

# A generic approach to topic models and its application to virtual communities

Gregor Heinrich

PhD presentation (English translation incl. backup slides, 45min)

Faculty of Mathematics and Computer Science  
University of Leipzig

28 November 2012

Version 2.9 EN BU

# A generic approach to topic models and its application to virtual communities

Gregor Heinrich

PhD presentation (English translation incl. backup slides, 45min)

Faculty of Mathematics and Computer Science  
University of Leipzig

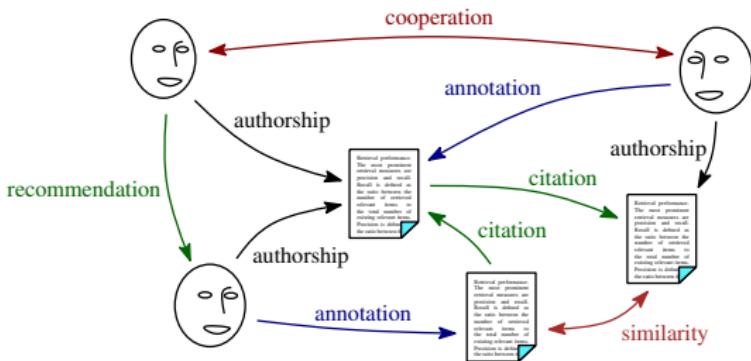
28 November 2012

Version 2.9 EN BU

# Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook

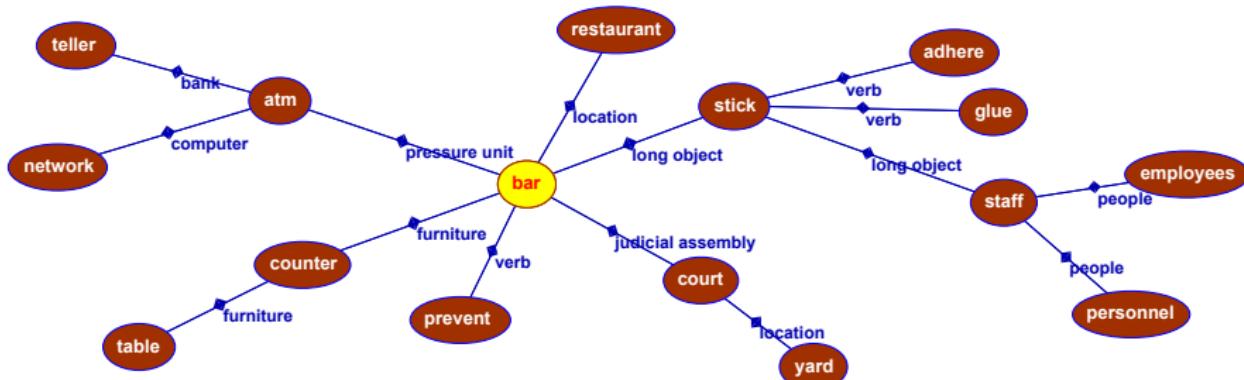
# Motivation: Virtual communities



- Virtual communities = groups of persons who exchange information and knowledge electronically
- Examples: organisations, digital libraries, “Web 2.0” applications incl. social networks
- Data are multimodal: text content; authorship, citation, annotations and recommendations; cooperation and other social relations
- Typical case: discrete data with high dynamics and large volumes

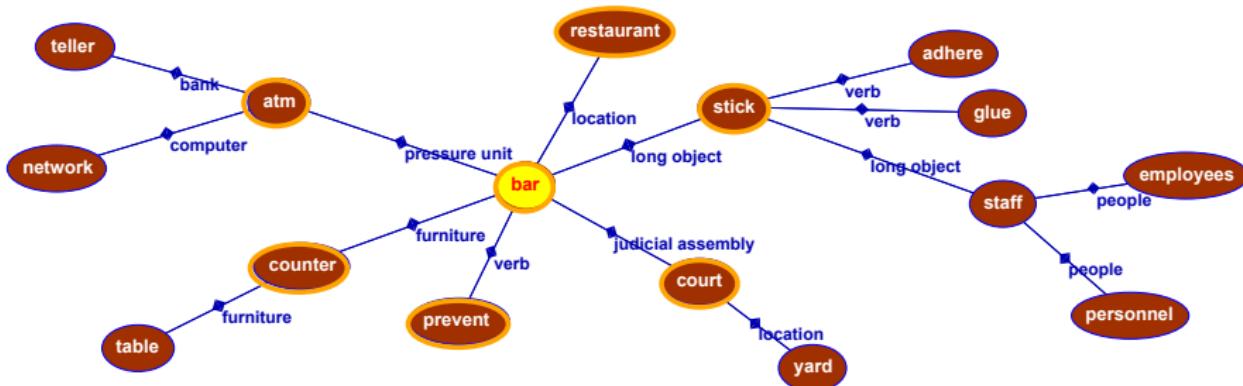
# Motivation: Unsupervised mining of discrete data

- Identification of relationships in large data volumes
- Only data (and possibly model) required (information retrieval, network analysis, clustering, NLP methods)
- **Density problem:** Features too sparse for analysis in high-dimensional feature space
- **Vocabulary problem:** Semantic similarity  $\neq$  lexical similarity (polysemy, synonymy, etc.)

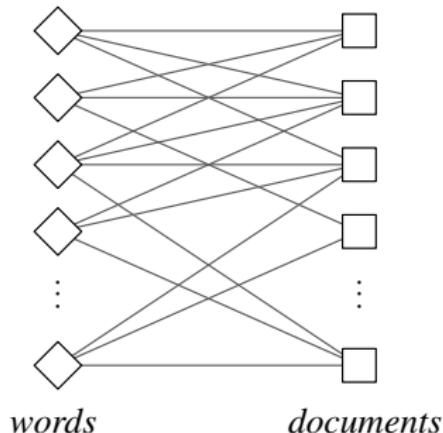


# Motivation: Unsupervised mining of discrete data

- Identification of relationships in large data volumes
- Only data (and possibly model) required (information retrieval, network analysis, clustering, NLP methods)
- **Density problem:** Features too sparse for analysis in high-dimensional feature space
- **Vocabulary problem:** Semantic similarity  $\neq$  lexical similarity (polysemy, synonymy, etc.)

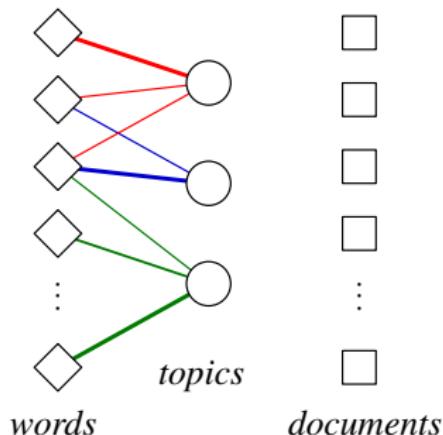


# Topic models as approach



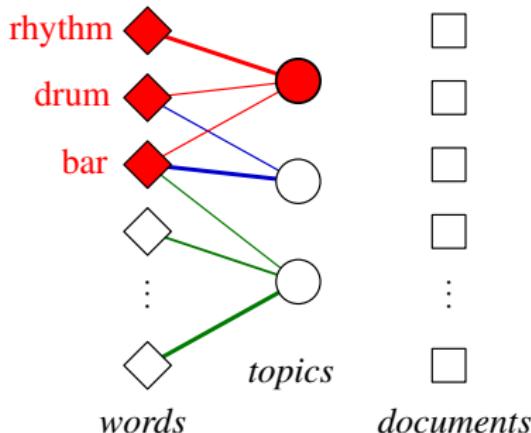
- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

# Topic models as approach



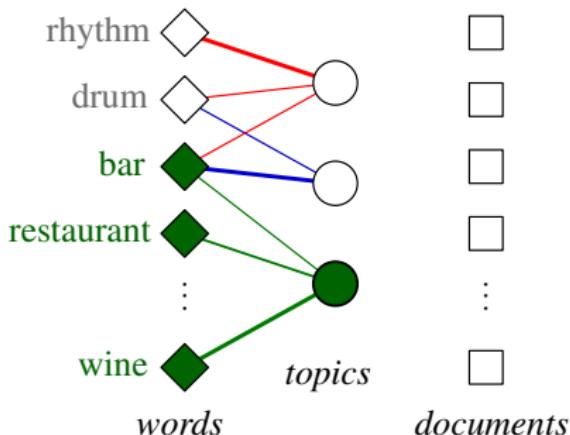
- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

# Topic models as approach



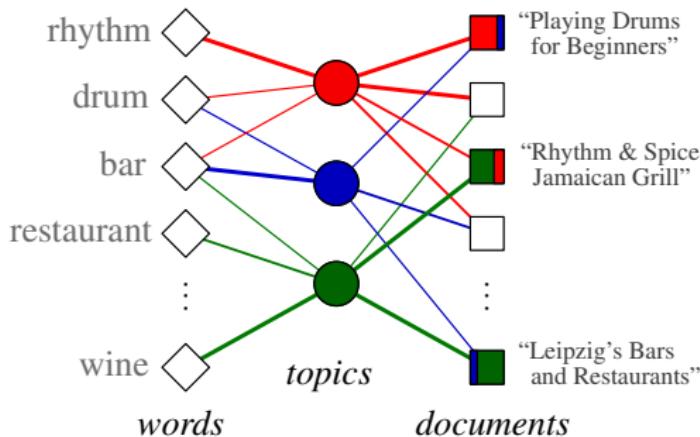
- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

# Topic models as approach



- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

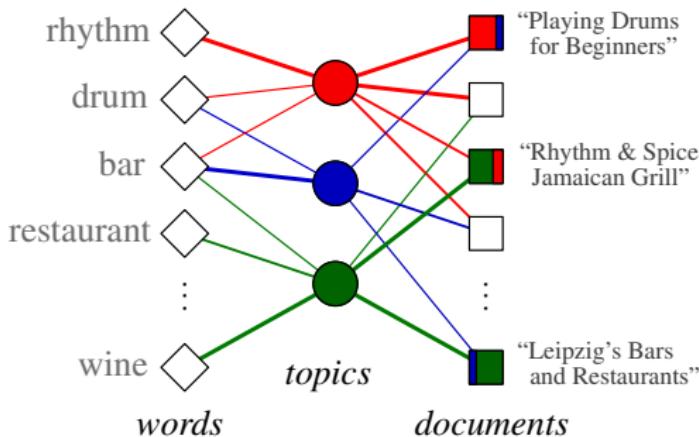
# Topic models as approach



- Probabilistic representations of grouped discrete data
- Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)

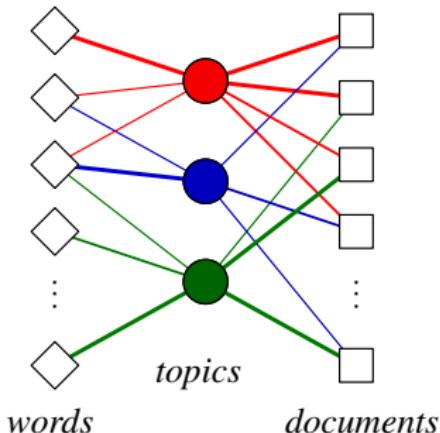
→ Reduce **vocabulary problem**: Find semantic relations  
→ Reduce **density problem**: Dimensionality reduction

# Topic models as approach



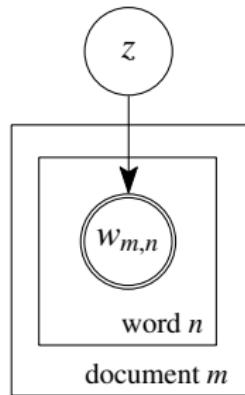
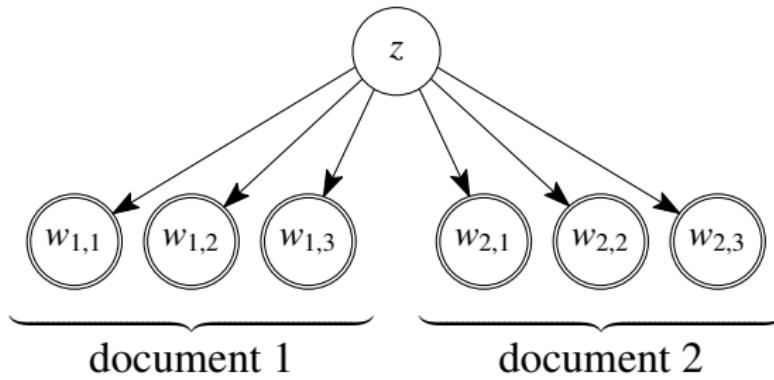
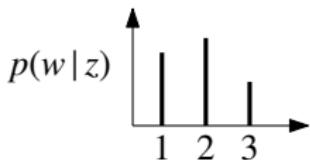
- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

