New Year’s Eve, Times Square
The Wish Corpus

[Goldberg et al., NAACL 2009]

- 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!
Corpus-wide word frequencies
Some Topics by Latent Dirichlet Allocation (LDA)

[Blei et al., JMLR 2003]

\[ p(\text{word} | \text{topic}) \]

“troops”

“election”

“love”
Some Topics by Latent Dirichlet Allocation (LDA)
[Blei et al., JMLR 2003]

\[ p(\text{word} \mid \text{topic}) \]

Major applications: exploratory data analysis
- Research trends [Wang & McCallum, 2006]
- Scientific influence [Gerrish & Blei, 2009]
- Matching papers to reviewers [Mimno & McCallum, 2007]
Quick Statistics Review

Dirichlet
[20, 5, 5]
Quick Statistics Review

Dirichlet
\[ [20, 5, 5] \]

Multinomial
\[ [0.6, 0.15, 0.25] \]
Quick Statistics Review

Dirichlet
[20, 5, 5]

Multinomial
[0.6, 0.15, 0.25]

Observed counts
[3, 1, 2]

A, A, B, C, A, C
Quick Statistics Review

Dirichlet
[20, 5, 5]

Multinomial
[0.6, 0.15, 0.25]

Observed counts
[3, 1, 2]

A, A, B, C, A, C

CONJUGACY!
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$
\[
\phi_t \sim \text{Dirichlet}(\beta)
\]

For each document $d$
\[
\theta \sim \text{Dirichlet}(\alpha)
\]

For each word position in $d$
\[
\text{topic } z \sim \text{Multinomial}(\theta) \quad \text{word } w \sim \text{Multinomial}(\phi_z)
\]

Inference goals: $p(z \mid w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$
(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

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

For each document $d$

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

For each word position in $d$

topic $z \sim \text{Multinomial}(\theta)$

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

Inference goals: $p(z \mid w, \alpha, \beta)$, $\arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w | \alpha, \beta)$:

For each topic $t$

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

For each document $d$

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

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z | w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta | w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:
For each topic $t$

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

For each document $d$

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

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$
(reminder on top)
A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$
$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$
$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$
$$\text{topic } z \sim \text{Multinomial}(\theta)$$
$$\text{word } w \sim \text{Multinomial}(\phi_z)$$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\text{argmax}_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

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

For each document $d$

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

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
A generative model for $p(\phi, \theta, z, w | \alpha, \beta)$:
For each topic $t$
\[ \phi_t \sim \text{Dirichlet}(\beta) \]
For each document $d$
\[ \theta \sim \text{Dirichlet}(\alpha) \]
For each word position in $d$
\[ \text{topic } z \sim \text{Multinomial}(\theta) \]
\[ \text{word } w \sim \text{Multinomial}(\phi_z) \]
Inference goals: $p(z | w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta | w, \alpha, \beta)$ (reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for \( p(\phi, \theta, z, w \mid \alpha, \beta) \):

For each topic \( t \)

\[ \phi_t \sim \text{Dirichlet}(\beta) \]

For each document \( d \)

\[ \theta \sim \text{Dirichlet}(\alpha) \]

For each word position in \( d \)

\[ \text{topic } z \sim \text{Multinomial}(\theta) \]
\[ \text{word } w \sim \text{Multinomial}(\phi_z) \]

Inference goals: \( p(z \mid w, \alpha, \beta) \), \( \arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta) \)

(reminder on top)
When LDA Alone is not Enough

- LDA is unsupervised
- Often domain experts have knowledge in addition to data, want better topics $\phi$
When LDA Alone is not Enough

- LDA is unsupervised
- Often domain experts have knowledge in addition to data, want better topics $\phi$
- There has been many specialized LDA variants
- This talk: how to do (general) “knowledge + LDA” with 3 models:
  1. topic-in-set
  2. Dirichlet forest
  3. Fold.all
Model 1: Topic-in-Set
Example Application: Statistical Software Debugging

[Andrzejewski et al., ECML 2007], [Zheng et al., ICML 2006]

- Insert predicates into a software:

```c
int x = my_func();
if (x > 5) {
    branch_42_true++
    ...
}
else {
    branch_42_false++
    ...
}
```
Example Application: Statistical Software Debugging

