

## Case Study 5: Mixed Membership Modeling

# Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data  
CSE547/STAT548, University of Washington

Emily Fox  
March 6<sup>th</sup>, 2014

©Emily Fox 2014

1

## Task 2: Cluster Documents

- Then examined:
  - Cluster documents based on topic



©Emily Fox 2014

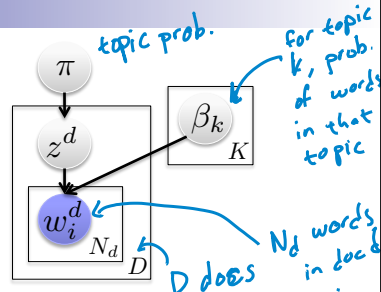
2

# A Generative Model

- Documents:  $x^1, \dots, x^D$
- Associated topics:  $z^1, \dots, z^D$
- Parameters:  $\theta = \{\pi, \beta\}$
- Generative model:

$z^d \sim \pi$  generate topic  
 $w_i^d | z^d \sim \beta_{z^d} \quad i=1, \dots, N_d$

Given topic  $z^d=k$  for doc  $d$ , draw each word from  $\beta_k \leftarrow$  word prob. for topic  $k$



©Emily Fox 2014

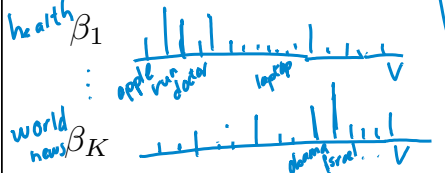
3

# Model In Pictures

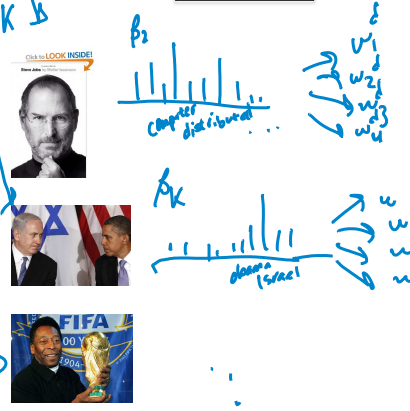
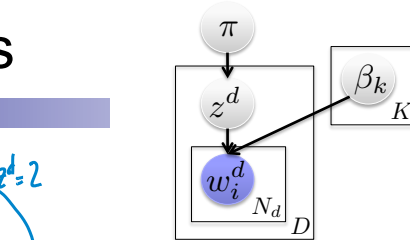
- Mixture weights (on topics)



- Topic distributions (on words)



- For each document,  
 $z^d \sim \pi$   
 $w_i^d | z^d \sim \beta_{z^d}$



©Emily Fox 2014

4

# Bayesian Document Model

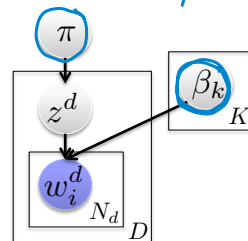
- Model parameters  $\pi, \{\beta_k\}$  unknown

← can use EM as in case study 2

- Bayesian approach

place priors on parameters

- Need distribution on pmf's



$$\sum_{k=1}^K \pi_k = 1$$

$$\sum_{k=1}^K \beta_k = 1$$

←  $\pi, \beta_k$  live on the simplex

- What is the simplex?
- What is a distribution on the simplex?

©Emily Fox 2014

5

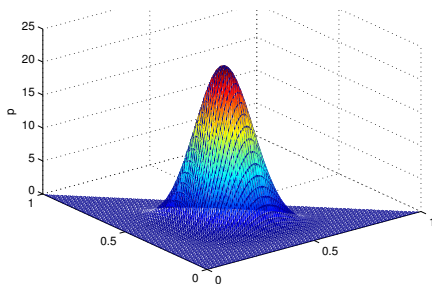
# Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