# Topic models as approach



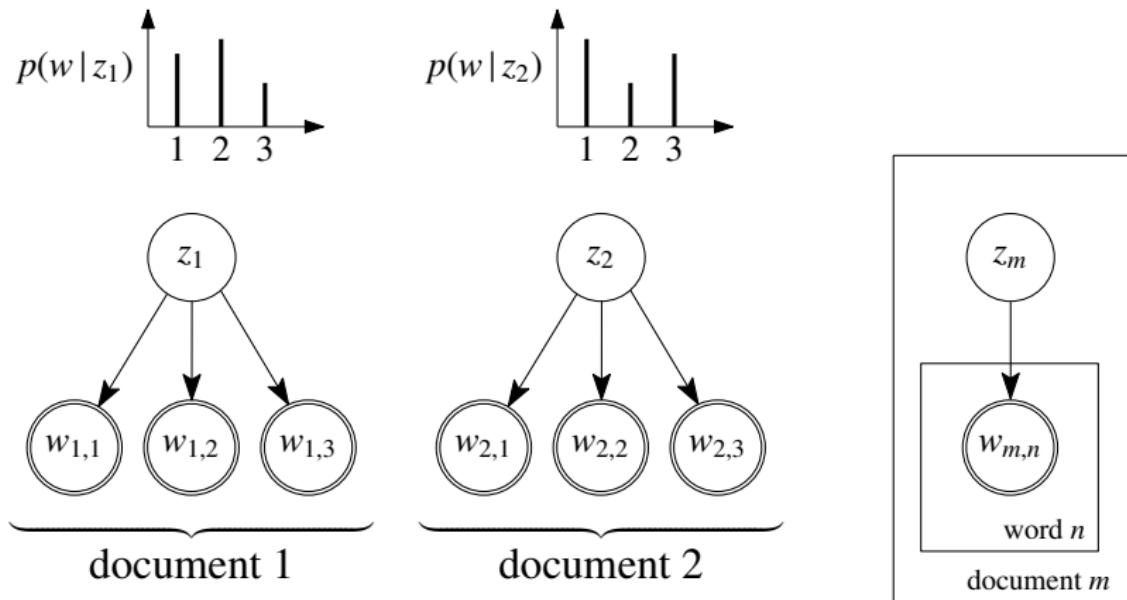
- Probabilistic representations of grouped discrete data
  - Illustrative for text: Words grouped in documents
    - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
    - Language = Mixture of topics (latent semantic structure)
- Reduce **vocabulary problem**: Find semantic relations
- Reduce **density problem**: Dimensionality reduction

# Language models: Unigram model



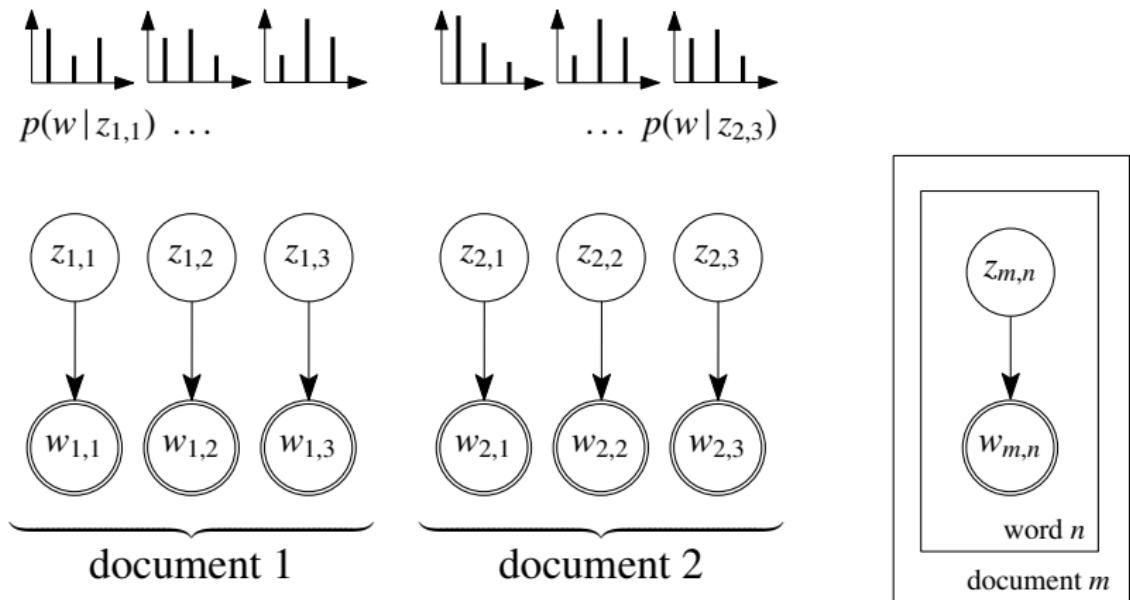
One distribution for all data

# Language models: Unigram mixture model



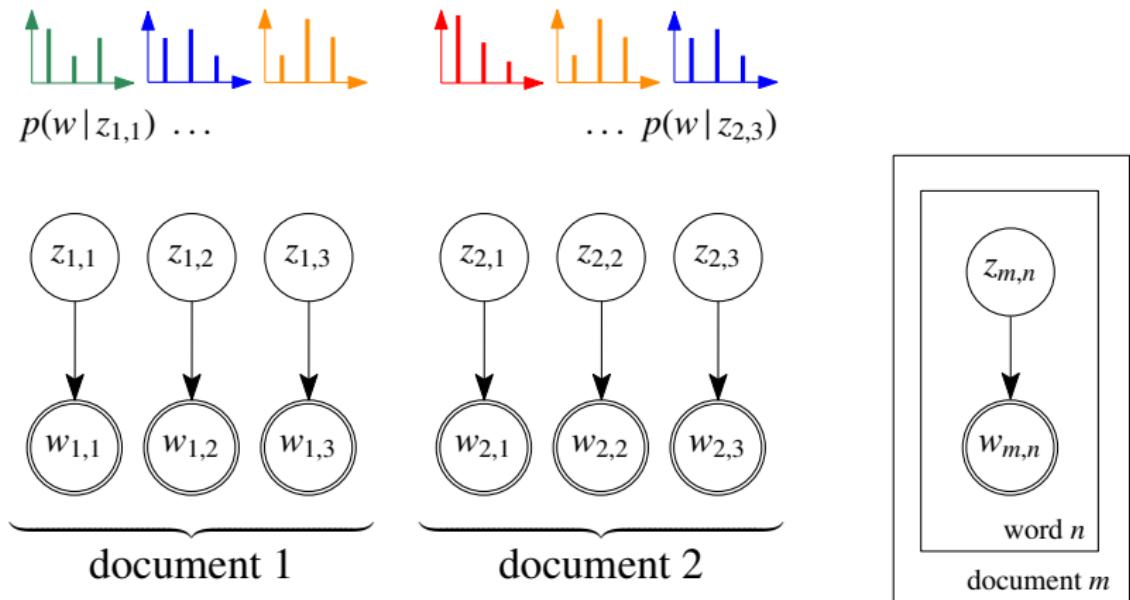
One distribution per document

# Language models: Unigram admixture model



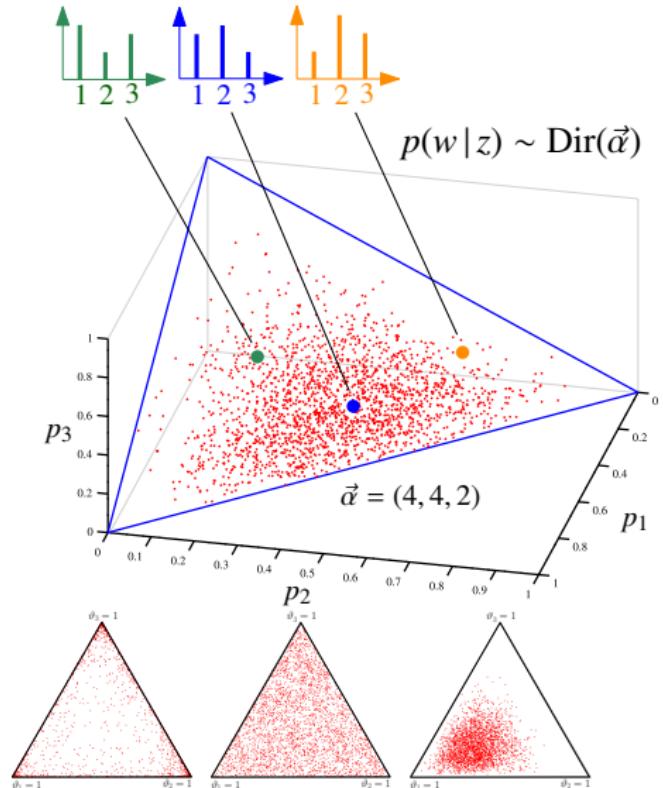
One distribution per word → basic topic model

# Language models: Unigram admixture model



One distribution per word → basic topic model

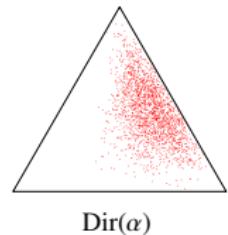
# Bayesian topic models: The Dirichlet distribution



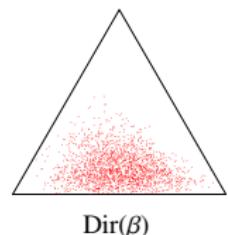
Bayesian methodology:

- Distributions generated from prior distributions
  - Speech + other discrete data: Dirichlet distribution important prior:
    - Defined on simplex: Surface containing all discrete distributions
    - Parameter  $\vec{\alpha}$  controls behaviour
- Bayesian topic model: Latent Dirichlet Allocation (LDA) (Blei et al. 2003)

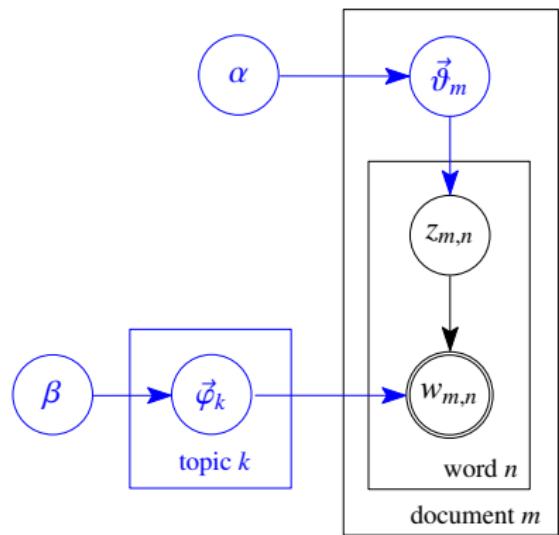
# Latent Dirichlet Allocation



Dir( $\alpha$ )



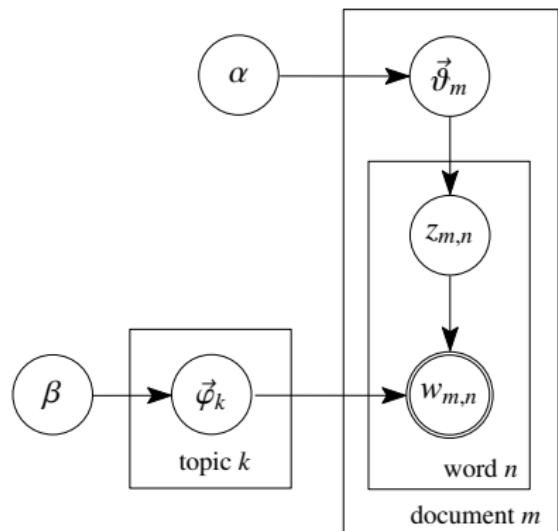
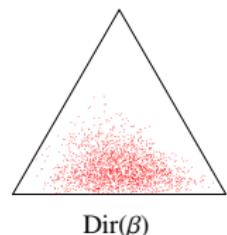
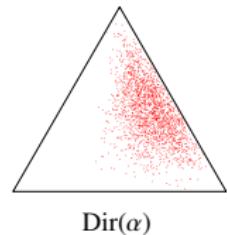
Dir( $\beta$ )



Latent Dirichlet Allocation (Blei et al. 2003)

# Latent Dirichlet Allocation

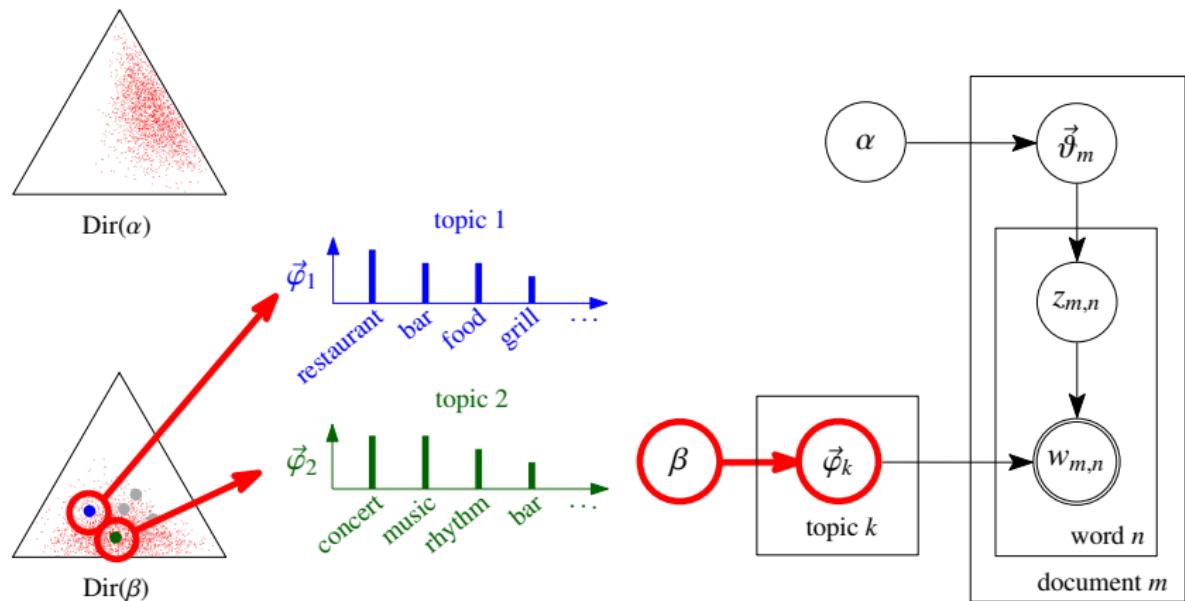
Concert tonight at Rhythm and Spice Restaurant ...



