topic models:
priors, stop words and languages

hanna m. wallach

university of massachusetts amherst

wallach@cs.umass.edu

Finding Needles in Haystacks

www.betaversion.org/~stefano/linotype/news/26/

- As more information becomes available, it can be harder and harder to find what we want
- We don't even always know what we want!
- Need new tools to help us organize, search and understand information

A Solution: Topic Models

Candida Hofer

- Use topic models to discover hidden topic-based patterns
- Use discovered topics to annotate the collection
- Use annotations to organize, understand, summarize, search...

Topics ↔ Words

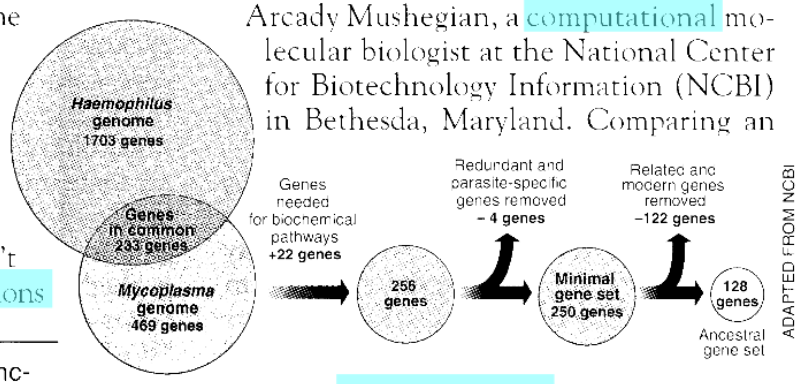
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Documents ⇔ Topics

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 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 "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.

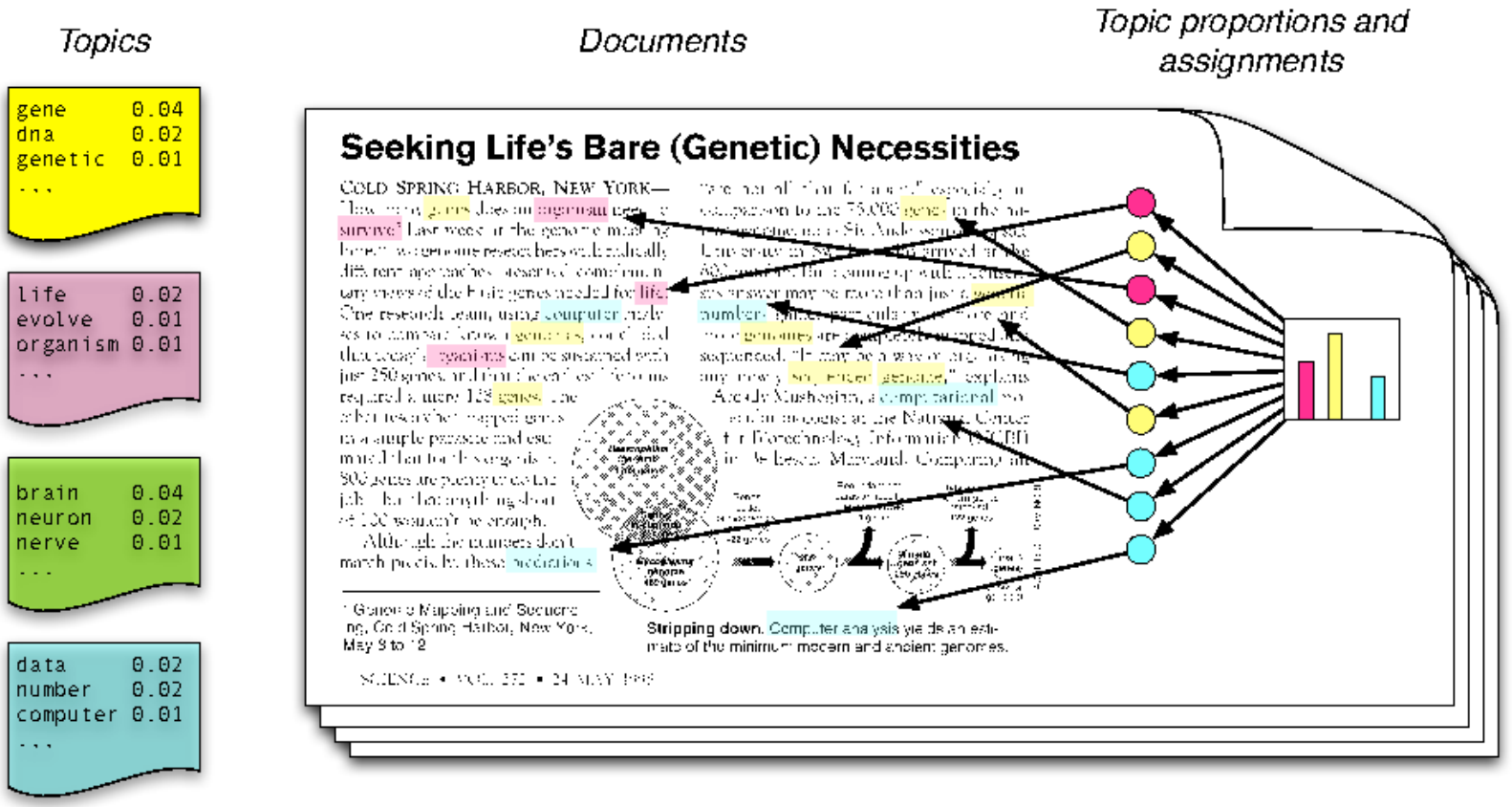
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Probabilistic Modeling

