

# Non-parametric Methods for Unsupervised Semantic Modelling

**Wray Buntine**

Monash University

Thanks Lan Du of Macquarie University and Swapnil Mishra and Kar Wai Lim of ANU for many great slides

Thursday 4<sup>th</sup> December, 2014

# Outline



- 1 Background
  - Motivation
    - Probability Vector Networks
    - Dirichlet distributions
    - Old School Probabilistic Reasoning
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

## Information Overload



langwitches' photostream @ Flickr

We need new tools to help us: **organize**, **search**, **summarise** and **understand** information. The field of **Information Access** serves this purpose.

# Information Warfare

Definition: "the use and management of information in pursuit of a competitive advantage over an opponent."

- Email spam, link spam, *etc.*
  - Whole websites are fabricated with fake content to trick search engines.
  - Spammers using social networks [to personalise attacks](#) (Nov. 2011).
- BBC reports trust in information on the web is being damaged "[by the huge numbers of people paid by companies to post comments](#)" (Dec. 2011).

**It's an information war out there** on the internet between consumers (*i.e.*, you), companies, not-for-profits, voters, parties, employees, bureaucrats, academics, ....

# Outline



- 1 Background
  - Motivation
  - Probability Vector Networks
  - Dirichlet distributions
  - Old School Probabilistic Reasoning
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

# Probability Vectors

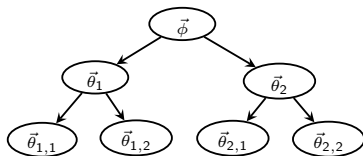
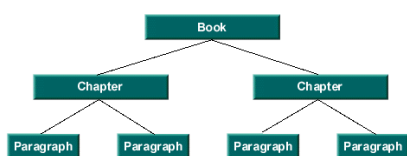
We often have **probability vectors** for:

- the next word given  $(n - 1)$  previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,
- hashtag in a tweet given the author.

We need to work with **distributions** over probability vectors for:

- inheritance and sharing of information;
- networks of probability vectors;
- inference and learning.

# Sharing/Inheritance with a Probability Hierarchy



We might model a set of vocabularies/documents hierarchically:

$$\begin{aligned}\vec{\theta}_1 &\sim \text{Dirichlet}(\alpha_0 \vec{\phi}) \\ \vec{\theta}_{1,2} &\sim \text{Dirichlet}(\alpha_1 \vec{\theta}_1)\end{aligned}$$

Statistical estimation with these generally difficult:

$$\dots \frac{1}{\text{Beta}(\alpha_1 \vec{\theta}_1)} \prod_k \theta_{1,2,k}^{\alpha_1 \theta_{1,k} - 1} \prod_k \theta_{1,k}^{\alpha_1 \phi_k - 1} \dots$$

**(N.B. fixed point MAP solutions exist for hierarchies)**

# Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.



# Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.
- New fast methods for training deep probability vector networks.

# Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.
- New fast methods for training deep probability vector networks.
- Allows modelling of latent semantics:
  - semantic resources to integrate (WordNet, sentiment dictionaries, *etc.*),
  - inheritance and shared learning across multiple instances,
  - hierarchical modelling,
  - deep latent semantics,
  - integrating semi-structured and networked content,

i.e. Same as deep neural networks!

# Outline



- 1 Background
  - Motivation
  - Probability Vector Networks
  - Dirichlet distributions
  - Old School Probabilistic Reasoning
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

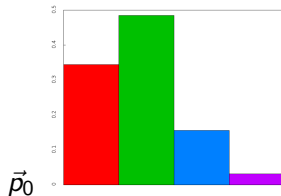
# Dirichlet distributions

The **Dirichlet distribution** is used to sample finite probability vectors.

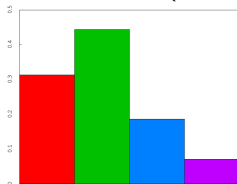
$$\vec{p} \sim \text{Dirichlet}_K(\vec{\alpha})$$

where  $\vec{\alpha}$  is a positive  $K$ -dimensional vector.

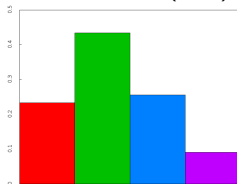
# 4-D Dirichlet samples



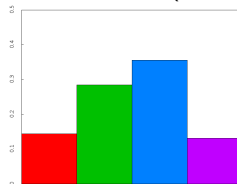
$$\vec{p}_1 \sim \text{Dirichlet}_4(500\vec{p}_0)$$



$$\vec{p}_2 \sim \text{Dirichlet}_4(5\vec{p}_0)$$



$$\vec{p}_3 \sim \text{Dirichlet}_4(0.5\vec{p}_0)$$



## Forms for 3-D Dirichlet

Consider  $\vec{p} = (p_1, p_2, p_3)$  where  $\sum_k p_k = 1$ .

$\vec{p} \sim \text{Dirichlet}_3(\vec{\alpha})$  means that  $p(\vec{p} | \vec{\alpha})$  is

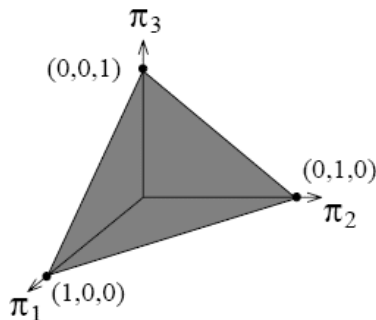
$$\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1},$$

where  $\Gamma(\cdot)$  is the Gamma function.

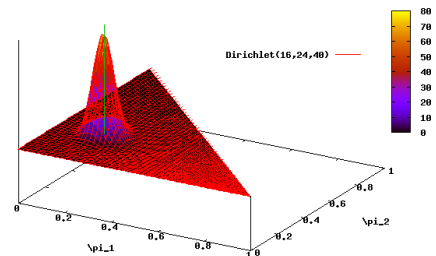
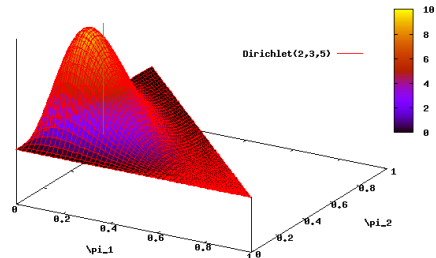
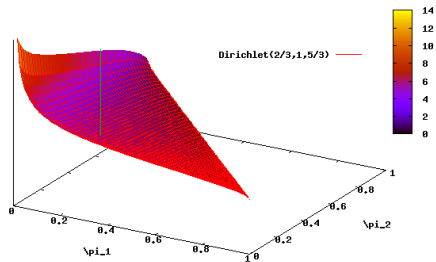
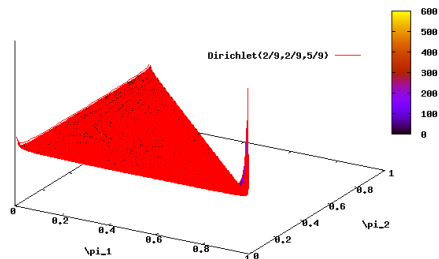
Also have derived parameters

$$\begin{aligned} \alpha_0 &= \alpha_1 + \alpha_2 + \alpha_3 \\ \vec{\mu} = (\mu_1, \mu_2, \mu_3) &= \left( \frac{\alpha_1}{\alpha_0}, \frac{\alpha_2}{\alpha_0}, \frac{\alpha_3}{\alpha_0} \right), \end{aligned}$$

where **mean of  $\vec{p}$  is  $\vec{\mu}$** , and  $\alpha_0$  is a concentration parameter.