Latent Dirichlet Allocation (Blei et al. 2003)

# Latent Dirichlet Allocation

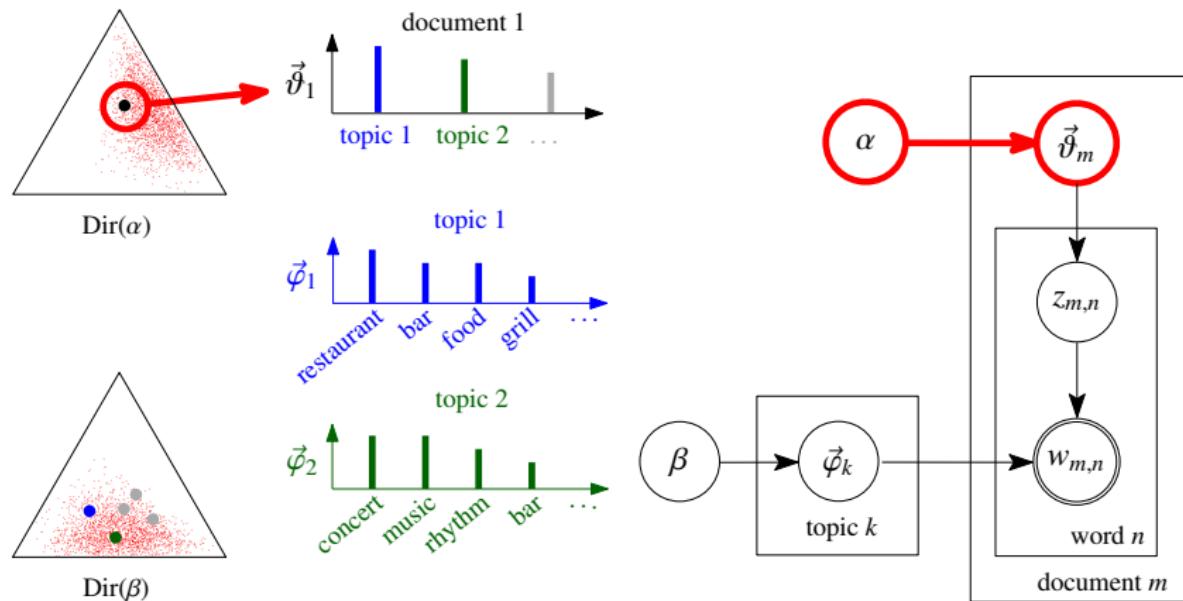
Concert tonight at Rhythm and Spice Restaurant ...



Generating word distributions for all topics

# Latent Dirichlet Allocation

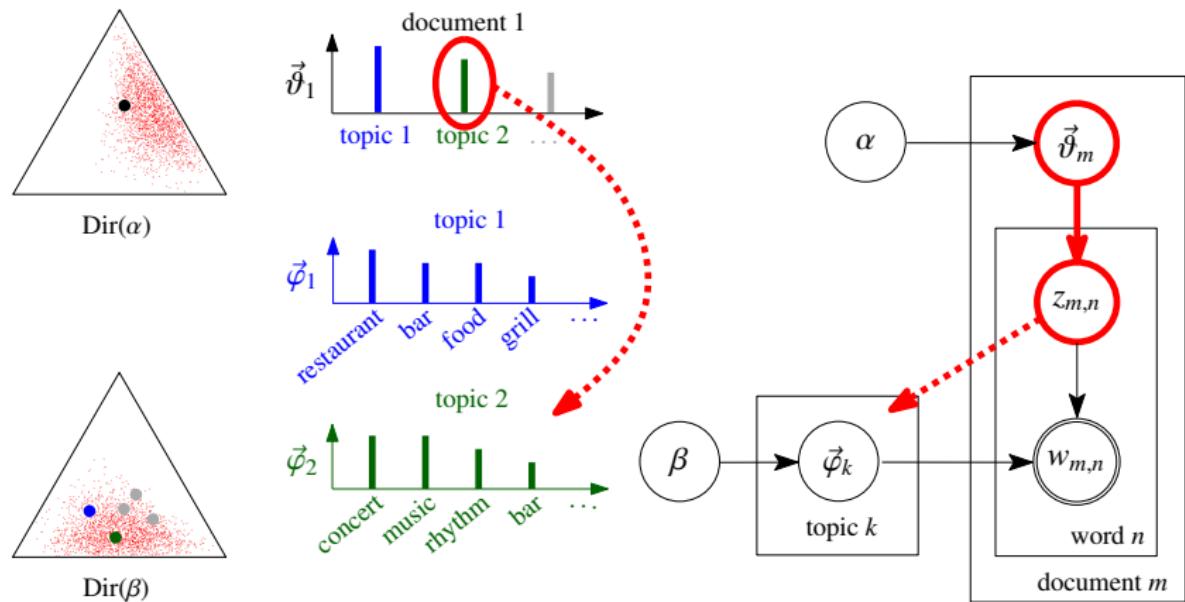
Concert tonight at Rhythm and Spice Restaurant ...



Generating topic distribution for document

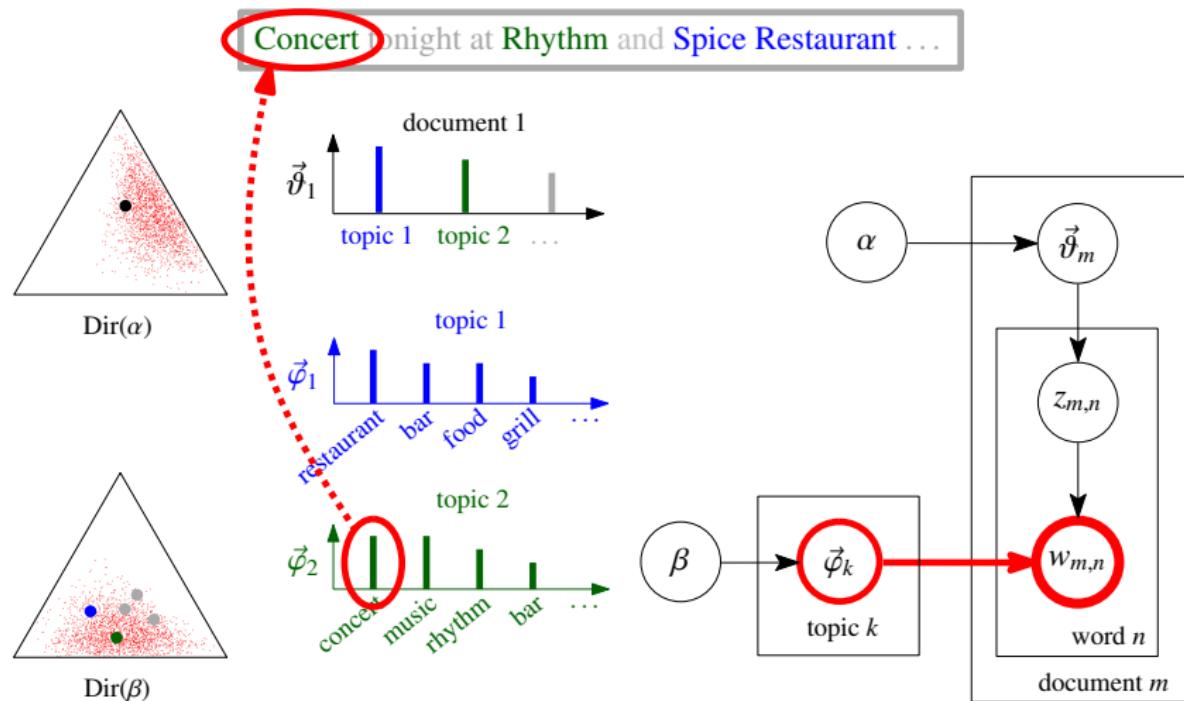
# Latent Dirichlet Allocation

Concert tonight at Rhythm and Spice Restaurant ...



Sampling the topic index for first word,  $z = 2$

# Latent Dirichlet Allocation



Sampling a word from term distribution for topic 2, "concert"

# State of the art

Large number of published models that extend LDA:

- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance

- But: No existing analysis as generic model class
- Partly tedious derivation, especially for inference algorithms

Conjecture:

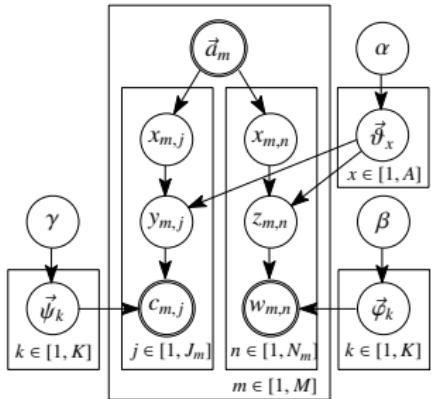
- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods

# State of the art

Large number of published models that extend LDA.

## Expert-tag-topic model

(Heinrich 2011)



$$\begin{aligned}
 p(\vec{\theta}, \vec{c}, \vec{d}, \vec{x}, \vec{z}, \vec{\varphi}, \vec{\psi} | \alpha, \beta, \gamma) &= p(\vec{\theta}|\vec{c}, \vec{\varphi}) p(\vec{c}|\vec{d}) \cdot p(\vec{x}|\vec{y}, \vec{y}) p(\vec{y}|\vec{w}) \\
 &\quad \cdot p(\vec{z}|\vec{y}, \vec{y}) p(\vec{y}|\vec{w}) p(\vec{w}|\vec{d}) \cdot p(\vec{x}|\vec{d}) \quad (E.1) \\
 &= \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{x_{m,n}}) p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,n,x_n} \right. \\
 &\quad \left. \cdot \prod_{j=1}^{J_m} p(c_{m,j} | \vec{\varphi}_{y_{m,j}}) p(y_{m,j} | \vec{\varphi}_{y_{m,j}}) a_{m,j,y_{m,j}} \right) \quad (E.2)
 \end{aligned}$$

$$\begin{aligned}
 p(\vec{\theta}, \vec{c}, \vec{d}, \vec{x}, \vec{z}, \vec{\varphi}, \vec{\psi} | \alpha, \beta, \gamma) &= \int \int \int \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{x_{m,n}}) p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,n,x_n} \right) a_{m,n,x_n} \\
 &\quad \cdot \prod_{j=1}^{J_m} p(c_{m,j} | \vec{\varphi}_{y_{m,j}}) p(y_{m,j} | \vec{\varphi}_{y_{m,j}}) a_{m,j,y_{m,j}} \quad (E.3)
 \end{aligned}$$

$$\begin{aligned}
 &\cdot d\mu(\vec{\varphi}|\alpha) \cdot d\mu(\vec{\varphi}|\beta) \cdot d\mu(\vec{\varphi}|\gamma) \\
 &= \int \prod_{m=1}^M \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{x_{m,n}}) \prod_{k=1}^K p(\vec{\varphi}_k | \beta) d\vec{\varphi}_k \\
 &\quad \cdot \int \prod_{m=1}^M \prod_{n=1}^{N_m} p(c_{m,j} | \vec{\varphi}_{y_{m,j}}) \prod_{k=1}^K p(\vec{\varphi}_k | \gamma) d\vec{\varphi}_k \\
 &\quad \cdot \int \prod_{m=1}^M p(\vec{C}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,n,x_n} \prod_{j=1}^{J_m} p(y_{m,j} | \vec{\theta}_{x_{m,n}}) a_{m,j,y_{m,j}} d\vec{\theta}_m \quad (E.4)
 \end{aligned}$$

$$\begin{aligned}
 &= \int \prod_{k=1}^K \frac{1}{\Delta_V(\beta)} \prod_{l=1}^V \theta_{k,l}^{n_{k,l}-\beta-1} d\vec{\varphi}_k \cdot \int \prod_{k=1}^K \frac{1}{\Delta_C(\gamma)} \prod_{c=1}^C \theta_{k,c}^{n_{k,c}+\gamma-1} d\vec{\varphi}_k \\
 &\quad \cdot \int \prod_{a=1}^A \frac{1}{\Delta_C(\alpha)} \prod_{k=1}^K \theta_{k,k}^{n_{k,k}^{(2)}+n_{k,k}^{(1)}-\alpha-1} d\vec{\theta}_a \cdot \prod_{m=1}^M \prod_{a=1}^A \frac{n_{m,a}^{(2)}+n_{m,a}^{(1)}}{\Delta_K(\alpha)} \quad (E.5) \\
 &= \prod_{k=1}^K \frac{\Delta(n_k^{(2)} + \beta)}{\Delta_V(\beta)} \cdot \frac{\Delta(n_k^{(3)} + \gamma)}{\Delta_C(\gamma)} \prod_{a=1}^A \frac{\Delta(n_a^{(2)} + n_a^{(1)} + \alpha)}{\Delta_K(\alpha)} \prod_{m=1}^M \frac{n_{m,a}^{(2)}+n_{m,a}^{(1)}}{\Delta_{K,a}^{(2)}+n_{m,a}^{(1)}}. \quad (E.6)
 \end{aligned}$$