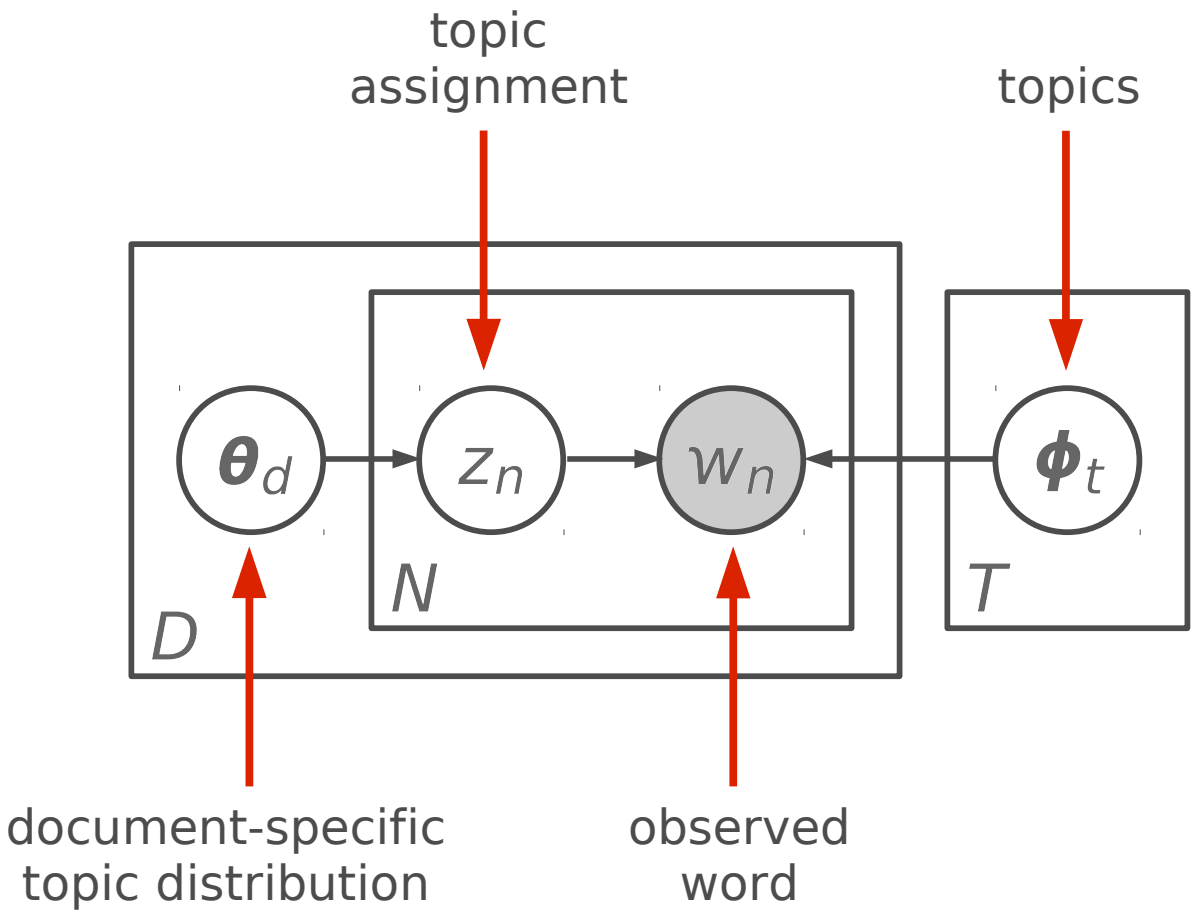
- Treat data as observations that arise from a generative probabilistic process that includes hidden variables:
  - For documents, the hidden variables represent the thematic structure of the collection
- Infer the hidden structure using posterior inference:
  - What are the topics that describe this collection?
- Situate new data into the estimated model:
  - Which topics best describe the new documents?