# The Dirichlet Distribution



# Dirichlet Details

- $\alpha_0 = \alpha_1 + \alpha_2 + \alpha_3$  is called the **concentration parameter**. Its significance is shown by the **variance**

$$\text{Var}_{\vec{\alpha}} [\vec{p}_i] = \frac{1}{\alpha_0 + 1} \mathbb{E}_{\vec{\alpha}} [\vec{p}_i] (1 - \mathbb{E}_{\vec{\alpha}} [\vec{p}_i]) .$$

- Let  $\vec{n} \sim \text{multinomial}(N, \vec{p})$ , then

$$p(\vec{n} | \vec{p}) = C_{\vec{n}}^N p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

So data sampled from a multinomial using  $\vec{p}$  has **the same functional form as the Dirichlet distribution** on  $\vec{p}$ , *i.e.*, conjugacy.

- Normalising constant for Dirichlet is called a **Beta function**:

$$\text{Beta}_3(\alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}$$

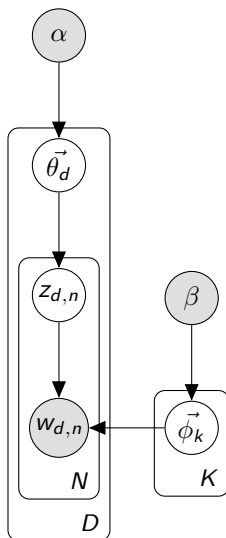


# Outline



- 1 Background
  - Motivation
  - Probability Vector Networks
  - Dirichlet distributions
  - Old School Probabilistic Reasoning
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

# Latent Dirichlet Allocation



## LDA Model

$$\begin{aligned}\forall_d \quad \vec{\theta}_d &\sim \text{Dirichlet}_K(\vec{\alpha}) \\ \forall_k \quad \vec{\phi}_k &\sim \text{Dirichlet}_W(\vec{\beta}) \\ \forall_{d,n} \quad z_{d,n} &\sim \text{Categorical}(\vec{\theta}_d) \\ \forall_{d,n} \quad w_{d,n} &\sim \text{Categorical}(\vec{\phi}_{z_{d,n}})\end{aligned}$$

## Collapsed Posterior for Gibbs Sampling

$$\prod_d \frac{\text{Beta}_K(\vec{n}_{\vec{\theta}_d} + \vec{\alpha})}{\text{Beta}_K(\vec{\alpha})} \prod_k \frac{\text{Beta}_W(\vec{n}_{\vec{\phi}_k} + \vec{\beta})}{\text{Beta}_W(\vec{\beta})}$$

# Learning Algorithms with Dirichlets

Common text-book algorithms/methods in modern machine-learning/statistics rely on Dirichlet distributions combined with:

- trees, tables;
- graphs, networks;
- context free grammars;

Algorithms on these combine Dirichlet normalizers with:

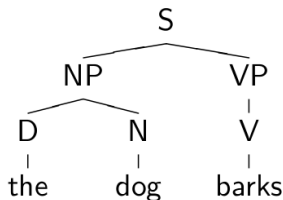
- model search;
- model averaging;
- EM algorithm;
- *etc.*

Arguably, many of these are of poor/mixed quality.

e.g., probabilistic context free grammars, decision trees

# Context Free Grammar

$$\mathcal{R} = \left\{ \begin{array}{lll} S \rightarrow NP \ VP & NP \rightarrow D \ N & VP \rightarrow V \\ D \rightarrow \text{the} & N \rightarrow \text{dog} & V \rightarrow \text{barks} \end{array} \right\}$$



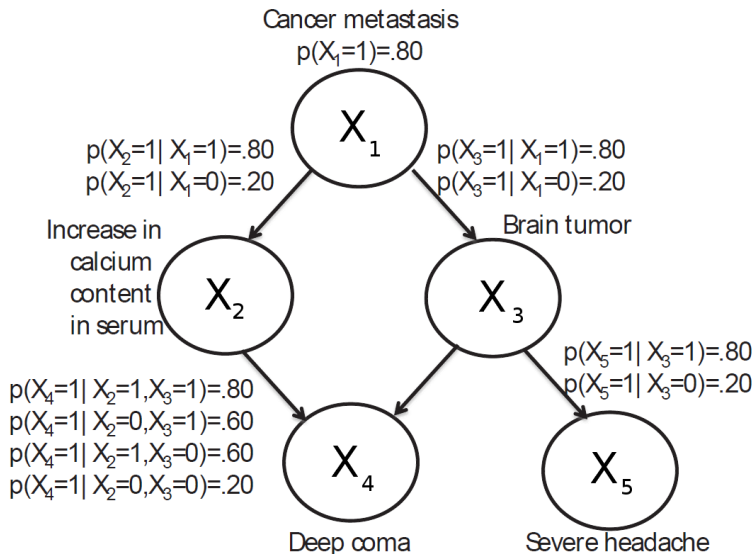
S  
 $\Rightarrow$  NP VP  
 $\Rightarrow$  D N VP  
 $\Rightarrow$  the N VP  
 $\Rightarrow$  the dog VP  
 $\Rightarrow$  the dog V  
 $\Rightarrow$  the dog barks

In a [probabilistic context free grammar](#), probabilities are associated with each rule, and rules apply independently of context.

# Probabilistic Context Free Grammar, cont.




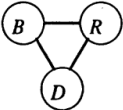
- Doesn't perform well in practice because **statistically, context matters**.  
*Google bought Youtube.*
- Previous words, or higher parts-of-speech do affect probabilities.
- State-of-the-art systems “hack” context by making probabilities dependent on:
  - previous few words in the input stream;
  - previous few parts-of-speech higher in the parse tree;
  - head (“main”) words for nodes higher in the parse tree.
- Alternatively, they introduce specialised parts-of-special to introduce context.
- This requires algorithmic sophistication!

# Bayesian Networks



# Bayesian Model Averaging (BMA) for Bayesian Networks

*Summaries of the posterior distributions of  $N$  for the spina bifida data for all models with posterior probability greater than 0.01.  $\hat{N}$  is a Bayes estimate, minimizing a relative squared error loss function*

Model	Posterior Prob.	$\hat{N}$	2.5%, 97.5%
	0.373	731	(701, 767)
	0.301	756	(714, 811)
	0.281	712	(681, 751)
	0.036	697	(628, 934)
<b>Model Averaging</b>	—	731	(682, 797)

From Madigan and York, 1995

# Bayesian Model Averaging (BMA) for Bayesian Networks, cont

Build a pool of “good” models (*i.e.*, the graphical structures) based on the training data:  $\text{Good-Models}(\mathcal{X})$ .

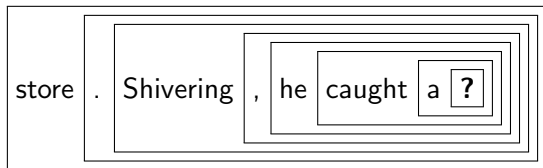
BMA estimate of probability for new data  $\vec{x}$  given training sample  $\mathcal{X}$  is

$$p(\vec{x}|\mathcal{X}) \approx \sum_{M \in \text{Good-Models}(\mathcal{X})} \frac{p(M|\mathcal{X})}{\sum_{M \in \text{Good-Models}(\mathcal{X})} p(M|\mathcal{X})} p(\vec{x}|M, \mathcal{X})$$

Question: how do we build “good” models given there are a combinatoric number?