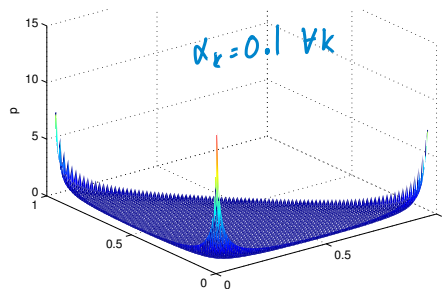
$$\alpha_k > 0 \quad \forall k$$

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$$

$$\Rightarrow \sum \pi_k = 1 \text{ and } \pi_k \geq 0 \quad \forall k$$



$$p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$



Moments:  $\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$

$$\text{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0+1)}$$

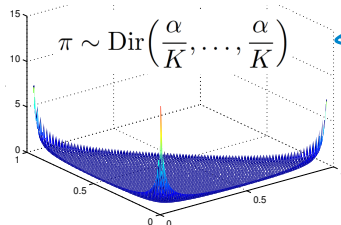
©Emily Fox 2014

6

# Model Summary

- Prior on model parameters

- E.g., symmetric Dirichlet for  $\pi$

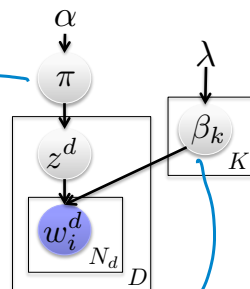


- Dirichlet prior for topic parameters  $\beta_k \sim \text{Dir}(\lambda_1, \dots, \lambda_v) \quad k=1, \dots, K$

- Sample observations as

$$z^d \sim \pi \quad d=1, \dots, D$$

$$w_i^d | z^d \sim \beta_{z^d} \quad i=1, \dots, N_d$$



©Emily Fox 2014

7

# Posterior Inference via Sampling

- Iterate between sampling

$$\pi \sim p(\pi | \{z^d\}, \{\beta_k\}, \{w_i^d\})$$

For  $k=1, \dots, K$

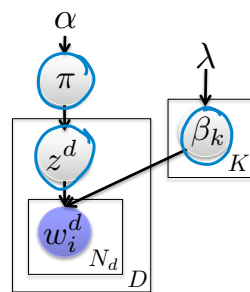
$$\beta_k \sim p(\beta_k | \pi, \{z^d\}, \{\beta_j, j \neq k\}, \{w_i^d\})$$

For  $d=1, \dots, D$

$$z^d \sim p(z^d | \pi, \{z^i, i \neq d\}, \{\beta_k\}, \{w_i^d\})$$

- What form do these complete conditionals take?

- First a look at statements of conditional independence in directed graphical models

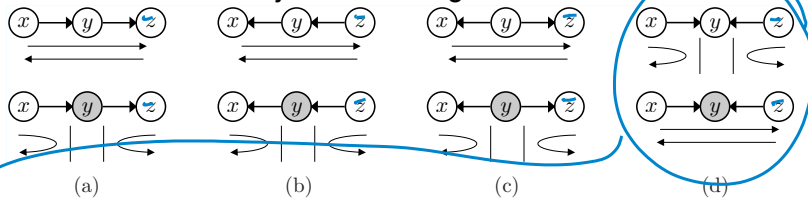


©Emily Fox 2014

8

# Conditional Independence in Bayes Nets

- Consider 4 different junction configurations



- Conditional versus unconditional independence:

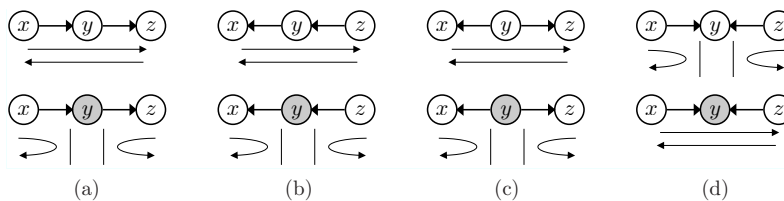
$P(x, y, z) = P(x)P(z)P(y|x, z)$  <sup>int. over y</sup>  $\Rightarrow P(x, z) = P(x)P(z) \Rightarrow x \perp\!\!\!\perp z$   
 $P(x, z|y) \neq P(x, y, z) = P(x)P(z)P(y|x, z)$   
 $\neq P(x|y)P(z|y) \leftarrow x \not\perp\!\!\!\perp z | y$   
 "explaining away":  $x = \text{earthquake}$ ,  $z = \text{burglar}$ ,  $y = \text{car alarm}$   
 If alarm ( $y=1$ ), an increase in earthquake  $p(x|y)$ , means  $p(z|y)$  lower  
ind. a priori