Documents ↔ Topics ↔ Words



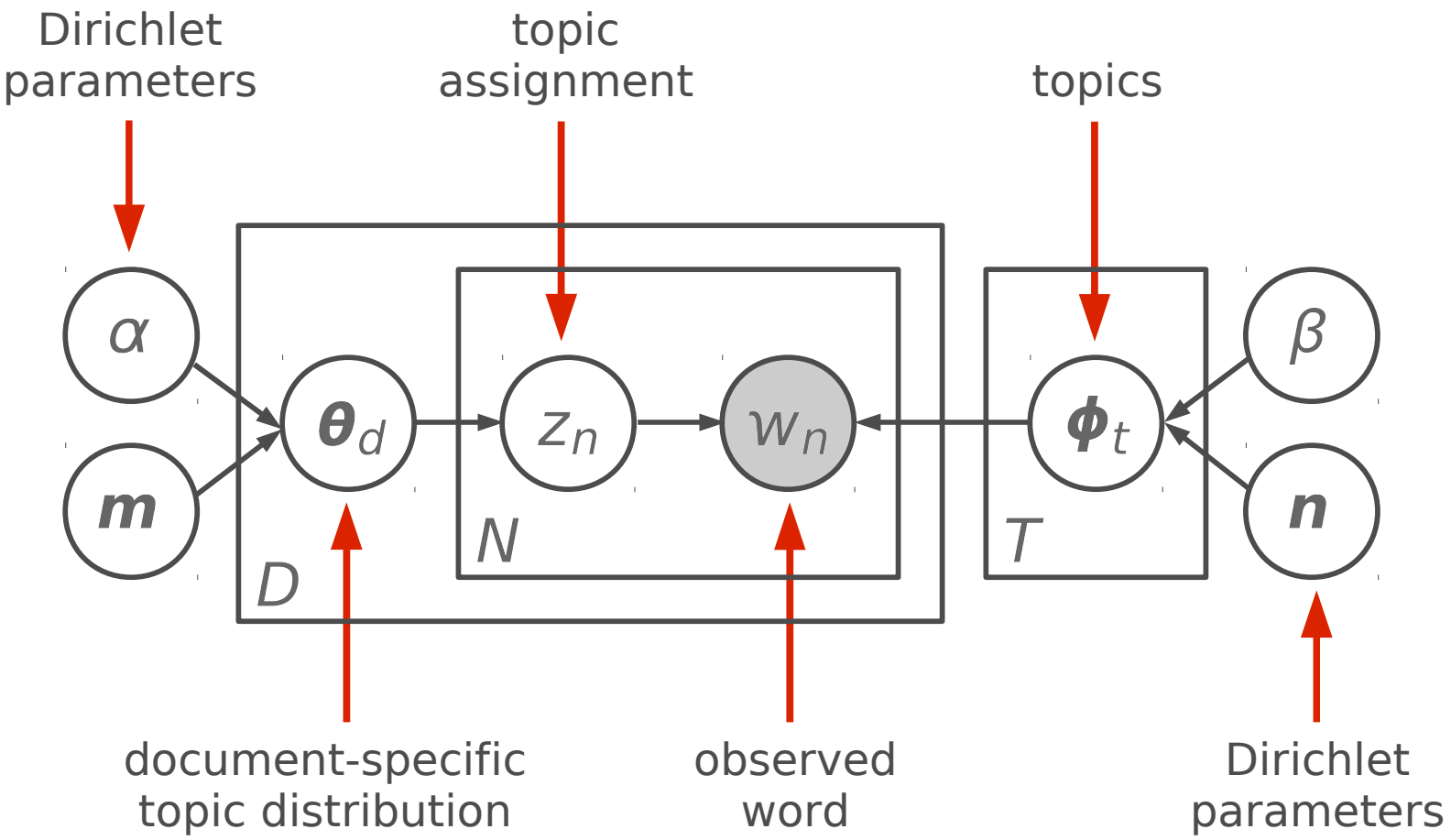
Probabilistic LSA

(Hofmann, 1999)



Latent Dirichlet Allocation

(Blei et al., 2003)