# N-grams



To model a sequence of words  $p(w_1, w_2, \dots, w_N)$  we can use:

Unigram (1-gram) model:

$$p(w_n | \vec{\theta}_1)$$

Bigram (2-gram) model:

$$p(w_n | w_{n-1}, \vec{\theta}_2)$$

Trigram (3-gram) model:

$$p(w_n | w_{n-1}, w_{n-2}, \vec{\theta}_3)$$

By BMA with data  $\mathcal{W}$ , we use the sequence of  $k$ -gram models  $M_k$ :

$$\sum_{k=1}^K p(w_n | w_{n-1}, \dots, w_{n-k+1}, \mathcal{W}, M_k) p(M_k | \mathcal{W})$$

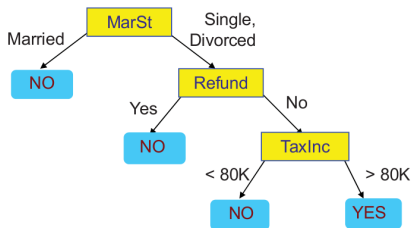
where the first probability is the  $k$ -gram estimate from the data.

For n-grams, the good models are easy.

# Decision Trees

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

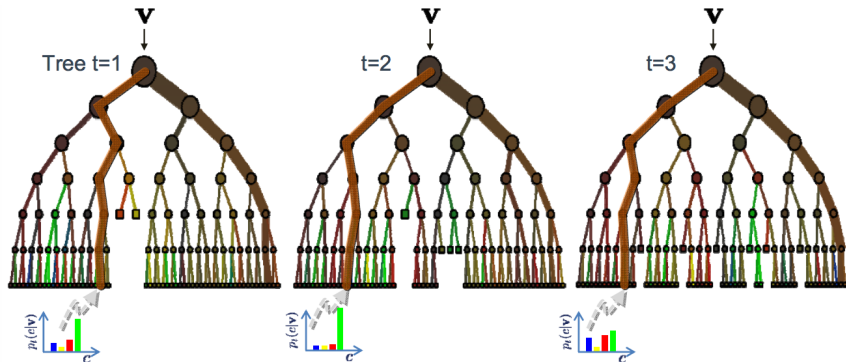
categorical  
categorical  
continuous  
class



There could be more than one tree that fits the same data!

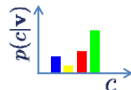
Classification on new data based on tree. Early classical non-parametric method made famous with CART (Breiman, Friedman and Stone) and ID3/C4.5 (Quinlan).

# Breiman's Random Forests or Bagging of Decision Trees



## The ensemble model

Forest output probability 
$$p(c|\mathbf{v}) = \frac{1}{T} \sum_t p_t(c|\mathbf{v})$$



(source unknown!)

# Bayesian Model Averaging and Non-parametrics Storyline

- 1990: My PhD thesis, **BMA** for decision trees.
- 1990: York and Madigan develop BMA for Bayesian networks.
- 1994: Breiman developed **bagging** (or random forests) for trees as a Frequentist response:
  - still one of the top performing classification algorithms
- 1995: Willems, Shtarkov, Tjalkens adapt BMA for n-grams, **context tree weighting (CTW)** for lossless compression.

Bayesian model averaging and Frequentist bagging became dominant paradigms.

# Bayesian Model Averaging and Non-parametrics Storyline, cont.

- 2006: Y.W. Teh develops [hierarchical Pitman-Yor model](#) for n-grams.
- 2009: Gasthaus, Wood, Teh and James develop [Sequence Memoizer](#) for n-grams for lossless compression. Beats CTW.
- 2009: Wood and Teh develop [Statistical Language Model Domain Adaptation](#). Further improves n-gram modelling by allowing adaptation.
- but the algorithm is impractical

Non-parametric Bayesian methods give new life to BMA because they are substantially better priors.

# Motivation

- While somewhat successful, the BMA paradigm based on standard (simple) conjugate priors has **reached a dead-end** for some models consisting of Dirichlets.
- Many, many problems in machine learning are ripe for improvement with better modelling of context.
- But this requires efficient non-parametric modelling.

# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
  - Hierarchical Dirichlet and Pitman-Yor Processes
  - PYPs on Discrete Data
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model
- 6 Conclusion

# Dirichlet Process

The **Dirichlet Process (DP)** has two arguments  $\text{DP}(\alpha, H(\cdot))$  :

- $H(\cdot)$  is another distribution that is the mean of the DP
- when  $H(\cdot) = \vec{\mu}$ , applied to a finite probability vector  $\vec{\mu}$  of dimension  $K$ , the **DP and the Dirichlet are identical**:

$$\text{Dirichlet}_K(\alpha, \vec{\mu}) = \text{DP}(\alpha, \vec{\mu}) .$$

- often the use of a DP is equivalent to the use of a Dirichlet.
- **why use the DP then?**
  - **nested/hierarchical DPs have fairly fast samplers.**

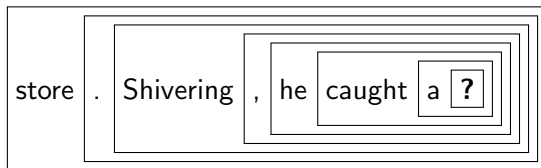


# Pitman-Yor Process

The **Pitman-Yor Process (PYP)**  $PYP(d, \alpha, H(\cdot))$  :

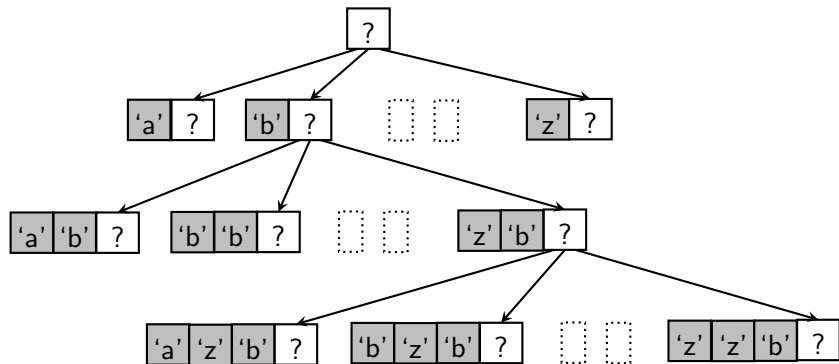
- extends the DP with an extra parameter, and is more suited to **Zipfian** data,
- **discount**  $d$  makes the expected probabilities converge faster,
- the PYP is not Dirichlet, unlike the DP
- same samplers as the DP

# Bayesian Idea: Similar Context Means Similar Word



- Words in a ? should be like words in ?
  - though no plural nouns
- Words in caught a ? should be like words in a ?
  - though a suitable object for “caught”
- Words in he caught a ? be *very like* words in caught a ?
  - “he” shouldn't change things much

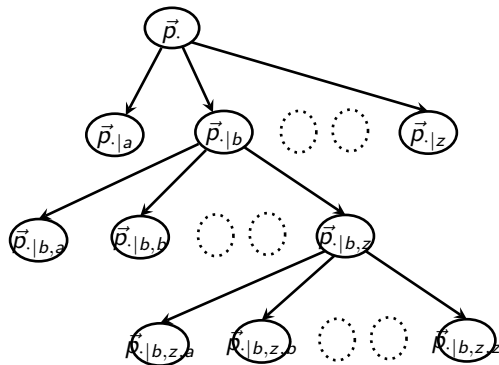
# Bayesian N-grams



Build all three together, making the 1-gram a prior for the 2-grams, and the 2-grams a prior for the 3-grams, etc.

For this, we need to say each probability vector in the hierarchy is a variant of its parent.