©Emily Fox 2014

9

# Bayes Ball Algorithm

- Consider 4 different junction configurations



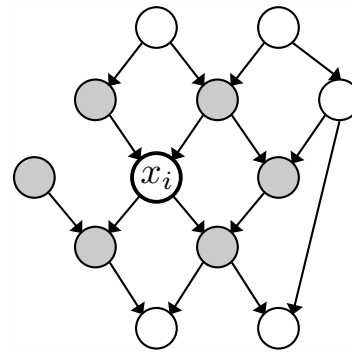
- Bayes ball algorithm

©Emily Fox 2014

10

# Markov Blanket

- A node is conditionally independent of all other nodes in the graph given its Markov blanket



- Gibbs sampling iterates between full conditionals

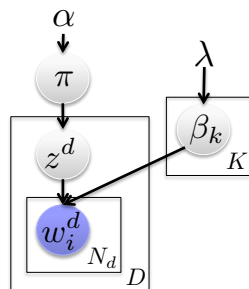
→ simplify to

©Emily Fox 2014

11

# Unplated Document Model

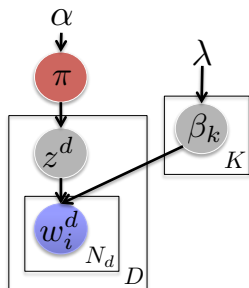
- Recall that the plate notation is really indicating



©Emily Fox 2014

12

# Complete Conditional for $\pi$



- Recall conjugate Dirichlet prior

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

- Likelihood:

- Dirichlet posterior

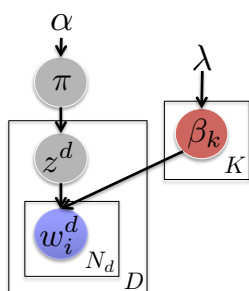
- Count occurrences of
- Then,

- Conjugacy: **Posterior** has same form as **prior**

©Emily Fox 2014

13

# Complete Conditional for $\beta_k$



- Again, Dirichlet prior

- Consider docs  $d$  such that

- For these observations,
- Do any other docs depend on  $\beta_k$ ?

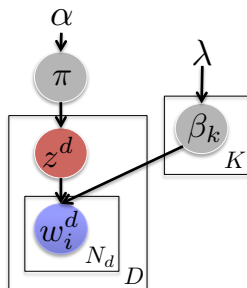
- Then,

- Again, **posterior** has same form as **prior**

©Emily Fox 2014

14

## Complete Conditional for $z^d$



- We have  $z^d \sim \pi$

$$w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d}$$

- Calculate the posterior for each value of  $z^d$  (“responsibility” of each topic to the doc):

$$r_{dk} = p(z^d = k \mid \{w_i^d\}, \pi, \beta) = \frac{\pi_k p(\{w_i^d\} \mid \beta_k)}{\sum_j \pi_j p(\{w_i^d\} \mid \beta_j)}$$

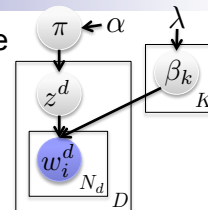
- Sample each cluster indicator as

©Emily Fox 2014

15

## Collapsed Gibbs Sampler

- In conjugate models, can analytically marginalize some variables and only sample remaining



- Can improve efficiency if marginalized variables are high-dim
  - Reduced dimension of search space
  - But, often introduces dependences!