Dirichlet Distribution

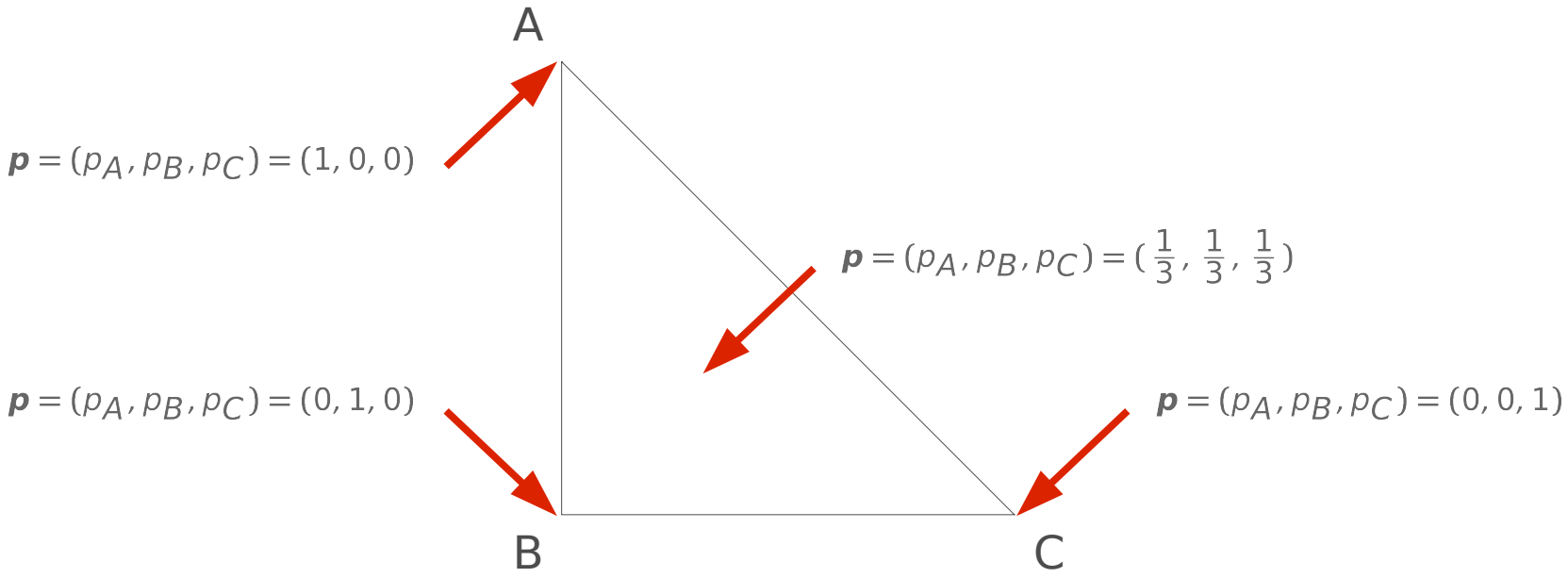
- Distribution over K-dimensional positive vectors that sum to one (i.e., points on the probability simplex)

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

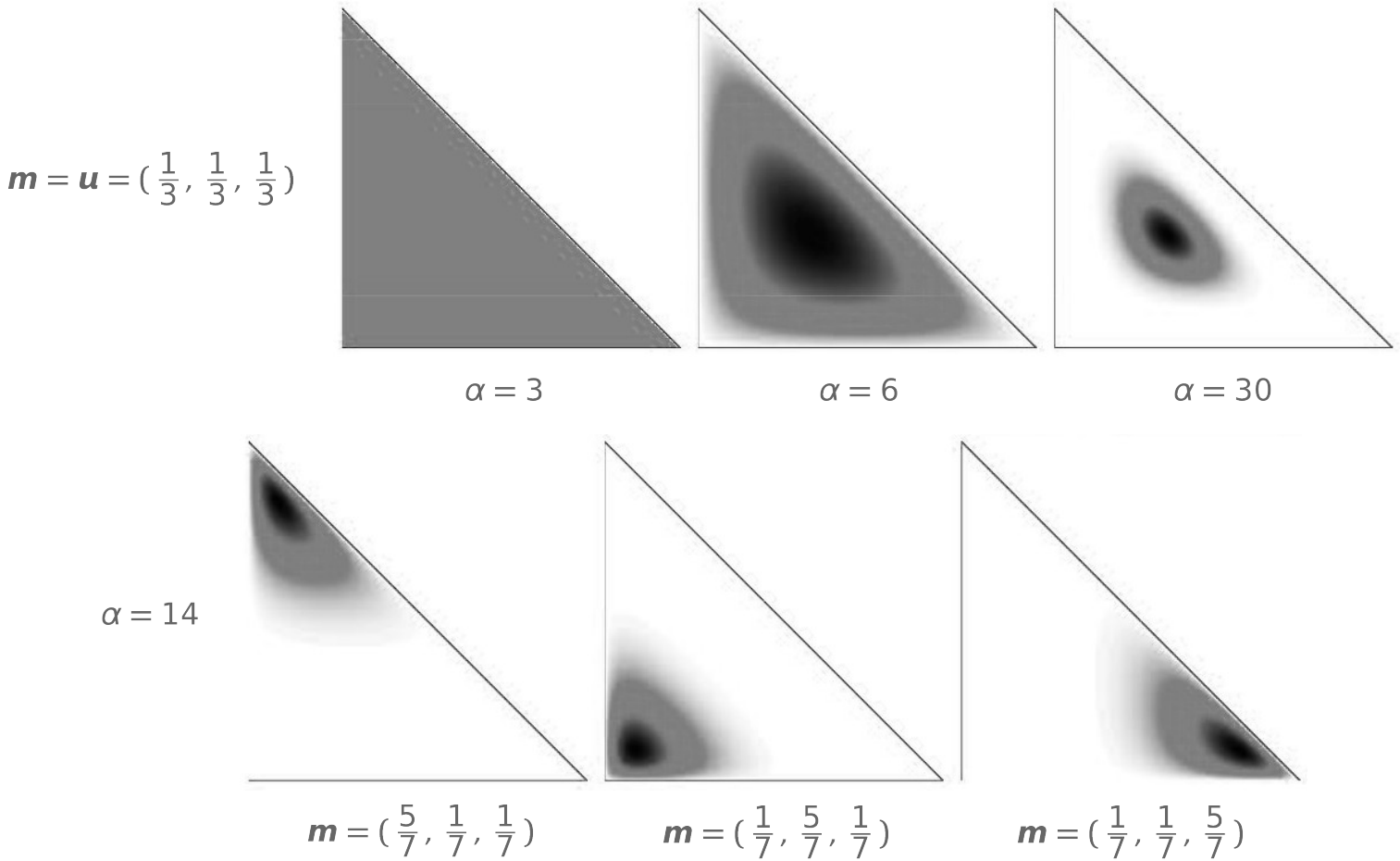
- Two parameters:
  - Base measure $\boldsymbol{m}$ (positive vector; sums to one)
  - Concentration parameter α (positive scalar)