# Bayesian N-grams, cont.



$\mathcal{S}$  = symbol set, fixed or possibly countably infinite

$\vec{p}_{\cdot} \sim$  prior on prob. vectors (initial vocabulary)

$\vec{p}_{\cdot|x_1} \sim$  dist. on prob. vectors with mean  $\vec{p}_{\cdot}$   $\forall x_1 \in \mathcal{S}$

$\vec{p}_{\cdot|x_1, x_2} \sim$  dist. on prob. vectors with mean  $\vec{p}_{\cdot|x_1}$   $\forall x_1, x_2 \in \mathcal{S}$

# Historical Context

- 1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*
- 2001-2003: Ishwaran and James develops and “translates” methods usable for machine learning.
- 2006: Teh develops hierarchical n-gram models using HPYs.
- 2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes (HDP), e.g. applied to LDA.
- 2006-2011: Chinese restaurant processes (CRPs) go wild!
- require dynamic memory in implementation,
  - Chinese restaurant franchise,
  - multi-floor Chinese restaurant process,
  - *etc.*

# Historical Context

- 1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*
- 2001-2003: Ishwaran and James develops and “translates” methods usable for machine learning.
- 2006: Teh develops hierarchical n-gram models using HPYs.
- 2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes (HDP), e.g. applied to LDA.
- 2006-2011: Chinese restaurant processes (CRPs) go wild!
- require dynamic memory in implementation,
  - Chinese restaurant franchise,
  - multi-floor Chinese restaurant process,
  - *etc.*
- 2011: Chen, Du, Buntine show Chinese restaurants not needed by introducing **block table indicator samplers**.

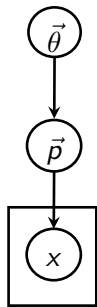
# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
  - Hierarchical Dirichlet and Pitman-Yor Processes
  - PYPs on Discrete Data
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model
- 6 Conclusion

# The Ideal Hierarchical Component?

We want a magic distribution that looks like a multinomial likelihood in  $\vec{\theta}$ .



$$\begin{aligned}\vec{p} &\sim \text{Magic}(\alpha, \vec{\theta}) \\ x_n &\sim \text{Discrete}(\vec{p}) \quad \forall_n\end{aligned}$$

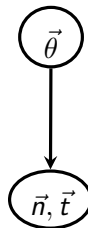


$$\begin{aligned}p(\vec{n} \mid \alpha, \vec{\theta}, N) \\ &= F_{\alpha}(\vec{n}) \prod_k \theta_k^{t_k} \\ \text{where } \sum_k n_k &= N\end{aligned}$$



# The PYP/DP is the Magic

- The PYP/DP plays the role of the magic distribution.
- However, the exponent  $t_k$  for the  $\theta$  now **becomes a latent variable**, so needs to be sampled as well.
- The  $t_k$  are **constrained**:
  - $t_k \leq n_k$
  - $t_k > 0$  iff  $n_k > 0$
- The  $\vec{t}$  act like data for the next level up involving  $\vec{\theta}$ .



$$p(\vec{n}, \vec{t} | d, \alpha, \vec{\theta}, N) = F_{d, \alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

$$\text{where } \sum_k n_k = N$$

# Interpreting the Auxiliary Counts

**Interpretation:**  $t_k$  is how much of the count  $n_k$  that affects the parent probability (i.e.  $\vec{\theta}$ ).

- If  $\vec{t} = \vec{n}$  then the sample  $\vec{n}$  affects  $\vec{\theta}$  100%.
- When  $n_k = 0$  then  $t_k = 0$ , no effect.
- If  $t_k = 1$ , then the sample of  $n_k$  affects  $\vec{\theta}$  minimally.



$$p(\vec{n}, \vec{t} \mid d, \alpha, \vec{\theta}, N)$$

$$= F_{d, \alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

$$\text{where } \sum_k n_k = N$$

# Why We Prefer DPs and PYPs over Dirichlets!

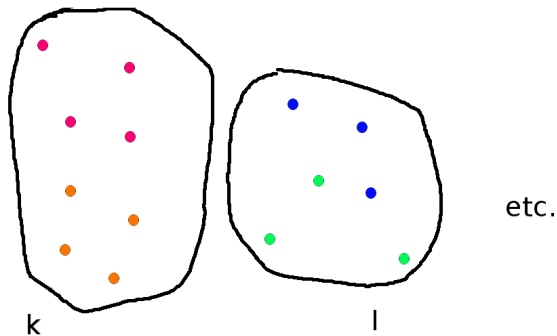
$$p\left(\vec{x}, \vec{t} \mid N, \text{MultPD}, d, \alpha, \vec{\theta}\right) \propto \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^K \mathcal{S}_{t_k, d}^{n_k} \theta_k^{t_k},$$

$$p\left(\vec{x} \mid N, \text{MultDir}, \alpha, \vec{\theta}\right) \propto \frac{1}{(\alpha)_N} \prod_{k=1}^K (\alpha \theta_k)_{n_k}.$$

For the PYP, the  $\theta_k$  just look like multinomial data, but you have to introduce a discrete latent variable  $\vec{t}$ .

For the Dirichlet, the  $\theta_k$  are in a complex gamma function.

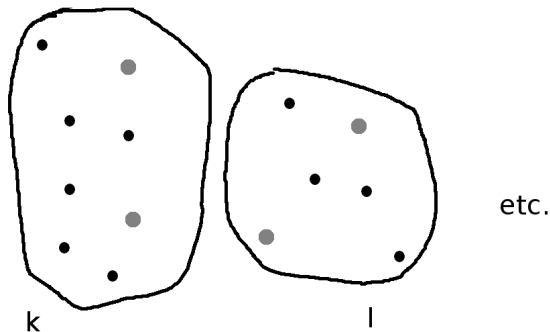
# Typed Data with Species



Within types there are separate *species*, pink and orange for type  $k$ , blue and green for type  $l$ .

Chinese restaurant samplers work in this space.

# Typed Data with New Species, cont.



Within types there are separate *species*, but we only know which data is the first of a new species. For other data, species is unknown.

Block table indicator samplers work in this space.

# Better Sampling Methods for HDP and HPYP

Sampling for hierarchical Dirichlet Processes and Pitman-Yor Processes:

The Old: [hierarchical Chinese Restaurant Processes](#) (CRP) from Teh *et al.* 2006.

The New: [block table indicator sampling](#) from Chen, Du and Buntine 2011.

- requires no dynamic memory
- more rapid mixing so leads to better models
- more easily applied to more complex models
- [demonstrated extensively on different problems!](#)

See <http://topicmodels.org> , “A tutorial on non-parametric methods”

# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
  - Topic Models (joint with Swapnil Mishra)
    - Background
    - Evolution of Models
    - Our Non-parametric Topic Model
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

# Component Models, Generally



*image*  
→



Prince, Elizabeth, son, ...	Queen, title,	school, student, college, education, year, ...
John, Michael, Paul, ...	David, Scott,	and, or, to , from, with, in, out, ...

*text*  
→

13 1995 accompany and(2) andrew at  
boys(2) charles close college day de-  
spite diana dr eton first for gayley  
harry here housemaster looking old on  
on school separation sept stayed the  
their(2) they to william(2) with year

Approximate faces/bag-of-words (RHS) with a linear combination of components (LHS).



# Matrix Approximation View

$$\mathbf{W} \approx \mathbf{L} * \mathbf{\Theta}^T$$

Diagram illustrating the matrix approximation view of Topic Models:

The matrix  $\mathbf{W}$  (Data) is approximated by the product of matrix  $\mathbf{L}$  (Components) and matrix  $\mathbf{\Theta}^T$  (Components).

Matrix  $\mathbf{L}$  (Components) is defined by  $I$  documents (rows) and  $K$  components (columns). Its elements are  $l_{i,j}$ .

Matrix  $\mathbf{\Theta}^T$  (Components) is defined by  $K$  components (rows) and  $J$  words (columns). Its elements are  $\theta_{j,k}$ .