$$\begin{aligned}
 p(z_i=k, x_i=x_i | \vec{v}_{-i}, \vec{y}_{-i}, \vec{x}_{-i}, \vec{w}_{-i}, \vec{d}, \vec{\vartheta}) &= \frac{p(\vec{w}| \vec{v}_{-i}, \vec{y}_{-i}, \vec{x}_{-i})}{p(\vec{w}| \vec{v}_{-i}, \vec{y}_{-i}, \vec{x}_{-i})} = \frac{p(\vec{w}| \vec{v}_{-i})}{p(\vec{w}| \vec{v}_{-i})} \cdot \frac{p(\vec{y}| \vec{v}_{-i})}{p(\vec{y}| \vec{v}_{-i})} \quad (E.7) \\
 &\propto \frac{\Delta(\vec{n}_{k,-i}^{(2)} + \beta)}{\Delta(\vec{n}_{k,-i}^{(2)} + \beta)} \cdot \frac{\Delta(\vec{n}_{k,-i}^{(3)} + \alpha)}{\Delta(\vec{n}_{k,-i}^{(3)} + \alpha)} \cdot a_{m,i} \quad (E.8)
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\Gamma(n_{k,-i}^{(2)} + \beta) \Gamma(n_{k,-i}^{(3)} + V\beta)}{\Gamma(n_{k,-i}^{(2)} + \beta) \Gamma(n_{k,-i}^{(3)} + V\beta)} \cdot \frac{\Gamma(n_{k,-i}^{(2)} + \alpha) \Gamma(n_{k,-i}^{(3)} + K\alpha)}{\Gamma(n_{k,-i}^{(2)} + \alpha) \Gamma(n_{k,-i}^{(3)} + K\alpha)} \cdot a_{m,i} \quad (E.9) \\
 &= \frac{n_{k,-i}^{(2)} + \beta}{n_{k,-i}^{(2)} + V\beta} \cdot \frac{n_{k,-i}^{(3)} + \alpha}{n_{k,-i}^{(3)} + K\alpha} \cdot a_{m,i} \quad (E.10)
 \end{aligned}$$

$$p(y_j=k, x_j=x_j | \vec{v}_{-j}, \vec{y}_{-j}, \vec{x}_{-j}, \vec{w}, \vec{d}, \vec{a}_{-j}) \propto \frac{n_{k,x,-j}^{(2)} + \gamma}{n_{k,x,-j}^{(2)} + V\gamma} \cdot \frac{n_{k,x,-j}^{(3)} + \alpha}{n_{k,x,-j}^{(3)} + K\alpha} \cdot a_{m,x} \quad (E.12)$$

Con-

algorithms and design methods

# State of the art

Large number of published models that extend LDA:

- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance

- But: No existing analysis as generic model class
- Partly tedious derivation, especially for inference algorithms

Conjecture:

- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods

# State of the art

Large number of published models that extend LDA:

- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance

- But: No existing analysis as generic model class
- Partly tedious derivation, especially for inference algorithms

Conjecture:

- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods

# Research questions

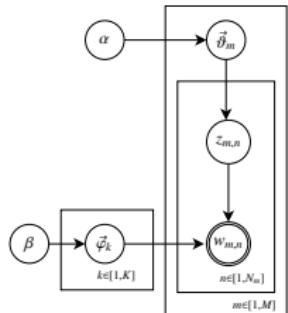
- “How can topic models be described in a generic way in order to use their properties across particular applications?”
- “Can generic topic models be implemented generically and, if so, can repeated structures be exploited for optimisations?”
- “How can generic models be applied to data in virtual communities?”

# Overview

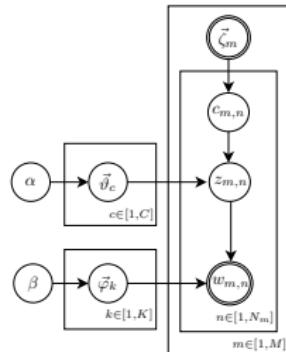
- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook

“How can topic models be described in a generic way in order to use their properties across particular applications?”

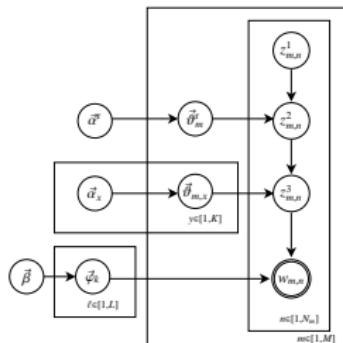
# Topic models: Example structures



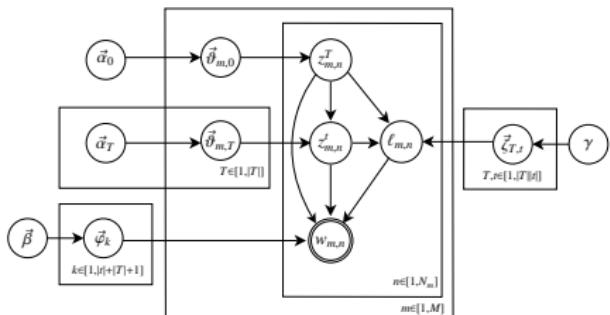
(a) Latent Dirichlet allocation (LDA)



(b) Author–topic model (ATM)



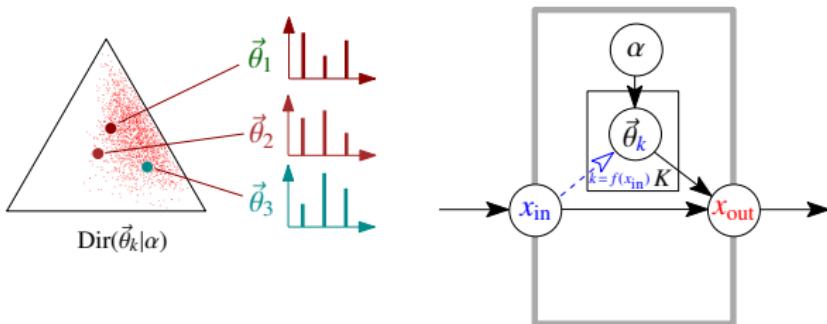
(c) Pachinko allocation model (PAM4)



(d) Hierarchical PAM (hPAM)

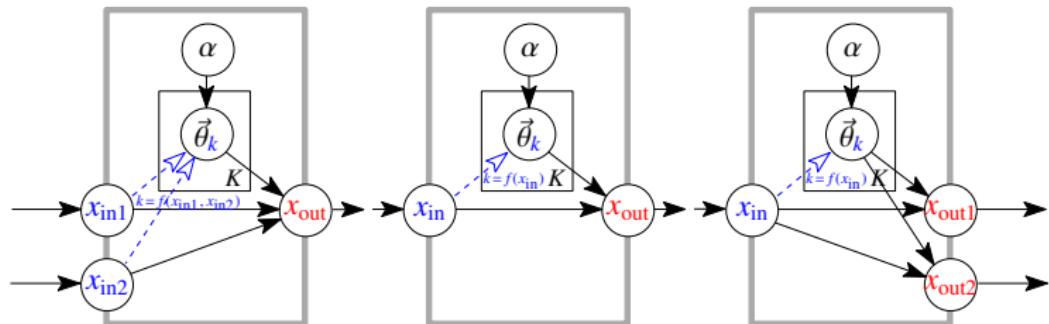
(Blei et al. 2003; Rosen-Zvi et al. 2004; Li and McCallum 2006; Li et al. 2007)

# Generic topic models – “NoMMs”



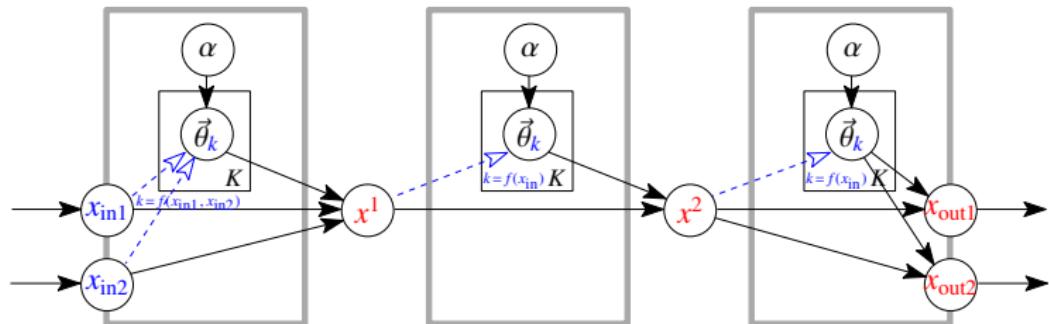
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



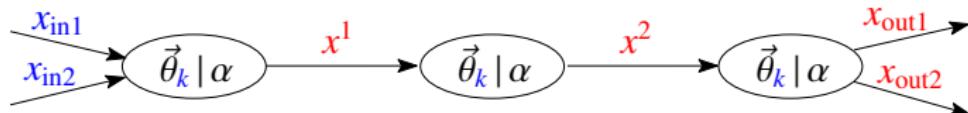
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



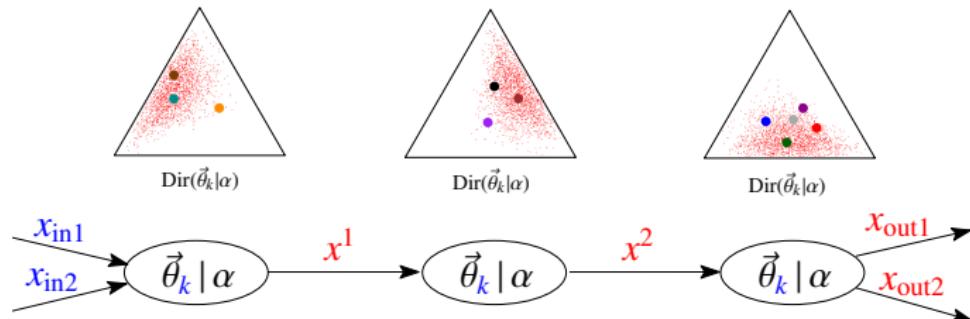
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



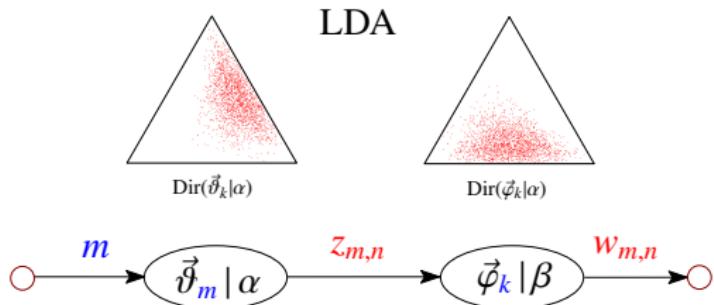
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



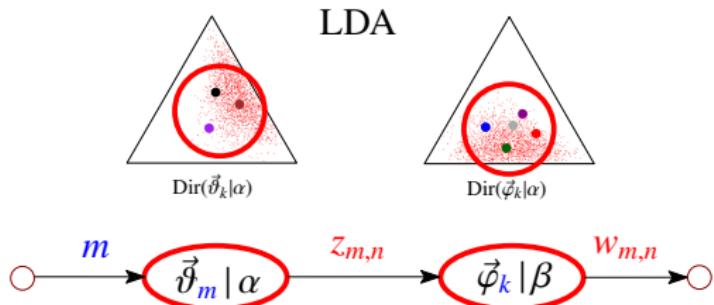
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



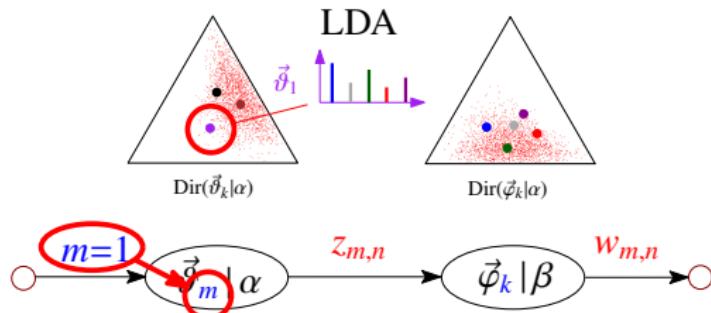
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



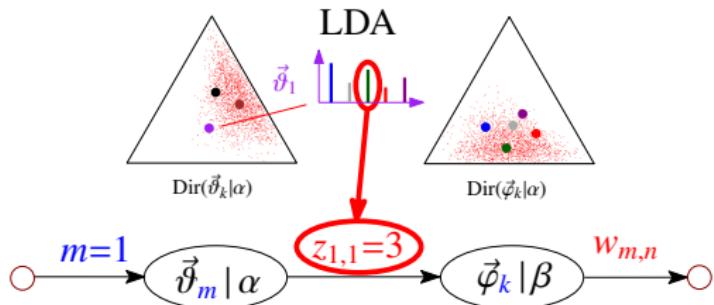
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