An Aside: The Simplex

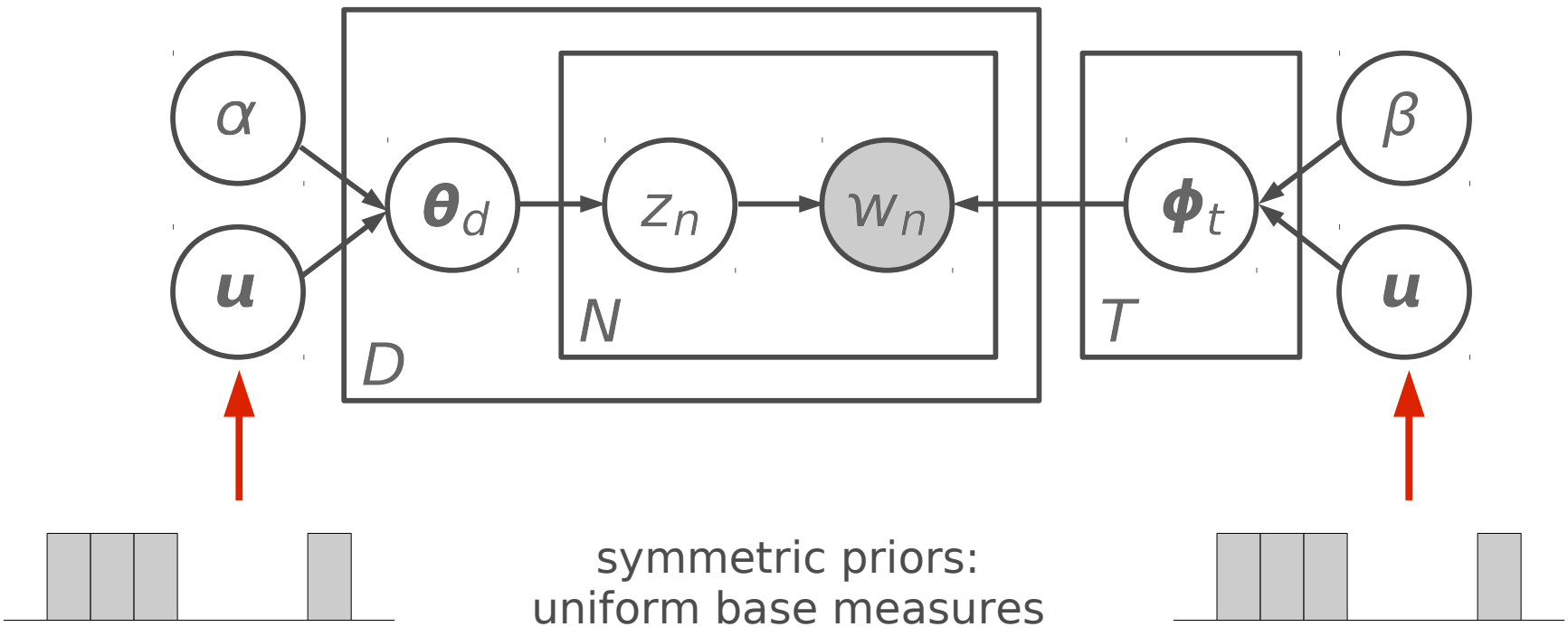
- K-dimensional probability distributions (i.e., points on the K-1 simplex) can be plotted in (K-1)-d:



Dirichlet Parameters



Latent Dirichlet Allocation



rethinking lda: why priors matter

hanna m. wallach, david mimno, andrew mccallum

Priors for LDA

- Almost all work on LDA uses symmetric Dirichlet priors
  - Two scalar concentration parameters: α and β
- Concentration parameters are usually set heuristically
- Some recent work on inferring optimal concentration parameter values from data (Asuncion et al., 2009)
- No rigorous study of the Dirichlet priors:
  - Base measure: asymmetric vs. symmetric
  - Treatment: optimize vs. integrate out

Topic Modeling in Practice

- Help! All my topics consist of “the, and of, to, a ...”
  - Preprocess data to remove stop words
- Now all my topics consist of “data, model, results ...”
  - Make a new corpus-specific stop word list
- Wait, but how do I choose the right number of topics T
  - Evaluate probability of held-out data for different T
- That sounds really time-consuming
  - Use a nonparametric model ...

Symmetric → Asymmetric

- Use prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ as a running example
- Uniform base measure → nonuniform base measure

$\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{u})$ $\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{m})$

- Asymmetric prior: some topics more likely a priori

Predictive Distributions

- Predictive probability of topic t in document d given $\mathcal{Z}$

$$P(t \mid d, \mathcal{Z}, \alpha \boldsymbol{m}) = \int \mathrm{d}\boldsymbol{\theta}_d \, P(t \mid \boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d \mid \mathcal{Z}, \alpha \boldsymbol{m})$$

$$= \frac{N_{t|d} + \alpha m_t}{N_d + \alpha}$$

- If t has not yet occurred in d then $P(t \mid d, \mathcal{Z}, \alpha \boldsymbol{m}) = m_t$
- $N_{t|d}$ is smoothed with topic-specific quantity αm_t

Handling Unknown $\boldsymbol{m}$

- Can take a fully Bayesian approach:
  - Give $\boldsymbol{m}$ a Dirichlet prior: $\boldsymbol{m} \sim \mathrm{Dir}(\alpha' \boldsymbol{u})$
  - Integrate $\boldsymbol{m}$ out thanks to conjugacy:

$$P(t \mid d, \mathcal{Z}, \alpha, \alpha' \boldsymbol{u}) = \int \mathrm{d}\boldsymbol{m}\, P(t \mid d, \mathcal{Z}, \alpha \boldsymbol{m}) P(\boldsymbol{m} \mid \mathcal{Z}, \alpha' \boldsymbol{u})$$

$$= \frac{N_{t|d} + \alpha \dfrac{\hat{N}_t + \frac{\alpha'}{T}}{\sum_t \hat{N}_t + \alpha'}}{N_d + \alpha}$$

An Observation

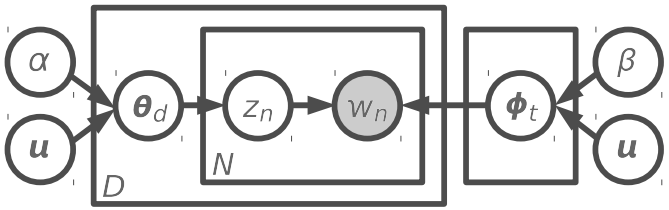
- As α' → ∞, the asymmetric hierarchical Dirichlet prior over Θ approaches a symmetric Dirichlet prior:

$$\frac{N_{t|d} + \alpha \dfrac{\hat{N}_t + \frac{\alpha'}{T}}{\sum_t \hat{N}_t + \alpha'}}{N_d + \alpha} \rightarrow \frac{N_{t|d} + \frac{\alpha}{T}}{N_d + \alpha}$$

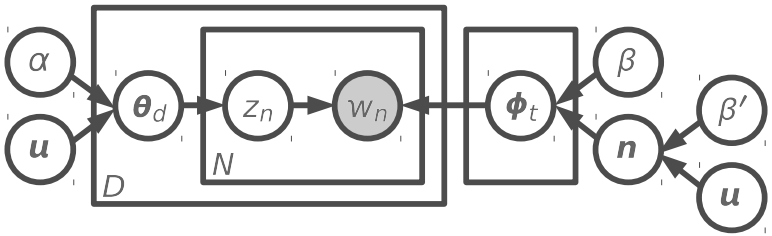
- Symmetric Dirichlet prior is a special case of the asymmetric hierarchical Dirichlet prior