The approximation is shown as:

$$\left\{ \begin{matrix} \text{I documents} \\ \left( \begin{matrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{matrix} \right) \end{matrix} \right\} \approx \left\{ \begin{matrix} \text{K components} \\ \left( \begin{matrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{matrix} \right) \end{matrix} \right\} * \left\{ \begin{matrix} \text{J words} \\ \left( \begin{matrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{matrix} \right) \end{matrix} \right\}$$

Different variants:

Data $\mathbf{W}$	Components $\mathbf{L}$	Error	Models
real valued	unconstrained	least squares	PCA and LSA
non-negative	non-negative	least squares	learning codebooks
non-neg integer	non-negative	cross-entropy	topic modelling
real valued	independent	small	ICA

# Why Topic Models?

- *Topic Models* discover hidden themes in text data to aid understanding.
  - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003).
- Recent research develops higher performance topic models.

# Why Topic Models?

- *Topic Models* discover hidden themes in text data to aid understanding.
  - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003).
- Recent research develops higher performance topic models.
- But why should you care?
- Moreover, why should I care?

# Topic Models: Potential for Semantics

Following sets of topic words created from the New York Times 1985-2005 news collection using `hca` (see Buntine and Mishra, KDD 2014):

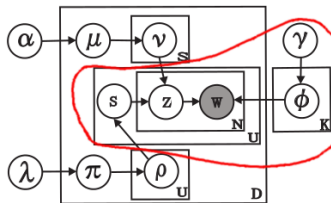
- career, born, grew, degree, earned, graduated, became, studied, graduate
- mother, daughter, son, husband, family, father, parents, married, sister
- artillery, shells, tanks, mortars, gunships, rockets, firing, tank
- clues, investigation, forensic, inquiry, leads, motive, investigator, mystery
- freedom, tyranny, courage, america, deserve, prevail, evil, bless, enemies
- viewers, cbs, abc, cable, broadcasting, channel, nbc, broadcast, fox, cnn
- anthrax, spores, mail, postal, envelope, powder, letters, daschle, mailed

Topic models yield high-fidelity semantic associations!

# Topic Models: Just an Intermediate Goal

Topic models are the leading edge of a new wave of deep latent semantic models applied to real NLP tasks:

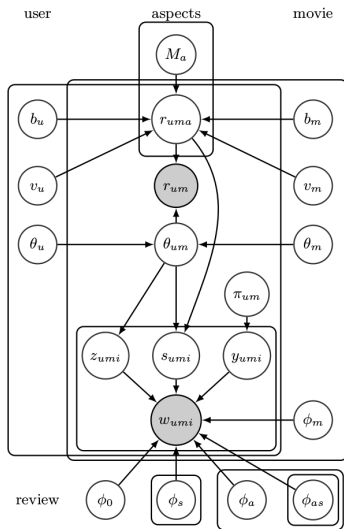
e.g., document segmentation, word sense disambiguation, facet discovery for sentiment analysis, unsupervised POS discovery, social networks, ...



in the middle of this segmentation model is a topic model

i.e., we don't care about topic models *per se*!

# ASIDE: Aspects, Ratings and Sentiments



“Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS),” Diao, Qiu, Wu, Smola, Jiang and Wang, KDD 2014.

State of the art sentiment model.

Typical methods currently lack probabilistic vector hierarchies.

Figure 1: Factorized rating and review model.

# Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test”  
(not his exact words .... but the same idea)

# Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test”  
(not his exact words .... but the same idea)

## Perplexity:

- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use “document completion,” see Wallach, Murray, Salakhutdinov, and Mimno, 2009;
- however **it is not a bonafide evaluation task**



# Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test”  
(not his exact words .... but the same idea)

## Perplexity:

- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use “document completion,” see Wallach, Murray, Salakhutdinov, and Mimno, 2009;
- however **it is not a bonafide evaluation task**

## PMI:

- measure of topic coherence: *“average pointwise mutual information between all pairs of top 10 words in the topic”*
- see Newman, Lau, Grieser, and Baldwin, 2010; Lau, Newman and Baldwin, 2014
- **but at least it corresponds to a semi-realistic evaluation task**

# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
  - Topic Models (joint with Swapnil Mishra)
  - **Background**
  - Evolution of Models
  - Our Non-parametric Topic Model
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

# Previous Work

## EXTENDING

- “Hierarchical Dirichlet Processes,” Teh, Jordan, Beal, Blei 2006.
- “Rethinking LDA: Why priors matter,” Wallach, Mimno, McCallum, 2009.
- “Accounting for burstiness in topic models,” Doyle and Elkan 2009.
- “Topic models with power-law using Pitman-Yor process,” Sato and Nakagawa 2010
- Sampling table configurations for the hierarchical Poisson-Dirichlet process,” Chen, Du and Buntine 2011.
- “Practical collapsed variational Bayes inference for hierarchical Dirichlet process,” Sato, Kurihara, and Nakagawa 2012.
- “Truly nonparametric online variational inference for hierarchical Dirichlet processes,” Bryant and Sudderth 2012.
- “Stochastic Variational Inference,” Hoffman, Blei, Wang and Paisley 2013.

# Text and Burstiness

**Original news article:**

Women may only account for 11% of all Lok-Sabha MPs but they fared better when it came to representation in the Cabinet. Six women were sworn in as senior ministers on Monday, accounting for 25% of the Cabinet. They include Swaraj, Gandhi, Najma, Badal, Uma and Smriti.

**Bag of words:**

11% 25% Badal Cabinet(2) Gandhi Lok-Sabha MPs Monday Najma Six Smriti Swaraj They Uma Women account accounting all and as better but came fared for(2) in(2) include it may ministers of on only representation senior sworn the(2) they to were when women

**NB.** "Cabinet" appears twice! It is **bursty**  
(see Doyle and Elkan, 2009)

## Aside: Burstiness and Information Retrieval

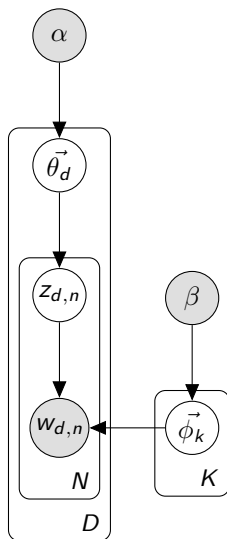
- burstiness and eliteness are concepts in information retrieval used to develop BM25 (*i.e.* dominant TF-IDF version)
- the two-Poisson model and the Pitman-Yor model can be used to justify theory (Sunehag, 2007; Puurula, 2013)
- relationships not yet fully developed

# Outline



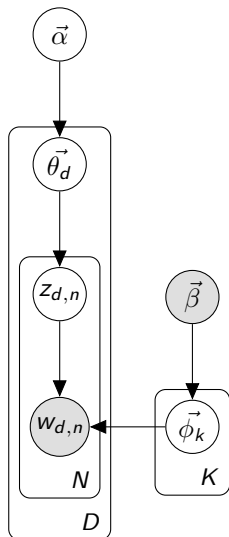
- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
  - Topic Models (joint with Swapnil Mishra)
  - Background
  - Evolution of Models
  - Our Non-parametric Topic Model
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model

# Evolution of Models



**LDA- Scalar**  
original LDA

# Evolution of Models



## LDA- Vector

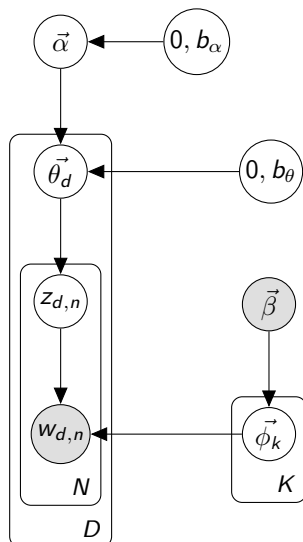
adds asymmetric Dirichlet prior like  
Wallach et al.; is also truncated  
HDP-LDA;

implemented by Mallet since 2008 as  
asymmetric-symmetric LDA

**no one knew!**



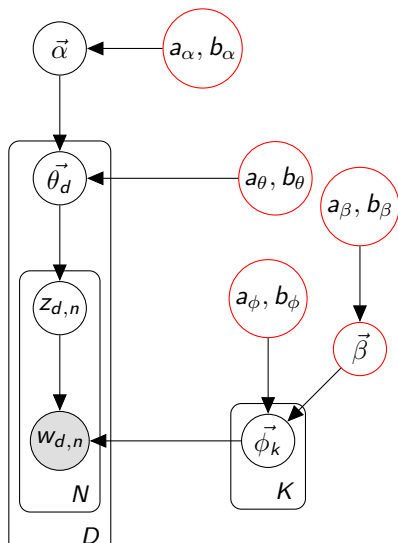
# Evolution of Models



## HDP-LDA

adds proper modelling of topic prior  
like Teh et al.