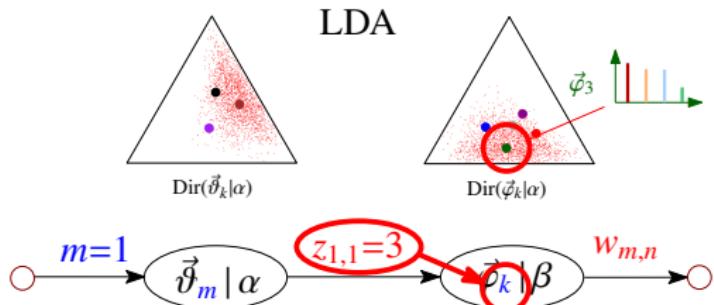
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



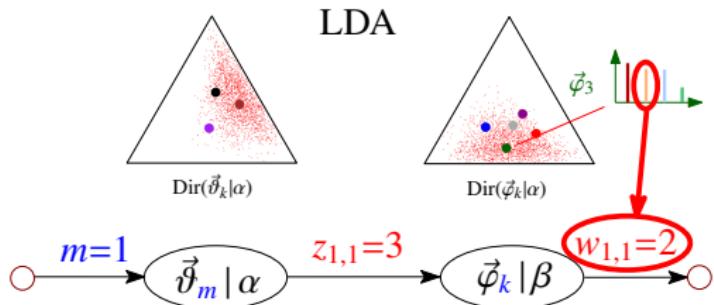
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



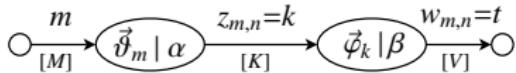
- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Generic topic models – “NoMMs”



- Generic characteristics of topic models:
  - Levels with discrete components  $\vec{\theta}_k$ , generated from Dirichlet distributions
  - Coupling via values of discrete variables  $x$
- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.

# Topic models as NoMMs



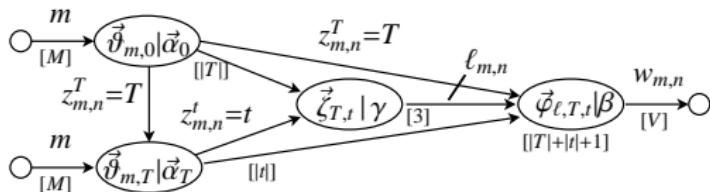
(a) Latent Dirichlet allocation, LDA



(b) Author–topic model, ATM

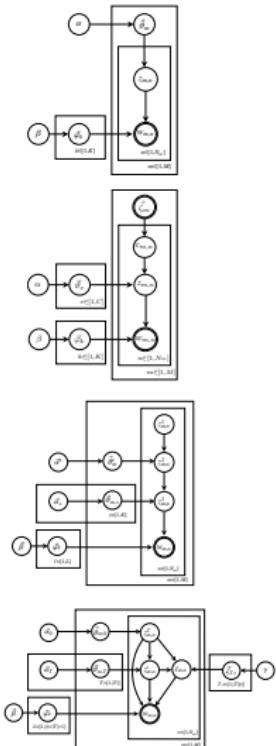


(c) Pachinko allocation model, PAM



(d) Hierarchical pachinko allocation model, hPAM

(Blei et al. 2003; Rosen-Zvi et al. 2004; Li and McCallum 2006; Li et al. 2007)



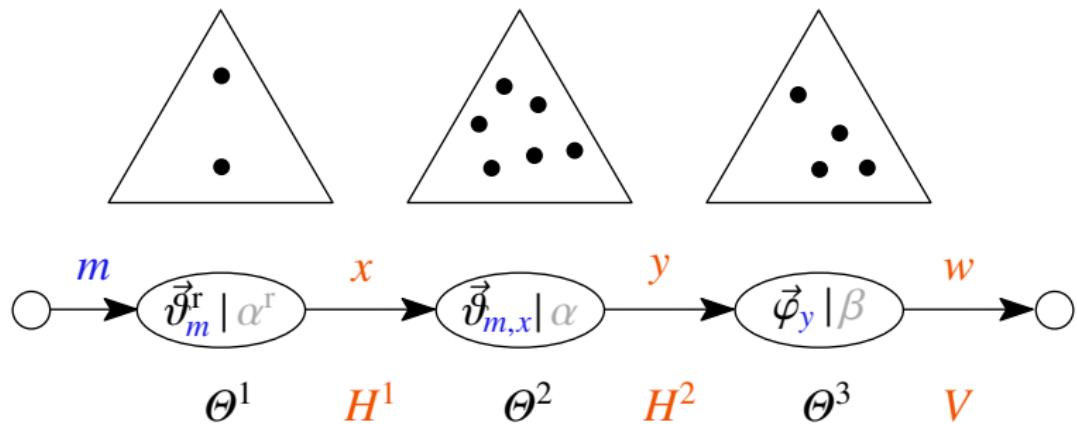
# Overview

- Introduction
- Generic topic models
- **Inference methods**
- Application to virtual communities
- Conclusions and outlook

“Can generic topic models be implemented generically...?”

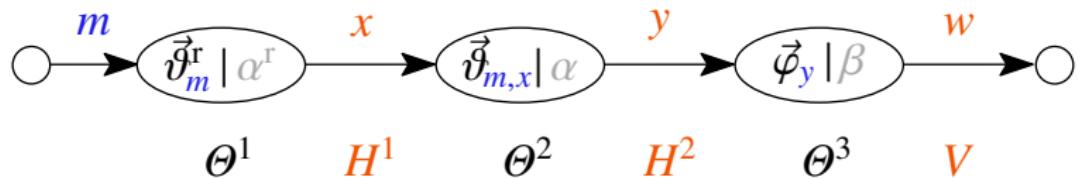
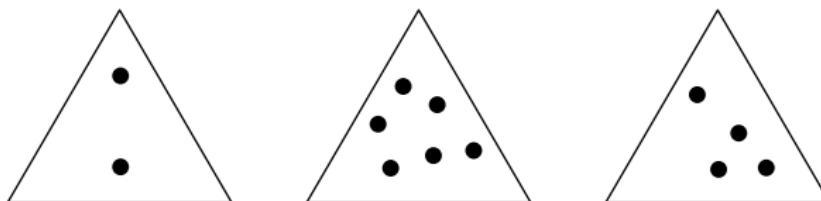
# Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters  $\Theta$  and latent variables/topics  $H$ , given observations  $V$  and Dirichlet parameters  $A$
  - = Determine posterior distribution  $p(H, \Theta | V, A)$
- Intractability  $\rightarrow$  approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters  $\Theta$  (“Collapsed” GS)
  - Sample topics  $H_i$  for each data point  $i$  in turn:  $H_i \sim p(H_i | H_{\neg i}, V, A)$



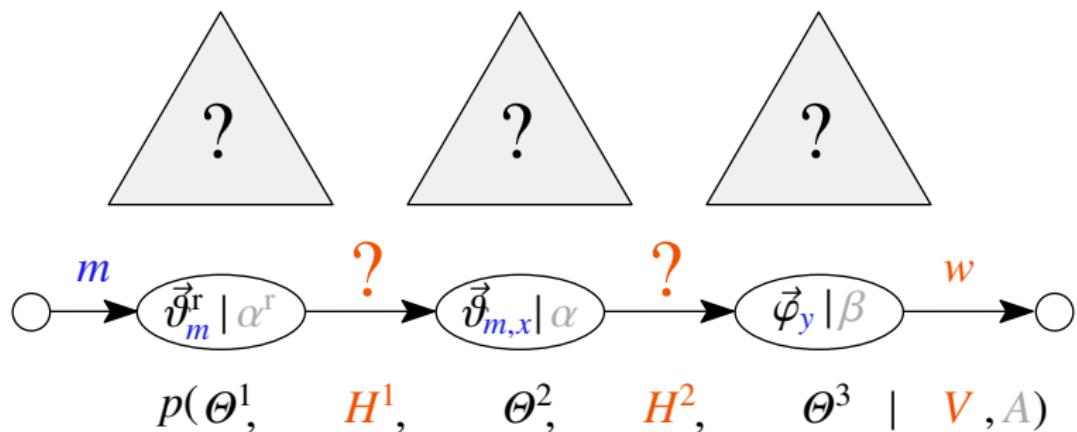
# Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters  $\Theta$  and latent variables/topics  $H$ , given observations  $V$  and Dirichlet parameters  $A$ 
    - = Determine posterior distribution  $p(H, \Theta | V, A)$
- Intractability  $\rightarrow$  approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters  $\Theta$  (“Collapsed” GS)
  - Sample topics  $H_i$  for each data point  $i$  in turn:  $H_i \sim p(H_i | H_{\neg i}, V, A)$



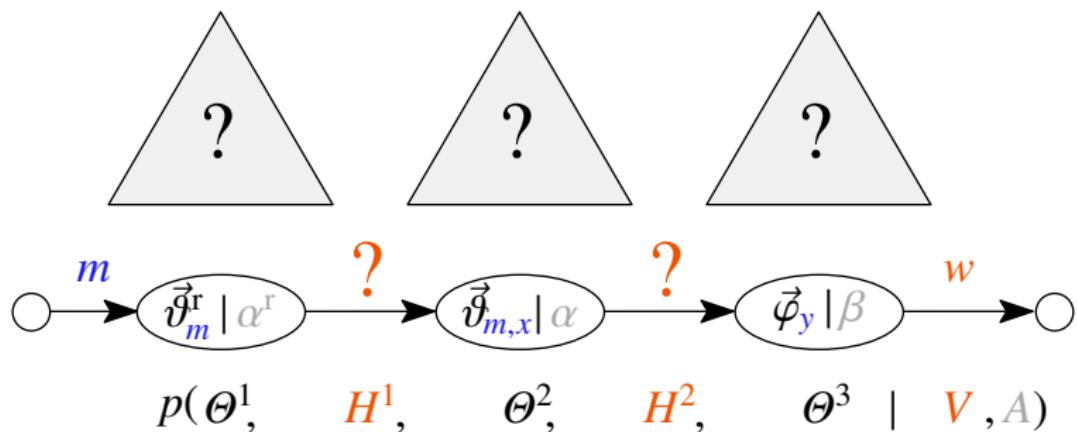
# Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters  $\Theta$  and latent variables/topics  $H$ , given observations  $V$  and Dirichlet parameters  $A$
  - = Determine posterior distribution  $p(H, \Theta | V, A)$
- Intractability → approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters  $\Theta$  (“Collapsed” GS)
  - Sample topics  $H_i$  for each data point  $i$  in turn:  $H_i \sim p(H_i | H_{\neg i}, V, A)$



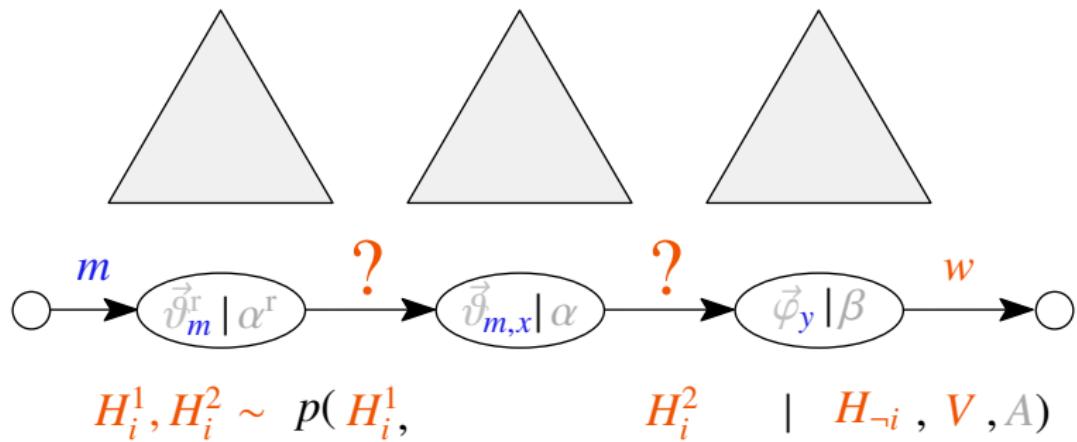
# Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters  $\Theta$  and latent variables/topics  $H$ , given observations  $V$  and Dirichlet parameters  $A$
  - = Determine posterior distribution  $p(H, \Theta | V, A)$
- Intractability → approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters  $\Theta$  (“Collapsed” GS)
  - Sample topics  $H_i$  for each data point  $i$  in turn:  $H_i \sim p(H_i | H_{-i}, V, A)$



# Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters  $\Theta$  and latent variables/topics  $H$ , given observations  $V$  and Dirichlet parameters  $A$
  - = Determine posterior distribution  $p(H, \Theta | V, A)$
- Intractability → approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters  $\Theta$  (“Collapsed” GS)
  - Sample topics  $H_i$  for each data point  $i$  in turn:  $H_i \sim p(H_i | H_{\neg i}, V, A)$