©Emily Fox 2014

16

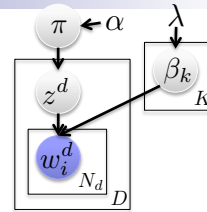


# Collapsed Sampler Full Conditional

$$p(\cdot) = p(\pi | \alpha) \prod_{d=1}^D p(z^d | \pi) \left( \prod_{k=1}^K p(\beta_k | \lambda) \prod_{d=1}^D \prod_{i=1}^{N_d} p(w_i^d | z^d, \beta) \right)$$

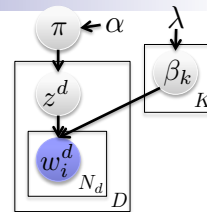
■ Derivation

$$p(z^d = k | z_{\setminus d}, \{w_i^d\}, \alpha, \lambda) \propto \int_{\pi} \int_{\beta_1} \dots \int_{\beta_K} p(\cdot)$$



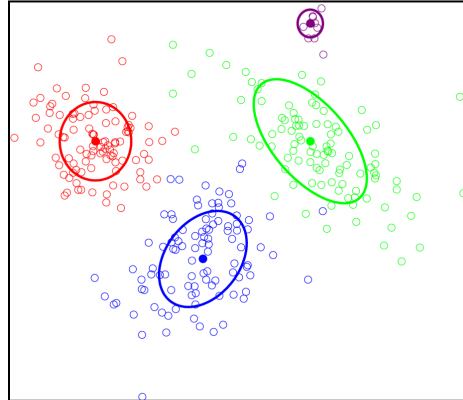
# Collapsed Sampler Full Conditional

$$p(z^d = k | z_{\setminus d}, \{w_i^d\}, \alpha, \lambda) \propto p(z^d = k | z_{\setminus d}, \alpha) p(\{w_i^d\} | \{w_i^c : z^c = k, c \neq d\})$$



# Collapsed Sampler Intuition (MoG)

- Previously,  $p(z^i = k | x^i, \pi, \theta) \propto \pi_k p(x^i | \theta_k)$
- If you're not told  $\pi, \theta_k$