[Andrzejewski et al., ECML 2007], [Zheng et al., ICML 2006]

- Insert predicates into a software:

```java
int x = my_func()
if (x > 5) {
    branch_42_true++
    ...
} else {
    branch_42_false++
    ...
}
```

- predicates $\rightarrow$ words $w$
- a software run $\rightarrow$ doc $d$
- we know which runs crashed and which didn’t (extra knowledge)
- the hope: run LDA on crashed runs, some topics $\phi$ will correspond to “buggy behaviors”
Hope Crashed

Normal software usage topics dominate. No “bug” topic.
Hope Crashed

Normal software usage topics dominate. No “bug” topic.
Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.
Hope Crashed

Normal software usage topics dominate. No “bug” topic.

Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.

- Synthetic success (left) and crashed (right) runs
Hope Crashed

Normal software usage topics dominate. No “bug” topic.

Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.

- Synthetic success (left) and crashed (right) runs

- LDA topics on crashed runs
Model success and crashed runs jointly with $T$ topics:
- fix $t < T$
- for all words in success runs, $z \in \{1 \ldots t\}$ (restricted)
- for all words in crashed runs, $z \in \{1 \ldots T\}$

New hope: $\phi_1 \ldots \phi_t$ usage topics, $\phi_{t+1} \ldots \phi_T$ bug topics
New Hope Succeeds

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.
- Synthetic success (left) and crashed (right) runs
- LDA topics on crashed runs
- ΔLDA topics on success and crashed runs
New Hope Succeeds

- exif
- grep
- gzip

Zhu (Wisconsin)
Generalize $\Delta$LDA to Topic-in-Set

[Andrzejewski & Zhu, NAACL’09 WS]

- The domain knowledge:
  - For each word position $i$ in corpus, we are given a set $C_i \subset \{1 \ldots T\}$, such that $z_i \in C_i$. 

$\text{Zhu (Wisconsin)}$
Generalize $\Delta$LDA to Topic-in-Set

[Andrzejewski & Zhu, NAACL’09 WS]

- The domain knowledge:
  - For each word position $i$ in corpus, we are given a set $C_i \subset \{1 \ldots T\}$, such that $z_i \in C_i$.

- Very easy to implement in collapsed Gibbs sampling:

  $$P(z_i = v | z_{-i}, w, \alpha, \beta) \propto$$

  $$\left( \frac{n_{-i,v}^{(d)} + \alpha}{\sum_u T (n_{-i,u}^{(d)} + \alpha)} \right) \left( \frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'} W (\beta + n_{-i,v}^{(w_i)})} \right) \delta(v \in C_i)$$

  - $n_{-i,v}^{(d)}$ is the number of times topic $v$ is used in document $d$
  - $n_{-i,v}^{(w_i)}$ is the number of times word $w_i$ is generated by topic $v$
  - both excluding position $i$

- Easy to relax the hard constraints
Model 2: Dirichlet Forest
Dirichlet Forest Enables Interactive Topic Modeling

[Andrzejewski et al. ICML 2009]

LDA Topics on Wish Corpus:

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|-----------------------------------------------------|
| 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       |
Dirichlet Forest Enables Interactive Topic Modeling
[Andrzejewski et al. ICML 2009]
LDA Topics on Wish Corpus:

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|---------------------------------------------------------|
| 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           |
**Model 2: Dirichlet Forest**

isolate(50 stopwords from 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(50 stopwords from 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             |
| 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

<table>
<thead>
<tr>
<th>Isolate</th>
<th>i to wish my for and a be that the in</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>god bless jesus loved know everyone love who loves</td>
</tr>
<tr>
<td>13</td>
<td>peace world earth win lottery around save</td>
</tr>
<tr>
<td>14</td>
<td>com call if 4 2 www u visit 1 3 email yahoo</td>
</tr>
</tbody>
</table>
**Model 2: Dirichlet Forest**

\[ \text{topic dice } \phi \sim \text{Dir}(\beta), \text{ doc dice } \theta \sim \text{Dir}(\alpha), \text{ topic } z \]

\[ \text{split}([\text{cancer free cure well}],[\text{go school into college}]) \]