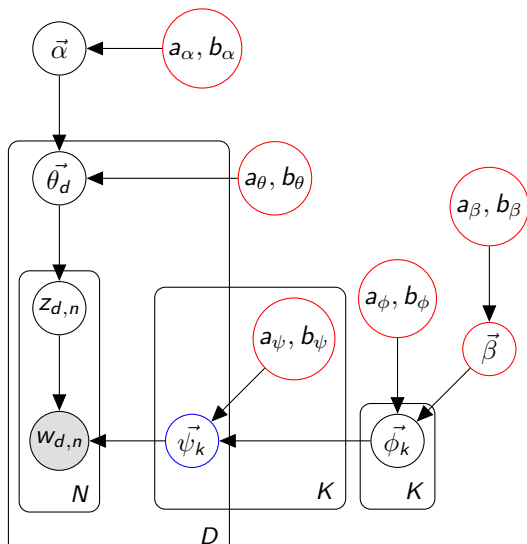
# Evolution of Models



## NP-LDA

adds power law on word distributions  
like Sato et al. and estimation of  
background word distribution

# Evolution of Models



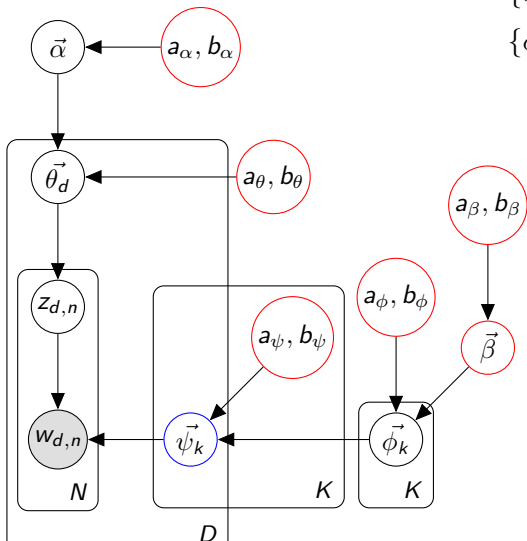
## NP-LDA with Burstiness

add's burstiness like Doyle and Elkan

- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
  - Topic Models (joint with Swapnil Mishra)
  - Background
  - Evolution of Models
  - Our Non-parametric Topic Model
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model



# Our Non-parametric Topic Model



$\{\vec{\theta}_d\} = \text{document} \otimes \text{topic matrix}$

$\{\vec{\phi}_k\} = \text{topic} \otimes \text{word matrix}$

- Full fitting of priors on topic $\otimes$ word and document $\otimes$ topic matrices (red nodes).
- Topic $\otimes$ word vectors  $\vec{\phi}_k$  specialised to the document to yield  $\vec{\psi}_k$ .
- This models burstiness (blue node).

# Design Notes

**Hierarchical priors:** whenever parts of the system seem similar, we give them a common prior and learn the similarity.

# Design Notes

**Hierarchical priors:** whenever parts of the system seem similar, we give them a common prior and learn the similarity.

**Estimating parameters:** whenever parameters cannot be reasonably set, we learn them instead.

# Design Notes

**Hierarchical priors:** whenever parts of the system seem similar, we give them a common prior and learn the similarity.

**Estimating parameters:** whenever parameters cannot be reasonable set, we learn them instead.

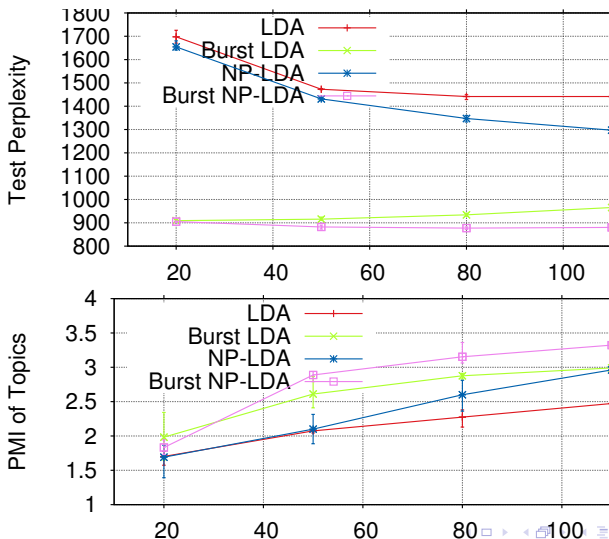
**Burstiness:** we developed a Gibbs sampler that acts as a front end to **any** LDA-style model with Gibbs, e.g.

- implemented as a C function that calls the Gibbs sampler
  - adds smallish memory (20%) and time (20%) overhead
- in all, double memory and time to regular LDA Gibbs sampling
- multi-core implementation

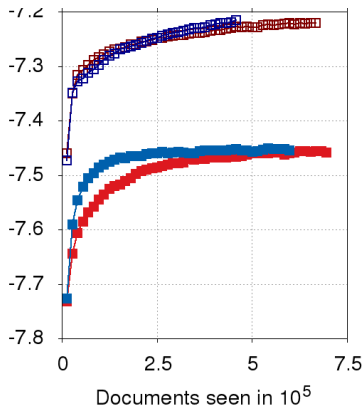
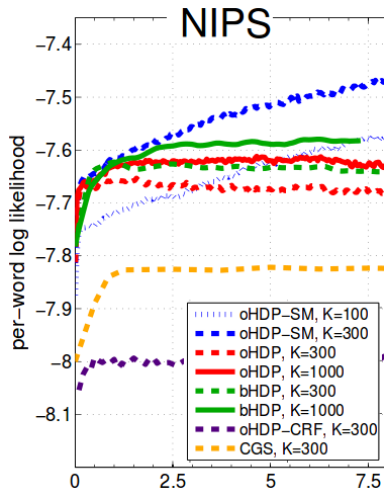


# Performance on Reuters-21578 ModLewis Split

Training on 11314 news articles with vocabulary of 16994.



# Comparison to Bryant+Sudderth (2012) on NIPS data



# Conclusion on Topic Models

Paper given at KDD 2014 with Swapnil Mishra.

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this

# Conclusion on Topic Models

Paper given at KDD 2014 with Swapnil Mishra.

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models

# Conclusion on Topic Models

Paper given at KDD 2014 with Swapnil Mishra.