©Emily Fox 2014

19

# Example – Uncollapsed Results

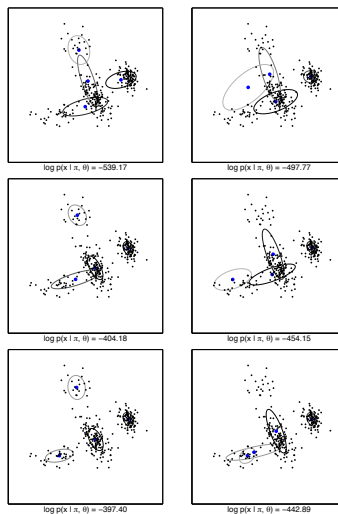


Figure courtesy of Erik Sudderth

©Emily Fox 2014

20

## Example – Collapsed Results

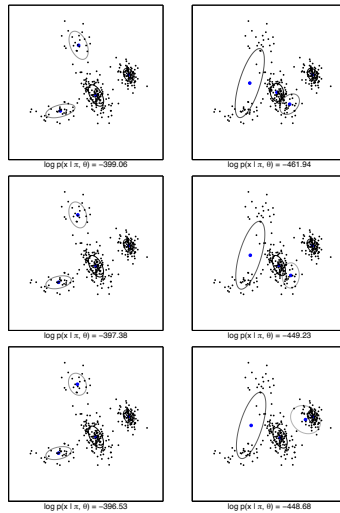


Figure courtesy of Erik Sudderth

©Emily Fox 2014

21

## Comparing Collapsed vs. Uncollapsed

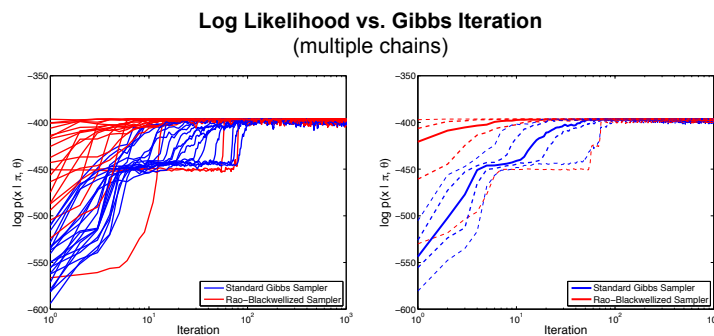


Figure courtesy of Erik Sudderth

©Emily Fox 2014

22

# Task 3: Mixed Membership Models

- **Now:** Document may belong to multiple clusters

The image shows a screenshot of a New York Times article titled "Students Rush to Web Classes, but Profits May Be Much Later" under the "Education" section. The article features a colorful graphic with the text "CALCULUS single variable" and "LECTURE 1 FUNCTIONS". Three blue arrows originate from the graphic and point to the labels "EDUCATION", "FINANCE", and "TECHNOLOGY", illustrating mixed membership.

©Emily Fox 2014

23

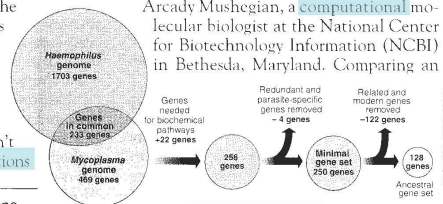
# Latent Dirichlet Allocation (LDA)

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



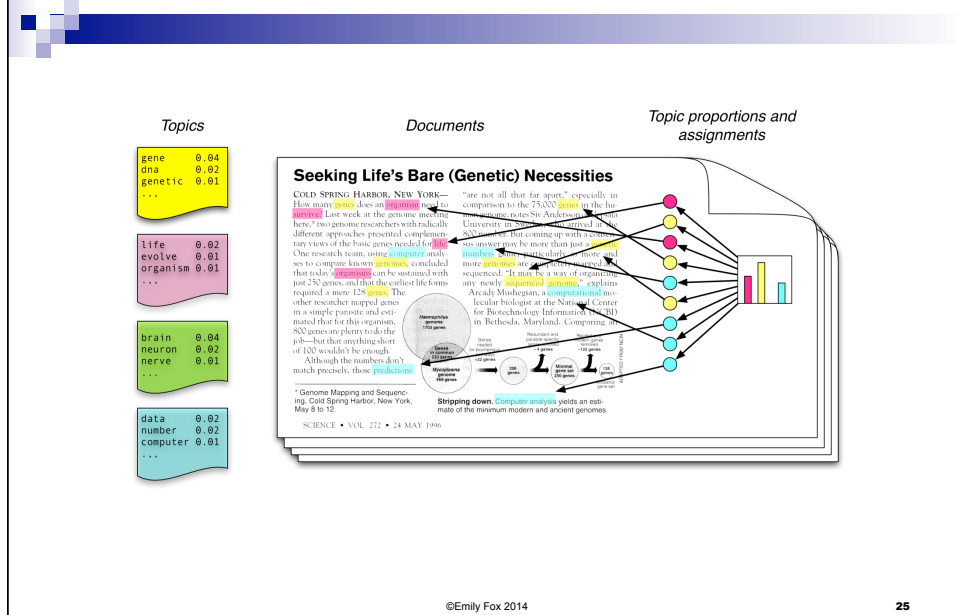
**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