<table>
<thead>
<tr>
<th>Split</th>
<th>Love forever happy together marry fall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Health happiness family good friends</td>
</tr>
<tr>
<td>1</td>
<td>Life happy best live love long time</td>
</tr>
<tr>
<td>2</td>
<td>As not do so what like much don was</td>
</tr>
<tr>
<td>3</td>
<td>Out make money house up work grow able</td>
</tr>
<tr>
<td>4</td>
<td>People no stop less day every each take</td>
</tr>
<tr>
<td>5</td>
<td>Home safe end troops iraq bring war husband</td>
</tr>
<tr>
<td>6</td>
<td>Love peace happiness true everyone joy</td>
</tr>
<tr>
<td>7</td>
<td>Happy healthy family baby safe prosperous</td>
</tr>
<tr>
<td>8</td>
<td>Better president hope paul ron than person</td>
</tr>
<tr>
<td>9</td>
<td>Lose meet man hope boyfriend weight finally</td>
</tr>
<tr>
<td>10</td>
<td>Isolate and to for a the year in new all my 2008</td>
</tr>
<tr>
<td>12</td>
<td>God bless jesus loved everyone know loves</td>
</tr>
<tr>
<td>13</td>
<td>Peace world earth win lottery around save</td>
</tr>
<tr>
<td>14</td>
<td>Com call if 4 www 2 u visit 1 email yahoo 3</td>
</tr>
</tbody>
</table>

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

| Topic     | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-----------|--------------------------------------------------------|
| 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        |
From LDA to Dirichlet Forest

For each topic $t$

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

For each doc $d$

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

For each word $w$

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

$w \sim \text{Multinomial}(\phi_z)$
Richer Knowledge Enabled by Dirichlet Forest

Two pairwise relational primitives:

- **Must-Link**($u, v$)
  - e.g., “school” and “college” should be in the same topic
  - encoded as $\phi_t(u) \approx \phi_t(v)$ for $t = 1 \ldots T$
  - can be both large or both small in a given topic
Richer Knowledge Enabled by Dirichlet Forest

Two pairwise relational primitives:

- **Must-Link** \((u, v)\)
  - e.g., “school” and “college” should be in the same topic
  - encoded as \(\phi_t(u) \approx \phi_t(v)\) for \(t = 1 \ldots T\)
  - can be both large or both small in a given topic

- **Cannot-Link** \((u, v)\)
  - e.g., “college” and “cure” should not be in the same topic
  - no topic has \(\phi_t(u), \phi_t(v)\) both high
Richer Knowledge Enabled by Dirichlet Forest

Two pairwise relational primitives:

- **Must-Link** \( (u, v) \)
  - e.g., “school” and “college” should be in the same topic
  - encoded as \( \phi_t(u) \approx \phi_t(v) \) for \( t = 1 \ldots T \)
  - can be both large or both small in a given topic

- **Cannot-Link** \( (u, v) \)
  - e.g., “college” and “cure” should not be in the same topic
  - no topic has \( \phi_t(u), \phi_t(v) \) both high

Complex knowledge:

- split = CL between groups, ML within group
- merge = ML between groups
- isolate = CL to all high prob words in all topics
The Dirichlet Distribution is Insufficient for Must-Link

- Must-Link(school, college) with vocabulary \{school, college, lottery\}
- desired topics $\phi$ distribution on 3-simplex
The Dirichlet Distribution is Insufficient for Must-Link

- Must-Link(school, college) with vocabulary \{school, college, lottery\}
- desired topics $\phi$ distribution on 3-simplex

cannot be represented by a Dirichlet

Dir(5,5,0.1)  Dir(50,50,1)  Dir(500,500,100)
The Dirichlet Tree

[Denis III 1991], [Minka 1999]

- Dirichlet variance $V(i) = \frac{E(i)(1-E(i))}{1+\sum_j \beta_j}$ tied to mean $E(i) = \frac{\beta_i}{\sum_j \beta_j}$
- More flexible: $\phi \sim \text{DirichletTree}(\gamma)$
  - Draw multinomial at each internal node
  - Multiply probabilities from leaf to root

$(\beta = 1, \eta = 50)$
The Dirichlet Tree

[Dennis III 1991], [Minka 1999]