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
- Grab our topic modelling code from

<https://github.com/wbuntine/topic-models>

<http://mloss.org/software/view/527/>

# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model
- 6 Conclusion

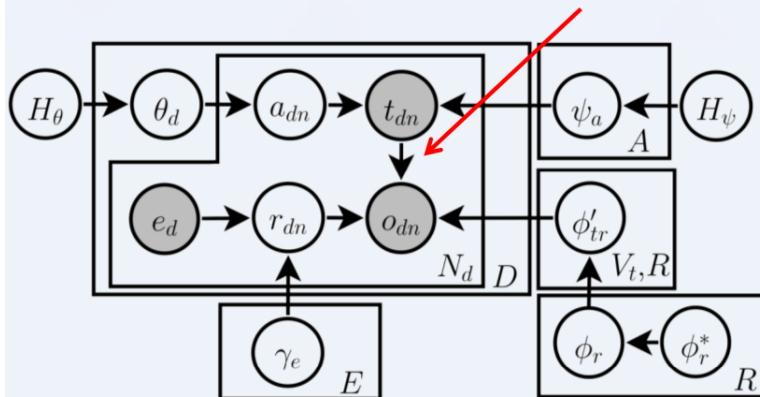
# Aspect-based Opinion Aggregation

- Opinion Aggregation for reviews.
  - A process to collect reviews of products and services to analyze in aggregate.
- Aspect-based.
  - Groups reviews based on “aspects”.
  - Example:
    - Product types
      - Game consoles
      - Mobile phones
    - Product specs
      - Computer specs
      - Flight quality

Aspect	Examples
Game console	PS4, Xbox One, Wii U...
Mobile phone	iPhone, Samsung Note...
Computer spec	CPU, RAM, GPU...
Flight quality	Food, customer service...

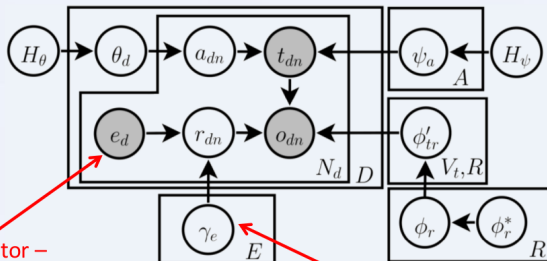
# Explaining the Model

Model target-opinion interaction directly.  
This improves opinion prediction significantly.





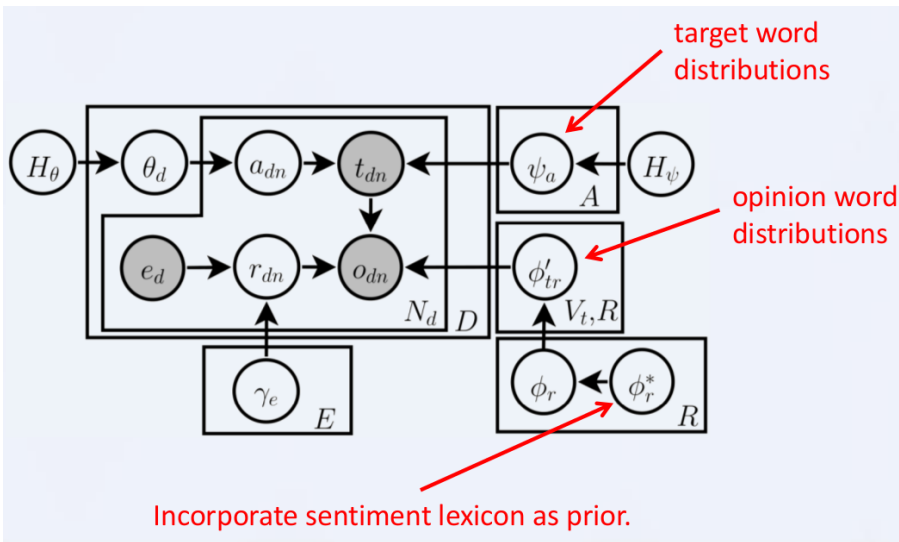
# Explaining the Model, cont.



Emotion indicator –  
determined by seen emoticons  
or strong sentiment words  
(such as 'happy', 'sad').

Positive opinions tend to  
associate with positive emotions.

## Explaining the Model, cont.



# Explaining the Model, cont.

Target ( $t$ )	+/-	Opinions ( $o$ )
phone	-	<b>dead</b> damn stupid bad crazy
	+	<b>mobile smart</b> good great f***ing
battery life	-	terrible poor bad horrible <b>non-existence</b>
	+	good <b>long</b> great <b>7hr ultralong</b>
game	-	<b>addictive</b> stupid free <b>full addicting</b>
	+	great good awesome favorite <b>cat-and-mouse</b>
sausage	-	silly <b>argentinian</b> cold blue stupid
	+	<b>hot grilled</b> good <b>sweet</b> awesome

\* Words in **bold** are more specific and can only describe certain targets.

# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model
  - Document Segmentation (with Lan Du and Mark Johnson)
  - With a Structured Topic Model
- 6 Conclusion

# Task, Roughly

## Unsegmented Text

The EPA has issued a rule requiring greater use of ethanol. Made from corn, ethanol helps gasoline burn cleaner and lowers some kinds of emissions. Well, if you drive a car, you'll really be interested in our next report. New rules about gasoline mean cleaner air but higher prices at the pump. CNN's Deborah Potter explains. The government is opening the way for more of this *corn* to wind up, not in the trough or on the table, but here, at the gas pump. To clear the air in the nation's smoggiest cities, the EPA has ordered the use of cleaner-burning gasoline. Adding renewable fuels to gasoline promotes products that are grown on American farms by American farmers. It promotes environmentally friendly jobs. It reduces air pollution. It protects the public's health.

## Segmented Text

The EPA has issued a rule requiring greater use of ethanol. Made from corn, ethanol helps gasoline burn cleaner and lowers some kinds of emissions.

Well, if you drive a car, you'll really be interested in our next report. New rules about gasoline mean cleaner air but higher prices at the pump. CNN's Deborah Potter explains.

The government is opening the way for more of this *corn* to wind up, not in the trough or on the table, but here, at the gas pump. To clear the air in the nation's smoggiest cities, the EPA has ordered the use of cleaner-burning gasoline.

Adding renewable fuels to gasoline promotes products that are grown on American farms by American farmers. It promotes environmentally friendly jobs. It reduces air pollution. It protects the public's health.

# Task, Roughly

**Passage:** contiguous text with no boundary, e.g., a sentence

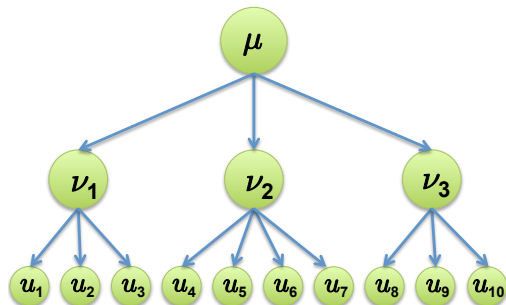
**Segment:** consecutive text passages that are semantically related.

**Document:** a sequence of topically coherent text segments.

**Document Segmentation Task:** (roughly) given a document as a monolithic block of text, where should we put the segment boundaries.

# Motivation–Structured Topic Modelling

Structured topic models (STM) by Du et al., (2010): hierarchical topic models with non-parametric Bayesian methods.

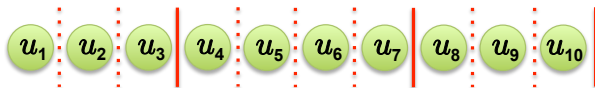


Has a hierarchy of topic probability vectors corresponding to document structure.

# Bayesian Segmentation

Bayesian word segmentation models (Goldwater et al., 2009)

- Learn to place boundaries after phonemes in an utterance.
- A pointwise boundary sampling algorithm: compute the probability of placing a word boundary after each phoneme.