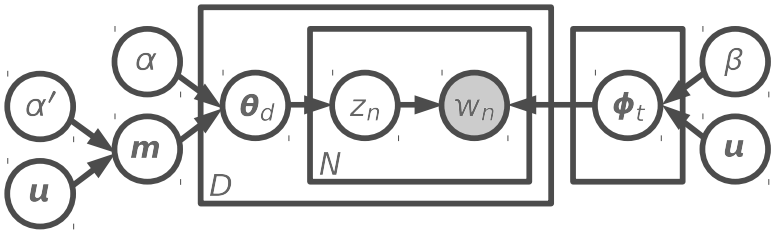
Four Combinations of Priors

"SS"

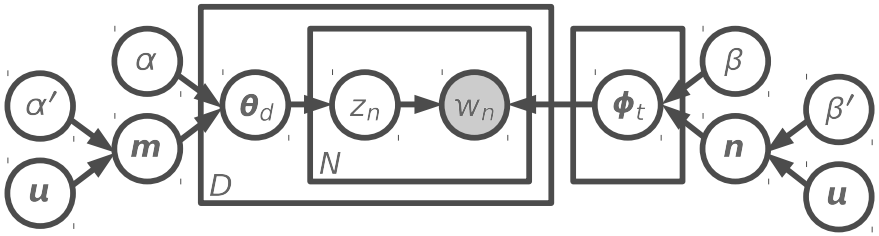


"SA"





"AS"

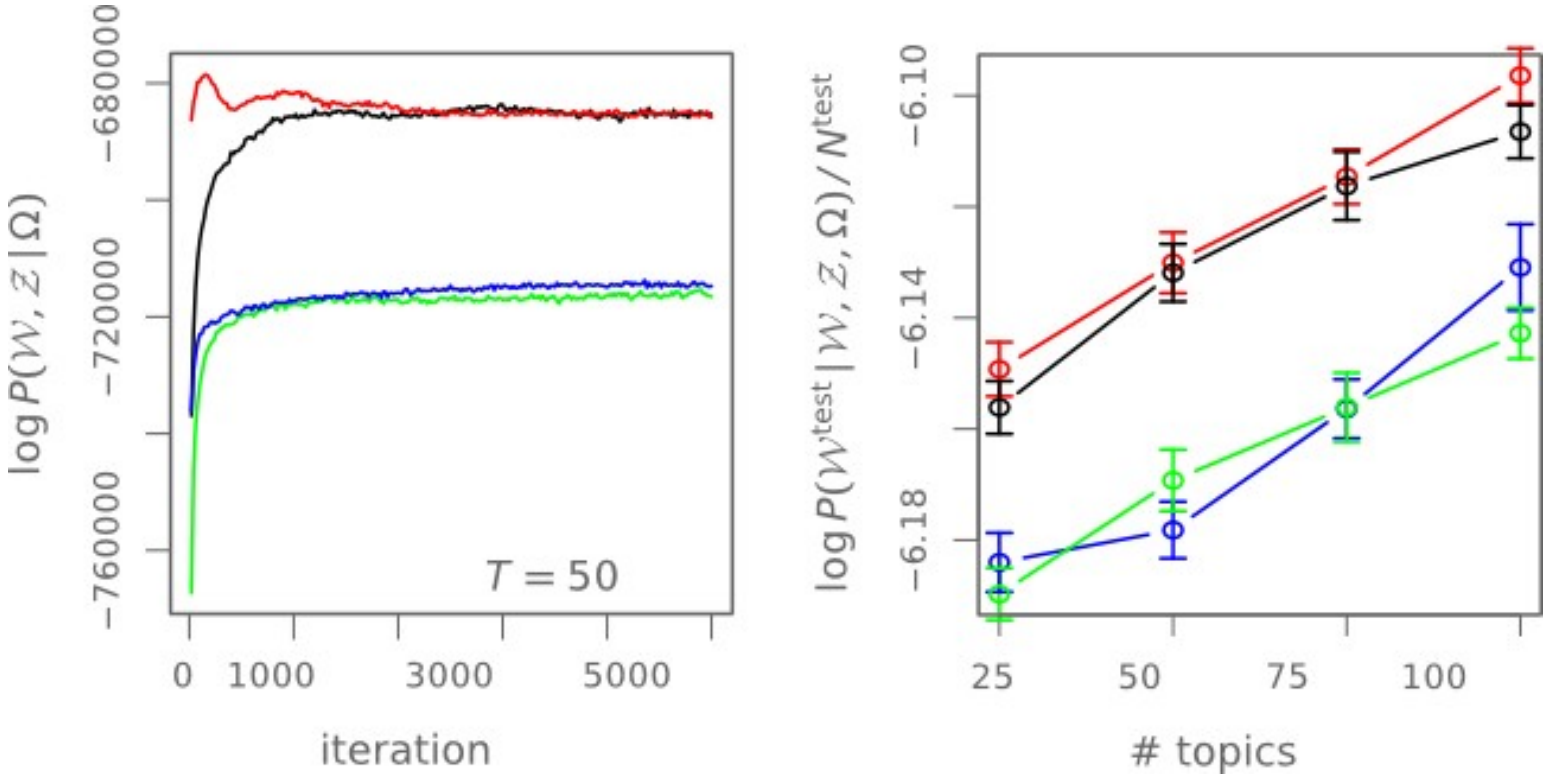


"AA"