ADAPTED FROM NCBI

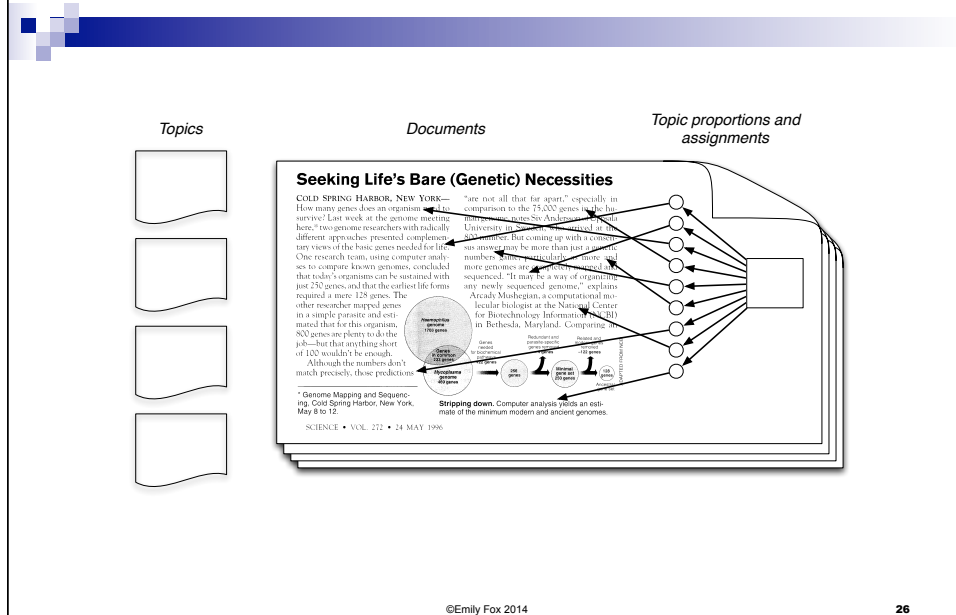
©Emily Fox 2014

24

# Latent Dirichlet Allocation (LDA)



# Latent Dirichlet Allocation (LDA)



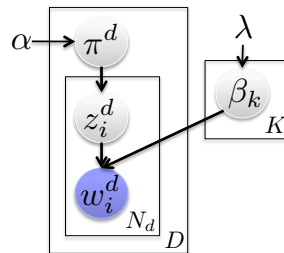
# LDA Generative Model

- Observations:  $w_1^d, \dots, w_{N_d}^d$
- Associated topics:  $z_1^d, \dots, z_{N_d}^d$
- Parameters:  $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

©Emily Fox 2014

27

# LDA Joint Probability



$$p(\cdot) = \prod_{k=1}^K p(\beta_k | \lambda) \prod_{d=1}^D p(\pi^d | \alpha) \left( \prod_{i=1}^{N_d} p(z_i^d | \pi^d) p(w_i^d | z_i^d, \beta) \right)$$

©Emily Fox 2014

28

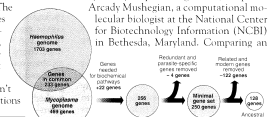
# Example Inference – Topic Weights

- **Data:** The OCR'ed collection of *Science* from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model

## Seeking Life's Bare (Genetic) Necessities

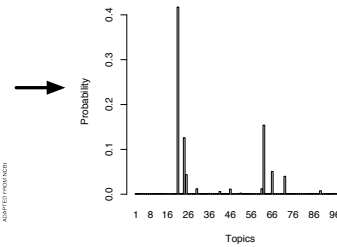
**COLD SPRING HARBOR, NEW YORK—** How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 120 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.



# Example Inference – Topic Words

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

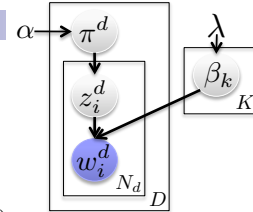
# Collapsed LDA Sampling

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$

$$\propto p(z_i^d = k \mid z_{\setminus id}, \alpha) p(w_i^d \mid z_i^d = k, z_{\setminus id}, w_{\setminus id}, \lambda)$$

- Unplate to see dependencies induced



©Emily Fox 2014

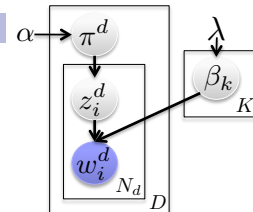
31

# Collapsed LDA Sampling

- Sample topic indicators for each word
  - Algorithm:

$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$

$$\propto p(z_i^d = k \mid \{z_j^d, j \neq i\}, \alpha) p(w_i^d \mid \{w_j^c : z_j^c = d, (j, c) \neq (i, d)\}, \lambda)$$