# Sampling distribution for NoMMs

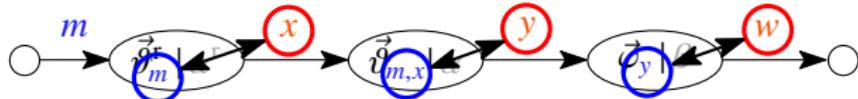


- Gibbs sampler can be generically derived (Heinrich 2009)
- Typical case: Quotients of factors over levels  $\ell$ :

$$p(H_i|H_{\neg i}, V, A) \propto \prod_{\ell} \left[ \frac{n_{k,t}^{-i} + \alpha}{\sum_t n_{k,t}^{-i} + \alpha} \right]^{[\ell]}$$

- $n_{k,t}$  = count of co-occurrences between **input** and **output** values of a level (**components** and **samples**)
- More complex variants covered by  $q(k, t) \triangleq \frac{\text{beta}(\{n_{k,t}\}_{t=1}^T + \alpha)}{\text{beta}(\{n_{k,t}^{-i}\}_{t=1}^T + \alpha)}$

# Sampling distribution for NoMMs



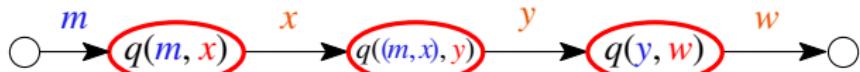
- Gibbs sampler can be generically derived (Heinrich 2009)
- Typical case: Quotients of factors over levels  $\ell$ :

$$p(\mathbf{H}_i | H_{\neg i}, V, A) \propto \prod_{\ell} \left[ \frac{n_{k,t}^{\neg i} + \alpha}{\sum_t n_{k,t}^{\neg i} + \alpha} \right]^{\ell}$$

- $n_{k,t}$  = count of co-occurrences between **input** and **output** values of a level (**components** and **samples**)

- More complex variants covered by  $q(k, t) \triangleq \frac{\text{beta}(\{n_{k,t}\}_{t=1}^T + \alpha)}{\text{beta}(\{n_{k,t}^{\neg i}\}_{t=1}^T + \alpha)}$

# Sampling distribution for NoMMs

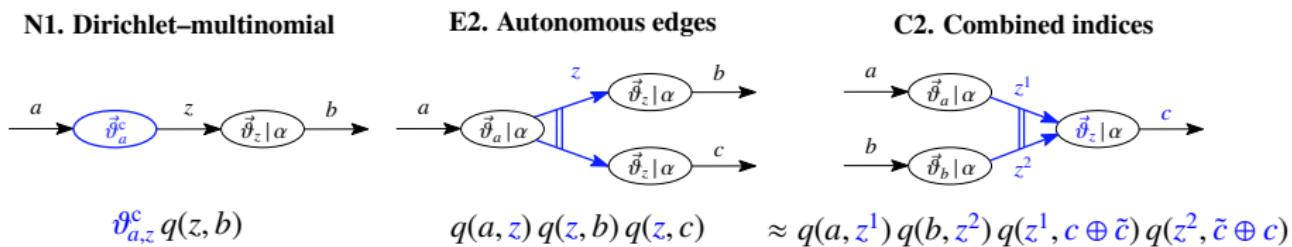
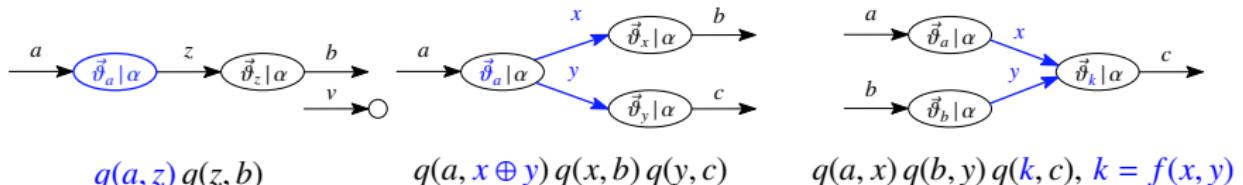


- Gibbs sampler can be generically derived (Heinrich 2009)
- Typical case: Quotients of factors over levels  $\ell$ :

$$p(H_i | H_{\neg i}, V, A) \propto \prod_{\ell} \left[ \frac{n_{k,t}^{-i} + \alpha}{\sum_t n_{k,t}^{-i} + \alpha} \right]^{[\ell]} = \prod_{\ell} [q(k, t)]^{[\ell]}$$

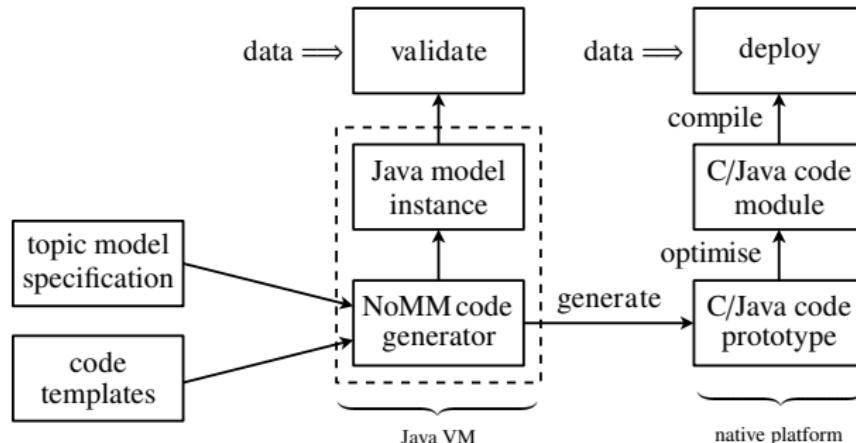
- $n_{k,t}$  = count of co-occurrences between **input** and **output** values of a level (**components** and **samples**)
- More complex variants covered by  $q(k, t) \triangleq \frac{\text{beta}(\{n_{k,t}\}_{t=1}^T + \alpha)}{\text{beta}(\{n_{k,t}^{-i}\}_{t=1}^T + \alpha)}$

# Typology of NoMM substructures



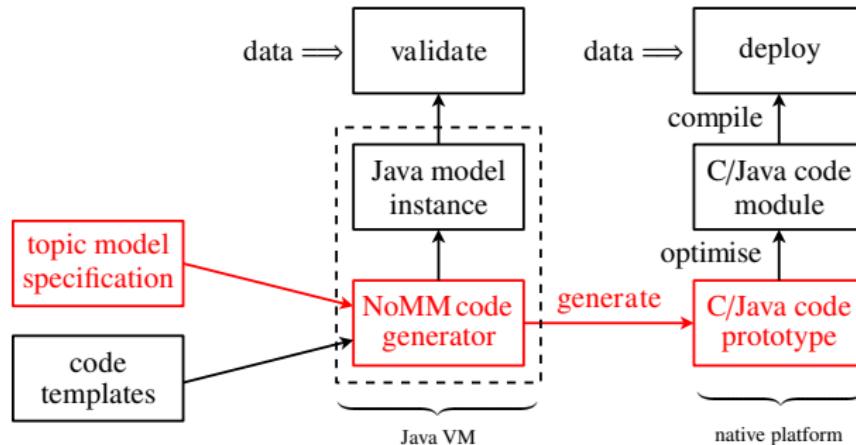
- NoMM substructures: Nodes, edges/branches, component indices/merging of edges:
  - Representation via  $q$ -functions and likelihood
  - Multiple samples per data point:  $q(a, x \oplus y)$  for respective level
- “Library” incl. additional structures: alternative distributions, regression, aggregation etc.  $\leftrightarrow q$ -functions + other factors

# Implementation: Gibbs “meta-sampler”



- Code generator for topic models in Java and C
- Separation of knowledge domains: topic model applications vs. machine learning vs. computing architecture

# Implementation: Gibbs “meta-sampler”

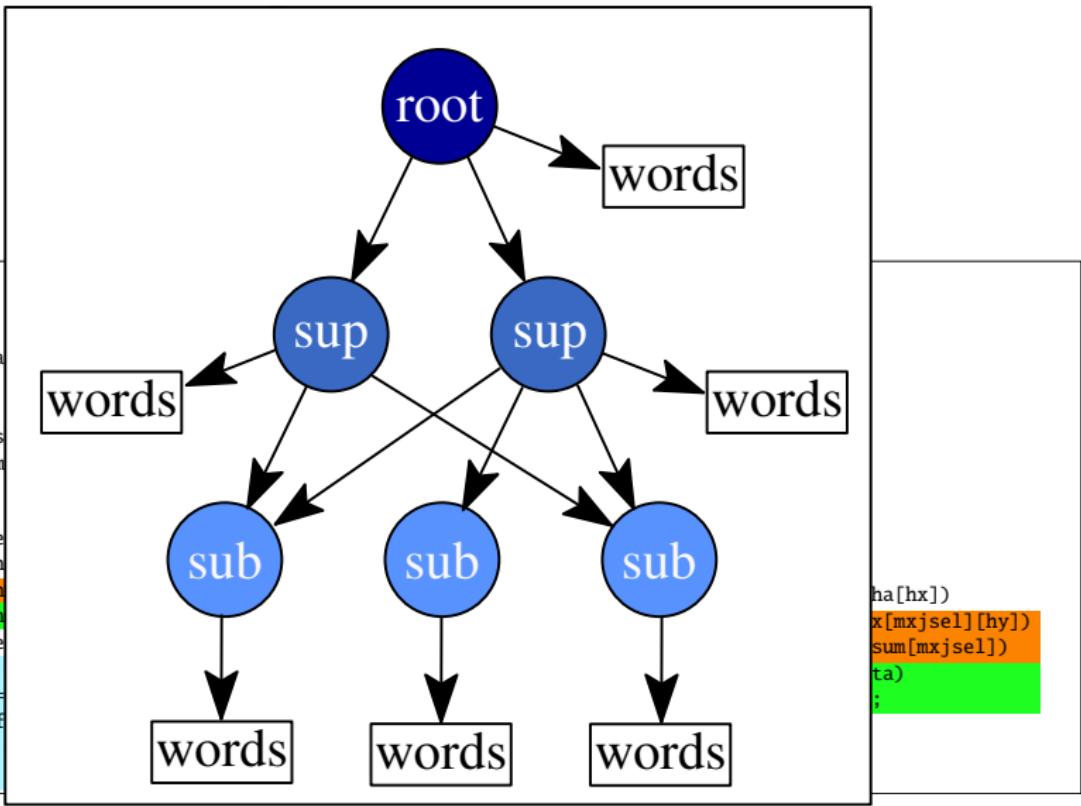


- Code generator for topic models in Java and C
- Separation of knowledge domains: topic model applications vs. machine learning vs. computing architecture

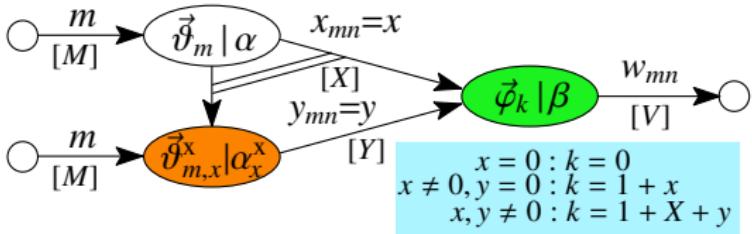
# Example NoMM script and generated kernel: hPAM2

```
model = HPAM2

description:
    Hierarchica
sequences:
    # variables
    w, x, y : m
network:
    # each line
    m >> th
    m,x >> th
    x,y >> ph
    # java code
    k : {
        if (x ==
        else if
        else k
    }.
```



# Example NoMM script and generated kernel: hPAM2



```

model = HPAM2

description:
    Hierarchical PAM model 2 (HPAM2)

sequences:
    # variables sampled for each (m,n)
    w, x, y : m, n

network:
    # each line one NoMM node
    m >> theta | alpha >> x
    m,x >> thetax | alphax[x] >> y
    x,y >> phi[k] >> w
    # java code to assign k
    k : {
        if (x==0) { k = 0; }
        else if (y==0) k = 1 + x;
        else k = 1 + X + y;
    }.
```



```

// hidden edge
for (hx = 0; hx < X; hx++) {
    // hidden edge
    for (hy = 0; hy < Y; hy++) {
        mxsel = X * m + hx;
        mxjsel = hx;
        if (hx == 0)
            ksel = 0;
        else if (hy == 0)
            ksel = 1 + hx;
        else
            ksel = 1 + X + hy;
        pp[hx][hy] = (nmx[m][hx] + alpha[hx])
                    * (nmxy[mxsel][hy] + alphax[mxjsel][hy])
                    / (nmxysum[mxsel] + alphaxsum[mxjsel])
                    * (nkw[ksel][w[m][n]] + beta)
                    / (nkwsun[ksel] + betasum);
        psum += pp[hx][hy];
    } // for h
} // for h
```