Inferred Topics and Stop Words

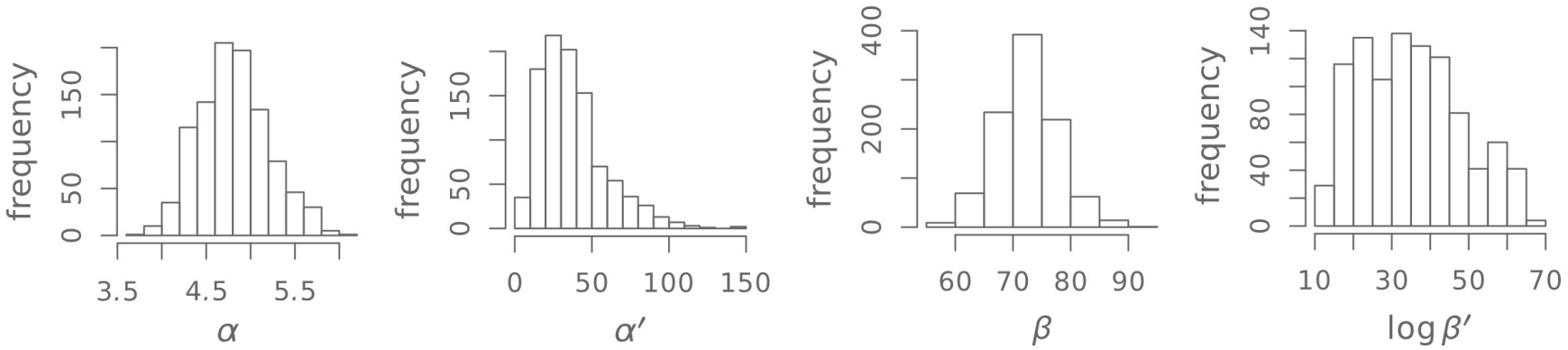
		symm. prior over Φ	asymm. prior over Φ
symm. Θ	0.080	a **field emission** an **electron** the	0.042 a **field** the **emission** and **carbon** is
	0.080	a the **carbon** and **gas** to an	0.042 the **carbon catalyst** a **nanotubes**
	0.080	the of a to and about at	0.042 a the of **susbtrate** to **material** on
	0.080	of a **surface** the with in **contact**	0.042 **carbon single wall** the **nanotubes**
	0.080	the a and to is of **liquid**	0.042 the a **probe tip** and of to
asymm. Θ	0.895	the a of to and is in	1.300 the a of to and is in
	0.187	**carbon nanotubes nanotube catalyst**	0.257 and are of for in as such
	0.043	sub is c or and n sup	0.135 a **carbon material** as **structure nanotube**
	0.061	**fullerene compound fullerenes**	0.065 **diameter swnt** about **nm** than **fiber swnts**
	0.044	**material particles coating inorganic**	0.029 **compositions polymers polymer contain**

Results: Log Probabilities



Sampled Concentration Parameters

- Sampled concentration parameters from “AA”



- β' Is large compared to $\sum_w \hat{N}_w$
- Prior over Φ is effectively symmetric: “AA” → “AS”

Intuition

- Topics are specialized distributions over words
  - Want topics to be as distinct as possible
  - Asymmetric prior over $\{ \boldsymbol{\phi}_t \}$ makes topics more similar to each other (and to the corpus word frequencies)
  - Want a symmetric prior to preserve topic "distinctness"
- Still have to account for power-law word usage:
  - Asymmetric prior over $\{ \boldsymbol{\theta}_d \}$ means some topics (e.g., "the, a, of, to ...") can be used more often than others

Conclusions

- Careful thinking about priors can yield new insights
  - e.g., priors and stop word handling are related
- For LDA the choice of prior is surprisingly important:
  - Asymmetric prior for document-specific topic distributions
  - Symmetric prior for topic-specific word distributions
- Rethinking priors for LDA facilitates new topic models
  - e.g., polylingual topic model ...

NESCAI 2010

- April 16-18 @ UMass Amherst
- http://nescai.cs.umass.edu/cfp.php

questions?