- Dirichlet variance \( V(i) = \frac{E(i)(1-E(i))}{1+\sum_j \beta_j} \) tied to mean \( E(i) = \frac{\beta_i}{\sum_j \beta_j} \)

- More flexible: \( \phi \sim \text{DirichletTree}(\gamma) \)
  - Draw multinomial at each internal node
  - Multiply probabilities from leaf to root

\[
\gamma \begin{array}{c}
\eta \beta \\
A \\
\eta \beta \\
B \\
\beta \\
C \\
2\beta \\
\end{array}
\]

\((\beta = 1, \eta = 50)\)
The Dirichlet Tree

[Dennis III 1991], [Minka 1999]

- Dirichlet variance $V(i) = \frac{E(i)(1-E(i))}{1+\sum_j \beta_j}$ tied to mean $E(i) = \frac{\beta_i}{\sum_j \beta_j}$

- More flexible: $\phi \sim \text{DirichletTree}(\gamma)$
  - Draw multinomial at each internal node
  - Multiply probabilities from leaf to root

$\gamma \left\{ \begin{array}{c}
2\beta \\
\eta\beta \\
A \\
B \\
C \\
\end{array} \right.$

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

$\phi = \left\{ \begin{array}{c}
0.53 \\
0.38 \\
0.09 \\
\end{array} \right.$

$0.91 \\
0.09 \\
0.58 \\
0.42$
The Dirichlet Tree

\[ p(\phi|\gamma) = \left( \prod_{k} \phi(k)^{\gamma(k)} - 1 \right) \left( \prod_{s} \frac{\Gamma \left( \sum_{k} C(s) \gamma(k) \right)}{\prod_{k} \Gamma \left( \gamma(k) \right)} \right) \left( \frac{L(s)}{\sum_{k} \phi(k)} \right)^{\Delta(s)} \]

- \( C(s) = \text{children of } s, \ L(k) = \text{leaves of subtree } k \)
- \( \Delta(s) = \text{InDegree}(s) - \text{OutDegree}(s) \) being zero recovers Dirichlet
The Dirichlet Tree

\[ p(\phi|\gamma) = \left( \prod_k \phi_k^{(k)} \gamma_k^{(k)} - 1 \right) \cdot \left( \prod_s \frac{\Gamma \left( \sum_k C(s) \gamma_k^{(k)} \right)}{\prod_k C(s) \Gamma \left( \gamma_k^{(k)} \right)} \cdot \frac{L(s)}{\sum_k \phi_k^{(k)}} \cdot \Delta(s) \right) \]

- \( C(s) \) = children of \( s \), \( L(k) \) = leaves of subtree \( k \)
- \( \Delta(s) = \text{InDegree}(s) - \text{OutDegree}(s) \) being zero recovers Dirichlet tree is conjugate to multinomial

\[ p(w|\gamma) = \prod_s \left( \frac{\Gamma \left( \sum_k C(s) \gamma_k^{(k)} \right)}{\Gamma \left( \sum_k C(s) \left( \gamma_k^{(k)} + n_k^{(k)} \right) \right)} \cdot \frac{C(s)}{\prod_k \Gamma \left( \gamma_k^{(k)} + n_k^{(k)} \right)} \right) \]
Encode Must-Links with a Dirichlet Tree

[Andrzejewski et al. ICML 2009]

1. Compute Must-Link transitive closures
2. Each transitive closures is a subtree with edges $\eta \beta$
3. The root connects to
   - each transitive closure with edge $|\text{closure}| \beta$
   - each individual word not in any closure with edge $\beta$

$(\beta = 1, \eta = 50)$
The Dirichlet Distribution is Insufficient for Cannot-Link

- Cannot-Link(school, cancer) with vocabulary \{school, cancer, cure\}
- desired topics \(\phi\) distribution on 3-simplex
- cannot be represented by a Dirichlet \((\beta < 1 \text{ not robust})\)
- cannot be represented by a Dirichlet Tree
- requires mixture of Dirichlet Trees (Dirichlet Forest)
Dirichlet Forest Example

- A toy example:
  - Vocabulary: \{A, B, C, D, E, F, G\}
  - Must-Link(A, B)
  - Cannot-Link(A, D), Cannot-Link(C, D), Cannot-Link(E, F)