# Document–Topic distribution in Gibbs sampler

Iteration 1

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

# Document–Topic distribution in Gibbs sampler

Iteration 5

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

# Document–Topic distribution in Gibbs sampler

Iteration 10

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

# Document–Topic distribution in Gibbs sampler



Iteration 15

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

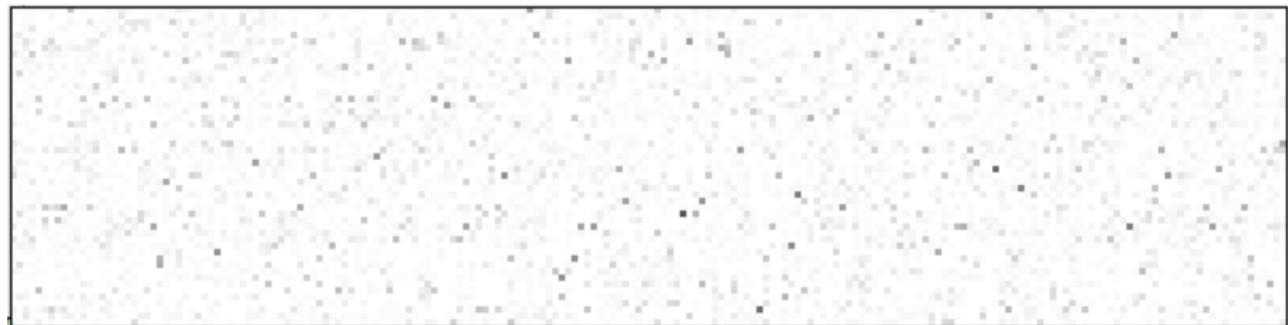
# Document–Topic distribution in Gibbs sampler



Iteration 20

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

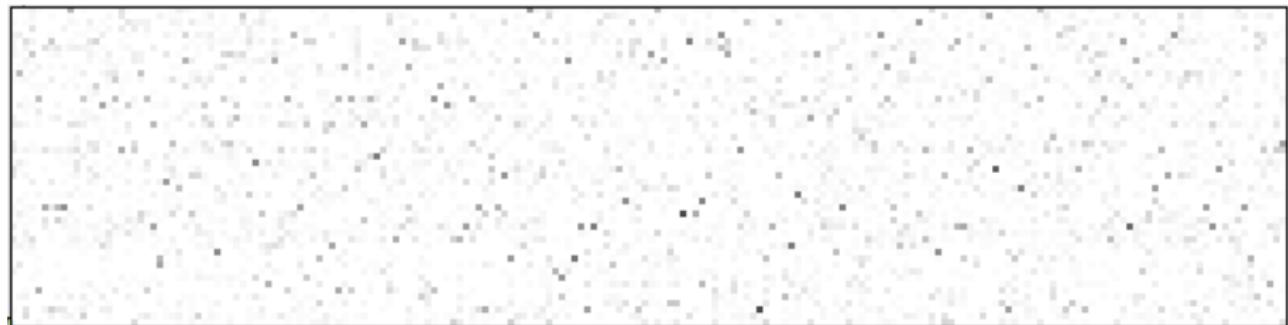
# Document–Topic distribution in Gibbs sampler



Iteration 30

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

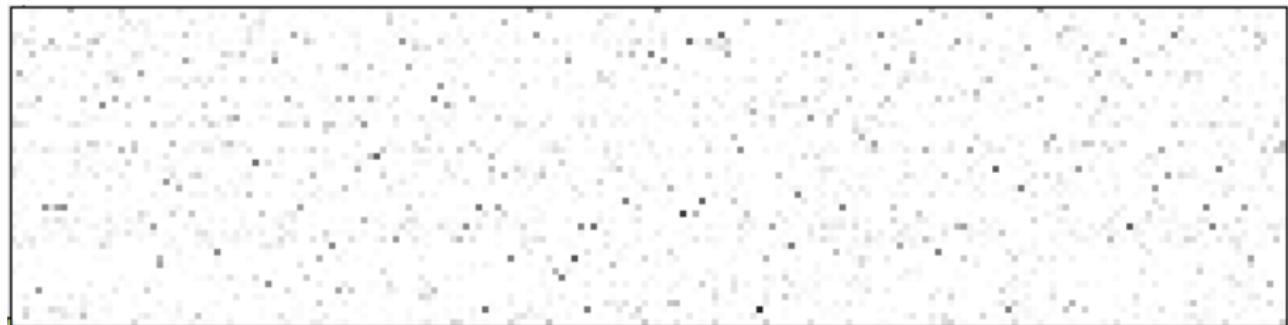
# Document–Topic distribution in Gibbs sampler



Iteration 40

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

# Document–Topic distribution in Gibbs sampler



Iteration 50

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

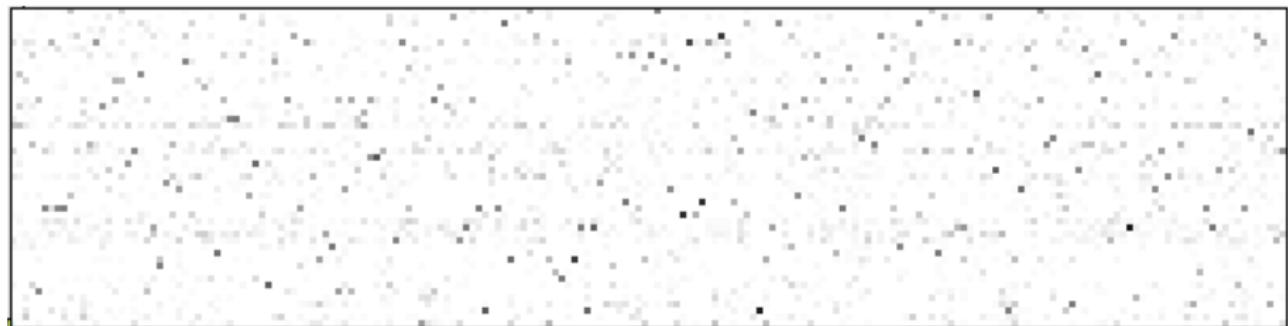
# Document–Topic distribution in Gibbs sampler



Iteration 60

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

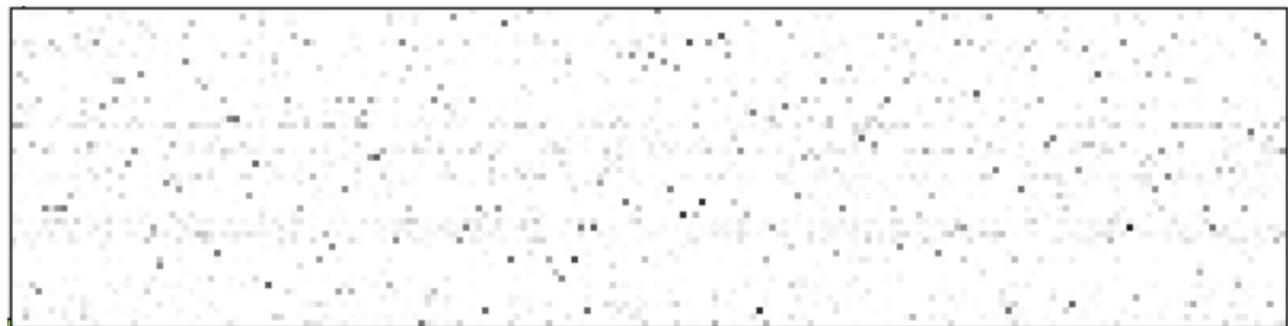
# Document–Topic distribution in Gibbs sampler



Iteration 80

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

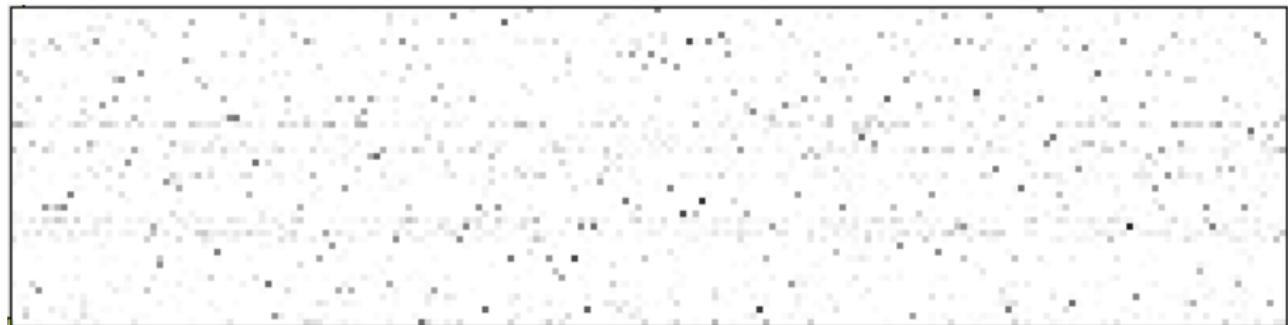
# Document–Topic distribution in Gibbs sampler



Iteration 100

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

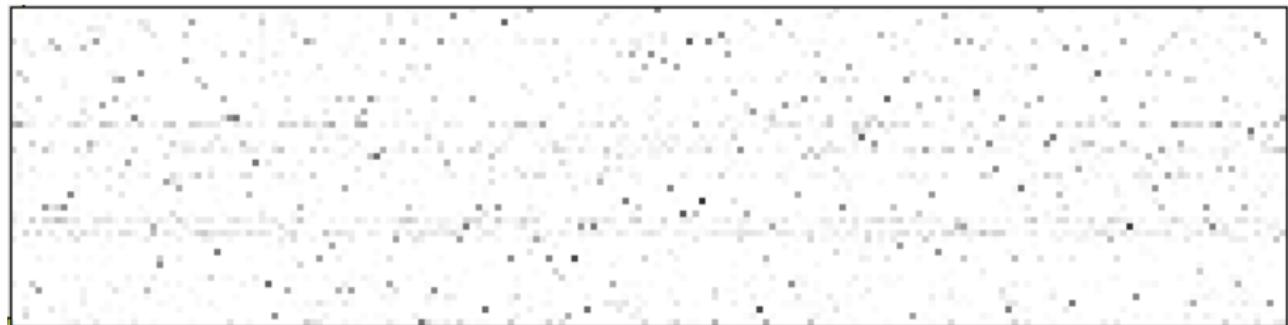
# Document–Topic distribution in Gibbs sampler



Iteration 120

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

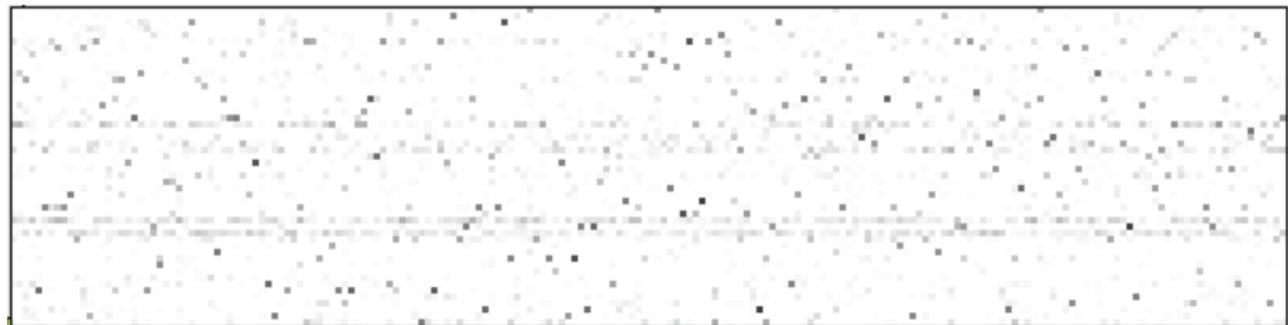
# Document–Topic distribution in Gibbs sampler



Iteration 150

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

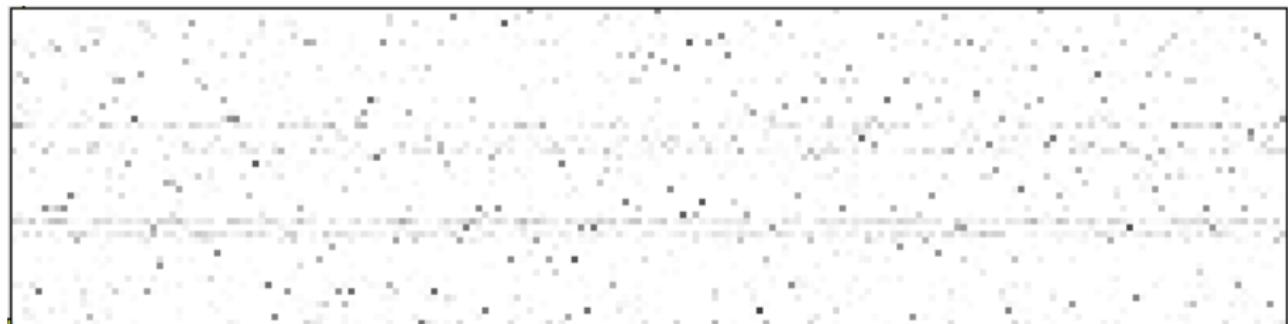
# Document–Topic distribution in Gibbs sampler