©Emily Fox 2014

32



## Select a Document

Etruscan	trade	price	temple	market

©Emily Fox 2014

33

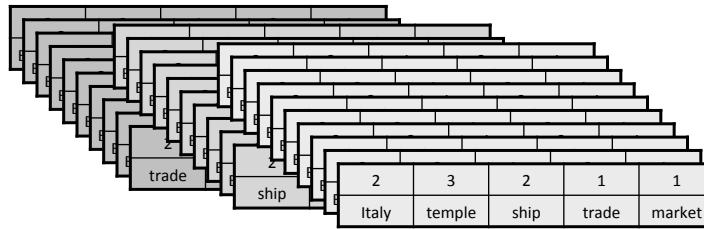
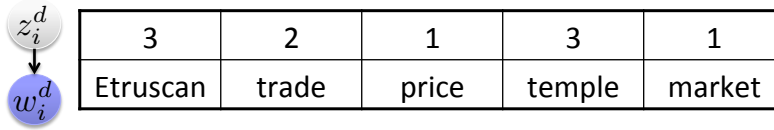
## Randomly Assign Topics

$z_i^d$	3	2	1	3	1
$w_i^d$	Etruscan	trade	price	temple	market

©Emily Fox 2014

34

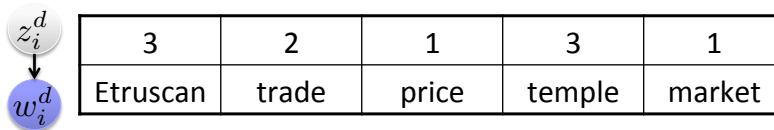
# Randomly Assign Topics



©Emily Fox 2014

35

# Maintain Local Statistics

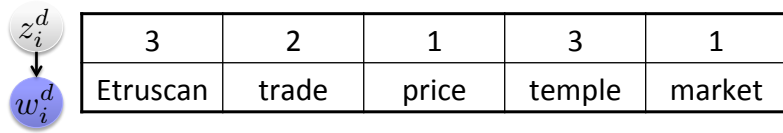


	Topic 1	Topic 2	Topic 3
Doc d			

©Emily Fox 2014

36

# Maintain Global Statistics



	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

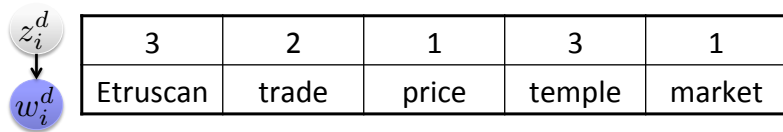
	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

Total counts from **all** docs

©Emily Fox 2014

37

# Resample Assignments



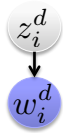
	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

©Emily Fox 2013

38


What is the conditional distribution for this topic?



3	?	1	3	1
Etruscan	trade	price	temple	market

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?