- Cannot-Link graph on Must-Link closures and other words
Picking a Dirichlet Tree from the Dirichlet Forest

Identify Cannot-Link connected components (they are independent)
Picking a Dirichlet Tree from the Dirichlet Forest

Flip edges: “Can”-Links within components

Model 2: Dirichlet Forest
Picking a Dirichlet Tree from the Dirichlet Forest

“Can”-maximal-cliques (no more words can have high prob together)

Model 2: Dirichlet Forest

\[ \text{topic dice } \phi \sim \text{Dir}(\beta), \text{ doc dice } \theta \sim \text{Dir}(\alpha), \text{ topic } z \]
Picking a Dirichlet Tree from the Dirichlet Forest

Pick a Can-clique (subtree) $q^{(1)}$ for the first connected component

$\eta\ (1) = ?$

Model 2: Dirichlet Forest
Picking a Dirichlet Tree from the Dirichlet Forest

We picked the first Can-clique $q^{(1)} = 1$: give $ABC$ most prob mass
Picking a Dirichlet Tree from the Dirichlet Forest

$D$ has little mass, will not occur with any of $A, B, C$: satisfying Cannot-Links
Picking a Dirichlet Tree from the Dirichlet Forest

Mass distribution between \((AB)\) and \(C\) flexible; \(A, B\) under Must-Link subtree

\[
q^{(1)} = 1
\]

\[
\eta \eta \eta
\]

\[
A \quad B \quad C \quad D \quad E \quad F \quad G
\]

\[
\text{Zhu (Wisconsin)} \quad \text{Knowledge} \rightarrow \text{LDA}
\]
Picking a Dirichlet Tree from the Dirichlet Forest

Pick a Can-clique $q^{(2)}$ for the second connected component

\[ q^{(1)} = 1 \]

\[ q^{(2)} = ? \]
Picking a Dirichlet Tree from the Dirichlet Forest

We picked the 2nd Can-clique $q^{(2)} = 2$
Picking a Dirichlet Tree from the Dirichlet Forest

Here is the chosen subtree: $F$ will get prob mass, not $E$
Inference with Dirichlet Forest

- In theory the number of Can-cliques is exponential, in practice the largest we’ve seen is 3
- Collapsed Gibbs sampling over topics $z_1 \ldots z_N$ and subtree indices $q_j^{(r)}$ for topic $j = 1 \ldots T$ and Cannot-Link connected components $r$

\[
p(z_i = v|z_{-i}, q_{1:T}, w) \propto (n_{-i,v}^{(d)} + \alpha) \prod_s I_v(\uparrow i) \frac{\gamma_v^{(C_v(s \downarrow i))} + n_{-i,v}^{(C_v(s \downarrow i))}}{\sum_k C_v(s) \left( \gamma_v^{(k)} + n_{-i,v}^{(k)} \right)}
\]

\[
p(q_j^{(r)} = q'|z, q_{-j}, q_j^{(-r)}, w) \propto \left( \sum_k \beta_k \right) \prod_s \left( \frac{\Gamma \left( \sum_k C_j(s) \gamma_j^{(k)} \right)}{\Gamma \left( \sum_k C_j(s) \left( \gamma_j^{(k)} + n_j^{(k)} \right) \right)} \prod_k \frac{C_j(s)}{\Gamma \left( \gamma_j^{(k)} + n_j^{(k)} \right)} \right)
\]
Model 3: Fold.all

Zhu (Wisconsin)
Logic can Encode Very General Knowledge

- Topic-in-set and Dirichlet Forest (ML, CL) not general enough
Logic can Encode Very General Knowledge

- Topic-in-set and Dirichlet Forest (ML, CL) not general enough
- Fold.all = First-Order Logic latent Dirichlet ALLocation
  - easy for domain experts to write rules
  - can describe very general domain knowledge
    - can encode many existing LDA variants
  - efficient inference
Logic can Encode Very General Knowledge

- Topic-in-set and Dirichlet Forest (ML, CL) not general enough
- Fold.all = First-Order Logic latent Dirichlet ALLocation
  - easy for domain experts to write rules
  - can describe very general domain knowledge
    - can encode many existing LDA variants
  - efficient inference
- A hybrid Markov Logic Network (MLN) [Wang & Domingos 2008] [Richardson & Domingos 2006], but with fast stochastic optimization
Domain Knowledge in Logic