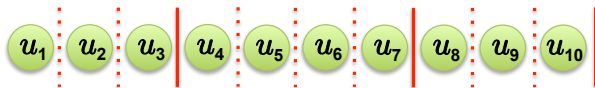




# Bayesian Segmentation

## Bayesian word segmentation models (Goldwater et al., 2009)

- Learn to place boundaries after phonemes in an utterance.
- A pointwise boundary sampling algorithm: compute the probability of placing a word boundary after each phoneme.



### Motivation:

- treat text passages like phonemes in the model,
- *i.e.*, estimate text passage boundaries as per phoneme boundaries.

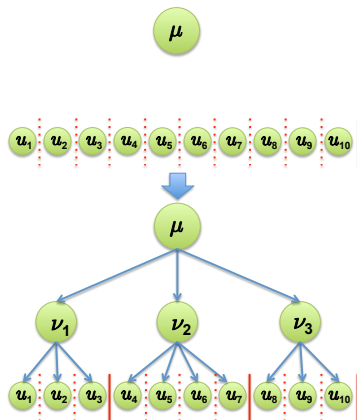
# Outline



- 1 Background
- 2 Non-parametric Bayesian Methods
- 3 High Performance Topic Models
- 4 Twitter Opinion Topic Model (with Kar Wai Lim)
- 5 Segmentation with a Structured Topic Model
  - Document Segmentation (with Lan Du and Mark Johnson)
  - With a Structured Topic Model
- 6 Conclusion

# Problem

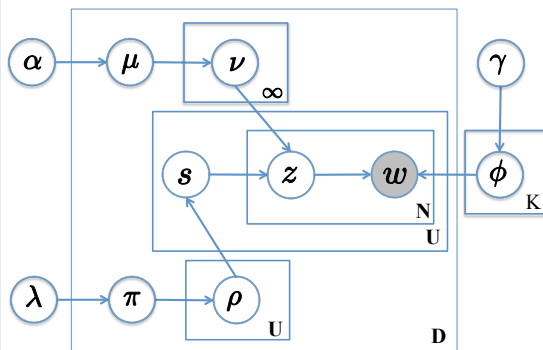
Problem:



**Hypothesis:** simultaneously learning topic segmentation and topic identification should allow better detection of topic boundaries.

# Segmentation Model–Generative process

## A segmentation model

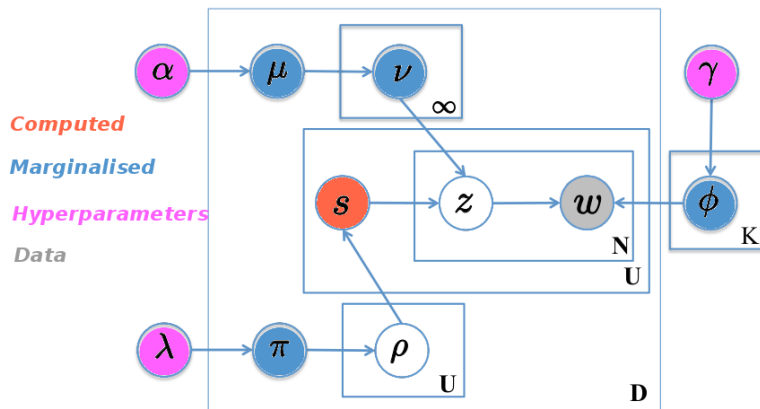


## Generative process

$$\begin{aligned}\vec{\phi} &\sim \text{Dirichlet}(\vec{\gamma}) \\ \vec{\mu} &\sim \text{Dirichlet}(\vec{\alpha}) \\ \pi &\sim \text{Beta}(\vec{\lambda}) \\ \vec{\nu} &\sim \text{PYP}(a, b, \vec{\mu}) \\ \rho &\sim \text{Bernoulli}(\pi) \\ z &\sim \text{Discrete}(\vec{\nu}_s) \\ w &\sim \text{Discrete}(\vec{\phi}_z)\end{aligned}$$

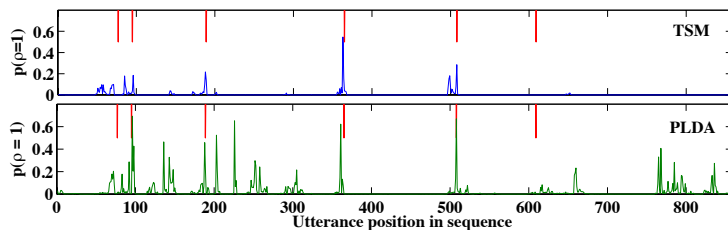
- $z$ : topic assignment of word  $w$ ;
- $N$ : the number of words in a passage.

# Posterior Inference—General Picture



- We need to sample the topic assignments  $z$  and segment boundaries  $\rho$ .

# Experiments on two meeting transcripts



**Figure:** Probability of a topic boundary, compared with gold-standard segmentation on one ICSI transcript.

Gold Standard	{77, 95, 189, 365, 508, 609, 860}
PLDA	{96, 136, 203, 226, 361, 508, 860}
TSM	{85, 96, 188, 363, 499, 508, 860}

# Conclusion of Segmentation with a Structured Topic Model

Paper given at NAACL 2013, main author Lan Du, also Mark Johnson

- A new hierarchical Bayesian model for unsupervised topic segmentation, using Bayesian segmentation + structured topic modelling.
- A novel sampling algorithm for splitting/merging restaurant(s) in CRP.

# Conclusion of Segmentation with a Structured Topic Model

Paper given at NAACL 2013, main author Lan Du, also Mark Johnson

- A new hierarchical Bayesian model for unsupervised topic segmentation, using Bayesian segmentation + structured topic modelling.
- A novel sampling algorithm for splitting/merging restaurant(s) in CRP.
- Code is available at Lan Du's website.
- Now running multi-core.



# Conclusion

- Latent Semantic Modelling with non-parametric Bayesians methods!
- See the individual papers.
- Read my blog and tutorials

<https://topicmodels.org>

*"A tutorial on non-parametric methods"*

## Thank You ... Questions?

# Alphabetic References

D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, **3**, 2003.

W.L. Buntine and S. Mishra, "Experiments with Non-parametric Topic Models," *KDD*, New York, 2014.

C. Chen, L. Du, L. and W. Buntine, "Sampling table configurations for the hierarchical Poisson-Dirichlet process," *ECML-PKDD*, Athens, 2011.

G. Doyle and C. Elkan, "Accounting for burstiness in topic models," *ICML*, Montreal, 2009.

L. Du, W. Buntine and H. Jin, "A segmented topic model based on the two-parameter Poisson-Dirichlet process," *Machine Learning Journal*, **81**(1), 2010.

L. Du, W. Buntine and M. Johnson, "Topic segmentation with a structured topic model," *NAACL-HLT*, Atlanta, 2013.

S. Goldwater, T. Griffiths and M. Johnson, "A Bayesian Framework for Word Segmentation: Exploring the Effects of Context," *Cognition*, **112**(1), 2009.

# Alphabetic References

- D. Newman, J.H. Lau, K. Grieser and T. Baldwin, "Automatic Evaluation of Topic Coherence," *NAACL-HLT*, Los Angeles, 2010.
- A. Puurula, "Cumulative Progress in Language Models for Information Retrieval," *ADCS*, Brisbane, 2013.
- S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, **3**(4), 2009.
- P. Sunehag, "Using two-stage conditional word frequency models to model word burstiness and motivating TF-IDF," *Artificial Intelligence and Statistics*, 2007.
- Y.W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," *ACL*, Sydney, 2006.
- Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei, "Hierarchical Dirichlet Processes," *JASA*, **101**(476), 2006.
- H.M. Wallach and I. Murray and R. Salakhutdinov and D. Mimno, "Evaluation Methods for Topic Models," *ICML*, Montreal, 2009.
- F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y.W. Teh. "A Stochastic Memoizer for Sequence Data," *ICML*, Montreal, 2009.