3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Topic 2

Topic 3

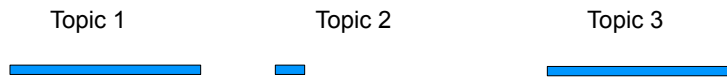


	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

## What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d$	3	?	1	3	1
$w_i^d$	Etruscan	trade	price	temple	market



	Topic 1	Topic 2	Topic 3
trade	10	7	1

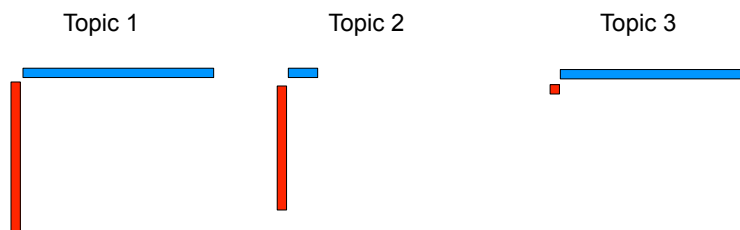
©Emily Fox 2014

41

## What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d$	3	?	1	3	1
$w_i^d$	Etruscan	trade	price	temple	market



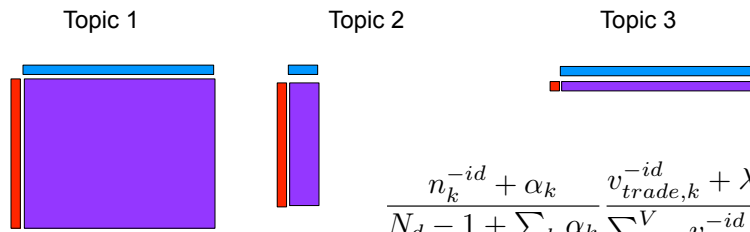
©Emily Fox 2014

42

## What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d$	3	?	1	3	1
$w_i^d$	Etruscan	trade	price	temple	market



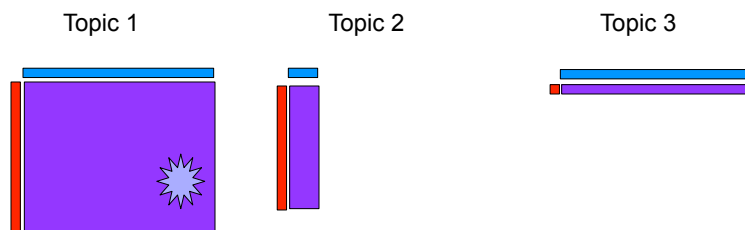
$$\frac{n_k^{-id} + \alpha_k}{N_d - 1 + \sum_k \alpha_k} \frac{v_{trade,k}^{-id} + \lambda_{trade}}{\sum_{\gamma=1}^V v_{\gamma,k}^{-id} + \lambda_{\gamma}}$$

©Emily Fox 2014

43

## Sample a New Topic Indicator

$z_i^d$	3	?	1	3	1
$w_i^d$	Etruscan	trade	price	temple	market



©Emily Fox 2014

44

# Update Counts

$z_i^d$   
 $w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			

	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

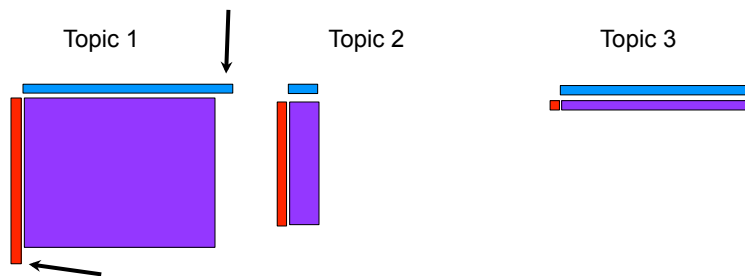
©Emily Fox 2014

45

# Geometrically...

$z_i^d$   
 $w_i^d$

3	1	1	3	1
Etruscan	trade	price	temple	market



©Emily Fox 2014

46

## Issues with Generic LDA Sampling

- Slow mixing rates → Need many iterations
- Each iteration cycles through sampling topic assignments for *all* words in *all* documents
- Modern approaches include:
  - Large-scale LDA. For example, [Mimno, David, Matthew D. Hoffman and David M. Blei. "Sparse stochastic inference for latent Dirichlet allocation." International Conference on Machine Learning, 2012.](#)
  - Distributed LDA. For example, [Ahmed, Amr, et al. "Scalable inference in latent variable models." Proceedings of the fifth ACM international conference on Web search and data mining \(2012\): 123-132](#)
  - And many, many more!
- Alternative: Variational methods instead of sampling
  - Approximate posterior with an optimized variational distribution

©Emily Fox 2014

47

## What you need to know...

- Bayesian specification of document clustering model
- Rules of conditional and unconditional independence in directed graphical models (Bayes nets)
  - Bayes' ball
  - Markov blanket
- Gibbs sampling for Bayesian document model
- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler

©Emily Fox 2014

48



# Reading

## ■ Mixed Membership Models: KM Sec. 27.3

- Basic LDA:  
[Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 \(2003\): 993-1022.](#)
- Introduction:  
[Blei, David M. "Probabilistic topic models." Communications of the ACM, vol. 55, no. 4 \(2012\): 77-84.](#)
- Sampling:  
[Griffith, Thomas L. and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America, Volume: 101, Supplement: 1 \(2004\): Pages: 5228-5235](#)

# Acknowledgements

- Thanks to Dave Blei, David Mimno, and Jordan Boyd-Graber for some material in this lecture relating to LDA