- Key hidden predicate: \( Z(i, t) \) TRUE if topic \( z_i = t \)
- Observed predicates (anything goes):
  - \( W(i, v) \) TRUE if word \( w_i = v \)
  - \( D(i, j) \) TRUE if word position \( i \) is in document \( j \)
  - \( \text{HasLabel}(j, l) \) TRUE if document \( j \) has label \( l \)
  - \( S(i, k) \) TRUE if word position \( i \) is in document \( k \)
  - \( \ldots \)
Domain Knowledge in Logic

- Key hidden predicate: \( Z(i, t) \) TRUE if topic \( z_i = t \)
- Observed predicates (anything goes):
  - \( W(i, v) \) TRUE if word \( w_i = v \)
  - \( D(i, j) \) TRUE if word position \( i \) is in document \( j \)
  - \( \text{HasLabel}(j, l) \) TRUE if document \( j \) has label \( l \)
  - \( S(i, k) \) TRUE if word position \( i \) is in document \( k \)
  - \( \ldots \)
- Domain knowledge-base \((\lambda_1, \psi_1) \ldots (\lambda_L, \psi_L)\)
  - rules \( \psi \)
  - positive weights \( \lambda \) indicate strength of rule

Example:

\[
\lambda_1 = 1, \psi_1 = "\forall i : W(i, \text{embryo}) \Rightarrow Z(i, 3)"
\]
\[
\lambda_2 = 100, \psi_2 = "\forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))"
\]
Propositionalization

- Let $G(\psi)$ be all ground formulas of $\psi$.
  - $\psi = \forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))$
  - One ground formula $g \in G(\psi)$ is
    $W(123, \text{movie}) \land W(456, \text{film}) \Rightarrow \neg(Z(123, 9) \land Z(456, 9))$
Propositionalization

- Let $G(\psi)$ be all ground formulas of $\psi$.
  - $\psi = "\forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))"$
  - One ground formula $g \in G(\psi)$ is $W(123, \text{movie}) \land W(456, \text{film}) \Rightarrow \neg(Z(123, 9) \land Z(456, 9))$
- $|G(\psi)|$ combinatorial.
- Let
  $$\mathbb{1}_g(z) = \begin{cases} 
1, & \text{if } g \text{ is TRUE under } z \\
0, & \text{otherwise.}
\end{cases}$$
Fold.all = LDA + MLN
[Andrzejewski et al. IJCAI 2011]

\[
p(z, \phi, \theta | w, \alpha, \beta) \\
\propto \left( \prod_t p(\phi_t | \beta) \right) \left( \prod_j p(\theta_j | \alpha) \right) \left( \prod_i \phi_{zi}(w_i) \theta_{di}(z_i) \right)
\]
Fold.all = LDA + MLN

[Andrzejewski et al. IJCAI 2011]
Fold.all Inference

MAP estimate, non-convex objective

\[ Q(z, \phi, \theta) \equiv \sum_t^{T} \log p(\phi_t|\beta) + \sum_j^{D} \log p(\theta_j|\alpha) \]

\[ + \sum_i^N \log \phi_{z_i}(w_i)\theta_{d_i}(z_i) + \sum_l^{L} \sum_{g \in G(\psi_l)} \lambda_l 1_g(z) \]

Alternating optimization. Repeat:

- fixing \( z \), let \((\phi^*, \theta^*) \leftarrow \text{argmax}_{\phi, \theta} Q(z, \phi, \theta) \) (easy)
- fixing \( \phi, \theta \), let \( z^* \leftarrow \text{argmax}_z Q(z, \phi, \theta) \) (integer)
Optimizing $z$ Step 1: Relax $1_g(z)$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$

$$\neg Z(i, 1) \land Z(j, 2)$$
Optimizing $z$ Step 1: Relax $1_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$

2. Remove negations $(\neg g)_+$

$\neg Z(i, 1) \land Z(j, 2)$

$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$
Optimizing $z$ Step 1: Relax $1_g(z)$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$

2. Remove negations $(\neg g)_+$

3. Numeric $z_{it} \in \{0, 1\}$

$$\neg Z(i, 1) \land Z(j, 2)$$

$$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$$

$$(z_{i2} + z_{i3})z_{j2}$$
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$
   
   $\neg Z(i, 1) \land Z(j, 2)$

2. Remove negations $(\neg g)_+$
   
   $(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$

3. Numeric $z_{it} \in \{0, 1\}$
   
   $(z_{i2} + z_{i3})z_{j2}$

4. Polynomial $\mathbb{1}_g(z)$
   
   $1 - (z_{i2} + z_{i3})z_{j2}$
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$
\[
\neg Z(i, 1) \land Z(j, 2)
\]

2. Remove negations $(\neg g)_+$
\[
(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)
\]

3. Numeric $z_{it} \in \{0, 1\}$
\[
(z_{i2} + z_{i3})z_{j2}
\]

4. Polynomial $\mathbb{1}_g(z)$
\[
1 - (z_{i2} + z_{i3})z_{j2}
\]

5. Relax discrete $z_{it}$
\[
z_{it} \in \{0, 1\} \rightarrow z_{it} \in [0, 1]
\]
Optimizing $\mathbf{z}$ Step 1: Relax $\mathbb{1}_g(\mathbf{z})$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$
2. Remove negations $(\neg g)_+$
3. Numeric $z_{it} \in \{0, 1\}$
4. Polynomial $1_g(\mathbf{z})$
5. Relax discrete $z_{it}$

$\neg Z(i, 1) \land Z(j, 2)$

$$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$$

$$(z_{i2} + z_{i3})z_{j2}$$

$$1 - (z_{i2} + z_{i3})z_{j2}$$

$z_{it} \in \{0, 1\} \rightarrow z_{it} \in [0, 1]$
Optimizing $\mathbf{z}$ Step 1: Relax $1_g(\mathbf{z})$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$
   
2. Remove negations $(\neg g)_+$
   
3. Numeric $z_{it} \in \{0, 1\}$

4. Polynomial $1_g(\mathbf{z})$

5. Relax discrete $z_{it}$

\[
1_g(\mathbf{z}) = 1 - \prod_{g_i \neq \emptyset} \left( \sum_{Z(i,t) \in (\neg g)_+} z_{it} \right)
\]
Optimizing \( z \) Step 2: Stochastic Optimization

- \( \sum_{l}^{L} |G(\psi_l)| + NT \) terms in \( Q \) related to \( z \):
  \[
  \max_{z} \sum_{l}^{L} \sum_{g \in G(\psi_l)} \lambda_l \mathbb{1}_g(z) + \sum_{i}^{N} \sum_{t}^{T} z_{it} \log \phi_t(w_i) \theta_d_i(t)
  \]
  s.t. \( z_{it} \geq 0, \sum_{t}^{T} z_{it} = 1. \)

- Entropic Mirror Descent [Beck & Teboulle, 2003]. Repeat:
  - select a term \( f \) at random
  - descent with decreasing step size \( \eta \)

\[
 z_{it} \leftarrow \frac{z_{it} \exp (\eta \nabla z_{it} f)}{\sum_{t'} z_{it'} \exp (\eta \nabla z_{it'} f)}
\]
Example: Movie Reviews

$$\forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))$$
Generalization and Scalability

Cross-validation

- Training: do Fold.all MAP inference to estimate $\hat{\phi}$
- Testing: use trainset $\hat{\phi}$ to infer testset $\hat{z}$ (no logic rules)
- Evaluation: testset objective $Q$
- “-”: runs more than 24 hours

| Data   | Fold.all | LDA  | Alchemy | $\sum_l |G(\psi_l)|$ |
|--------|----------|------|---------|-----------------|
| Synth  | 9.86     | -2.18| -1.73   | $10^5$          |
| Comp   | 2.40     | 1.19 | -       | $10^4$          |
| Con    | 2.51     | 1.09 | -       | $10^3$          |
| Pol    | 5.67     | 5.67 | -       | $10^9$          |
| HDG    | 10.66    | 3.59 | -       | $10^8$          |
Summary

- “Knowledge + data” for latent topic models
- Increasingly more general models
  - Topic-in-set
  - Dirichlet forest (Must & Cannot links)
  - Fold.all (logic)
- Easy for users
- Scalable inference