Iteration 200

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

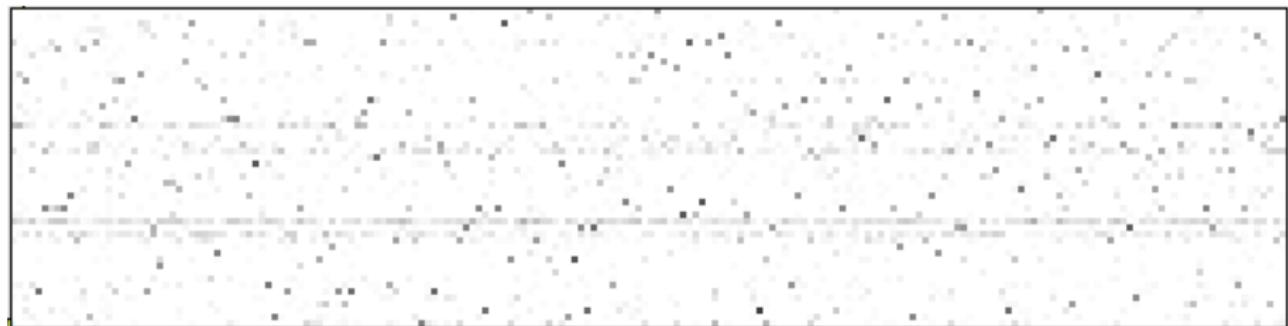
# Document–Topic distribution in Gibbs sampler



Iteration 300

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

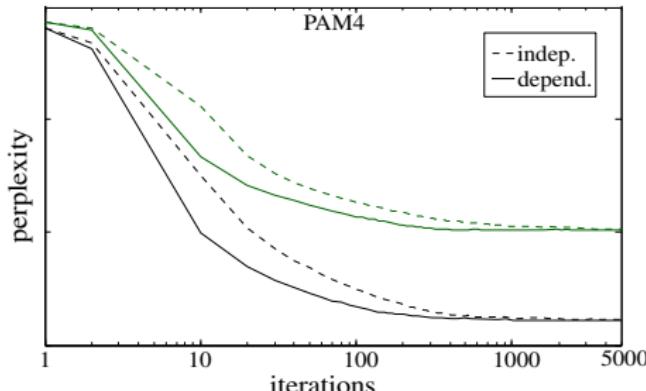
# Document–Topic distribution in Gibbs sampler



Iteration 500, converged  $\leftrightarrow$  stationary state

Document–topic matrix  $\vartheta$  (200 documents, 50 topics)

# Fast sampling: Hybrid scaling methods



model	dim.	ser.	par.	indep.	speedup
LDA	500	✓	✓	—	30.2
PAM4	$40 \times 40$			✓	7.4
PAM4	$40 \times 40$			✓	24.1
PAM4	$40 \times 40$	✓	✓	✓	49.8

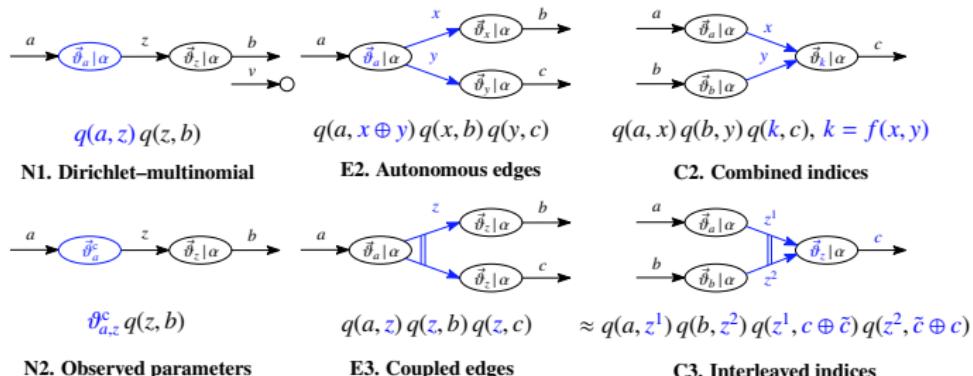
- Serial and parallel scaling methods:
  - Generalised results for LDA to generic NoMMs, specifically (Porteous et al. 2008; Newman et al. 2009) + novel approach
- Problem: Sampling space for stat. dependent variables:  $K \times L \times \dots$ 
  - Independence assumption: Separate samplers with dimensions  $K + L + \dots \ll K \times L \times \dots$
  - Empirical result: Iterations  $\uparrow$ , but topic quality comparable
- Hybrid approaches with independent samplers highly effective
  - Implementation: complexity covered by meta-sampler

# Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook

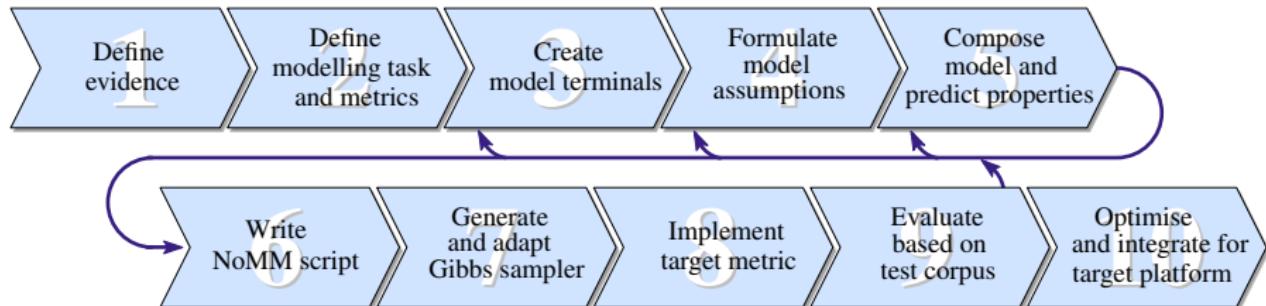
“How can generic models be applied to data in virtual communities?”

# NoMM design process



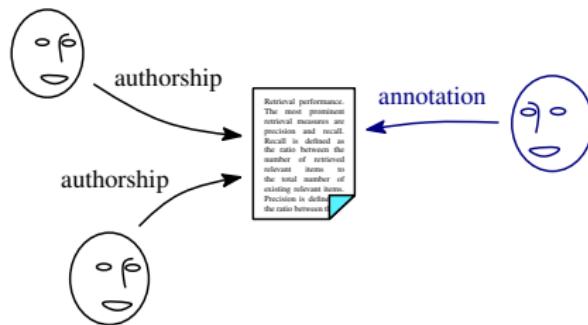
- Typology → “Library” of NoMM substructures
- Idea: Construct models from simple substructures that connect terminal nodes:
  - Terminal nodes ↔ multimodal data (virtual communities...)
  - Substructures ↔ relationships in data; latent semantics
- Process:
  - Assumptions on dependencies in data
  - Iterative association to structures in model (usage of typology)
    - Gibbs distribution known! ↔ model behaviour:  $q(x, y) = \text{“rich get richer”}$
  - Implementation and test with Gibbs meta-sampler; possibly iteration

# NoMM design process



- Typology → “Library” of NoMM substructures
- Idea: Construct models from simple substructures that connect terminal nodes:
  - Terminal nodes ↔ multimodal data (virtual communities...)
  - Substructures ↔ relationships in data; latent semantics
- Process:
  - Assumptions on dependencies in data
  - Iterative association to structures in model (usage of typology)
    - Gibbs distribution known! ↔ model behaviour:  $q(x, y) = \text{“rich get richer”}$
  - Implementation and test with Gibbs meta-sampler; possibly iteration

# Application: Expert finding with tag annotations



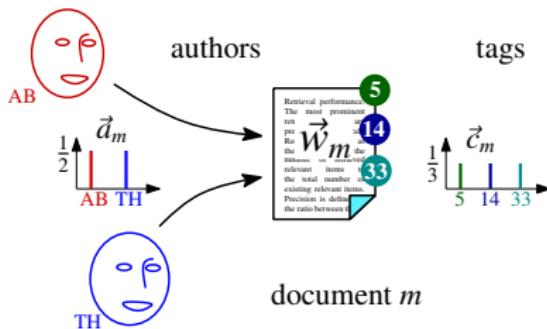
- Scenario: Expert finding via documents with tag annotations
  - Authors of relevant documents → experts
  - Frequently documents with additional annotations, here: tags

→ Goal: Enable tag queries, improve quality of text queries

- Problem: Tags often incomplete, partly wrong
  - Connection of tags and experts via topics

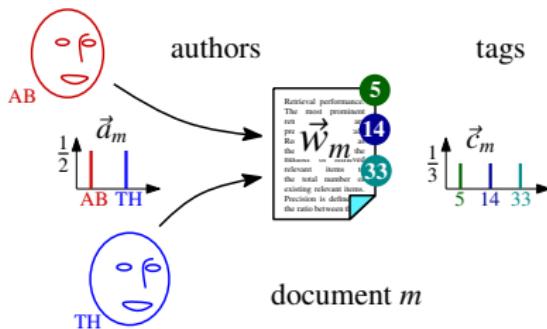
- (1) Data: For each document  $m$ : text  $\vec{w}_m$ , authors  $\vec{d}_m$ , tags  $\vec{c}_m$
- (2) Goal: Tag query  $\vec{c}'$ :  $p(\vec{c}'|a) = \max$ , word query  $\vec{w}'$ :  $p(\vec{w}'|a) = \max$
- (3) Terminal nodes: Authors in input, words and tags in output

# Application: Expert finding with tag annotations



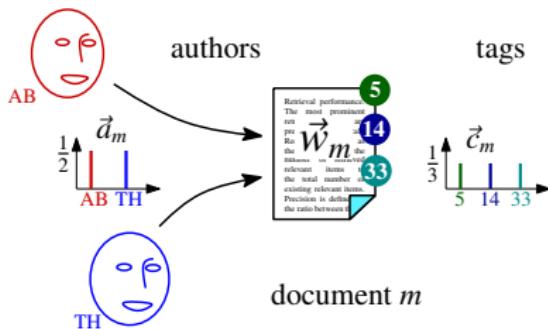
- Scenario: Expert finding via documents with tag annotations
  - Authors of relevant documents → experts
  - Frequently documents with additional annotations, here: tags
- Goal: Enable tag queries, improve quality of text queries
  - Problem: Tags often incomplete, partly wrong
    - Connection of tags and experts via topics
- (1) Data: For each document  $m$ : text  $\vec{w}_m$ , authors  $\vec{d}_m$ , tags  $\vec{c}_m$
- (2) Goal: Tag query  $\vec{c}'$ :  $p(\vec{c}'|a) = \max$ , word query  $\vec{w}'$ :  $p(\vec{w}'|a) = \max$
- (3) Terminal nodes: Authors in input, words and tags in output

# Application: Expert finding with tag annotations



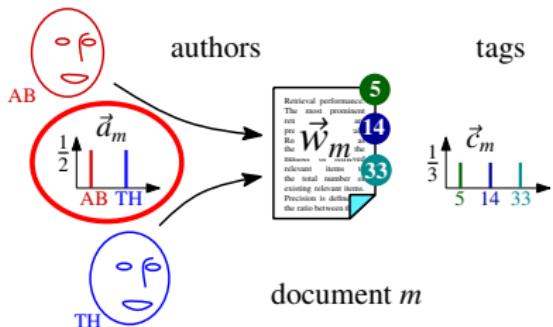
- Scenario: Expert finding via documents with tag annotations
  - Authors of relevant documents → experts
  - Frequently documents with additional annotations, here: tags
- Goal: Enable tag queries, improve quality of text queries
  - Problem: Tags often incomplete, partly wrong
    - Connection of tags and experts via topics
- (1) Data: For each document  $m$ : text  $\vec{w}_m$ , authors  $\vec{d}_m$ , tags  $\vec{c}_m$
- (2) Goal: Tag query  $\vec{c}'$ :  $p(\vec{c}'|a) = \max$ , word query  $\vec{w}'$ :  $p(\vec{w}'|a) = \max$
- (3) Terminal nodes: Authors in input, words and tags in output

# Application: Expert finding with tag annotations



- Scenario: Expert finding via documents with tag annotations
  - Authors of relevant documents → experts
  - Frequently documents with additional annotations, here: tags
- Goal: Enable tag queries, improve quality of text queries
  - Problem: Tags often incomplete, partly wrong
    - Connection of tags and experts via topics
- (1) Data: For each document  $m$ : text  $\vec{w}_m$ , authors  $\vec{d}_m$ , tags  $\vec{c}_m$
- (2) Goal: Tag query  $\vec{c}'$ :  $p(\vec{c}'|a) = \max$ , word query  $\vec{w}'$ :  $p(\vec{w}'|a) = \max$
- (3) Terminal nodes: Authors in input, words and tags in output

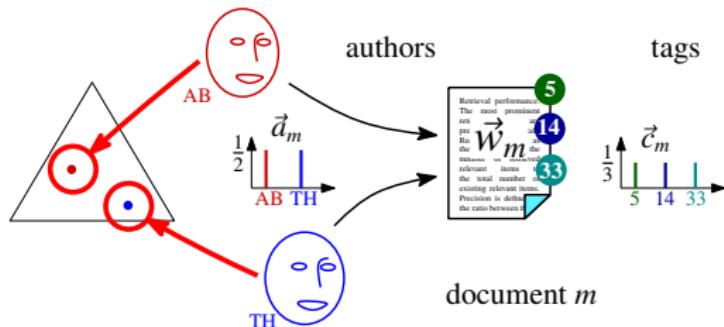
# Model assumptions



## (4) Model assumptions:

- (a) Expertise of an author is weighted with the portion of authorship
- (b) Semantics of expertise expressed by topics  $z$ . Each author has a single field of expertise (topic distribution).
- (c) Semantics of tags expressed by topics  $y$

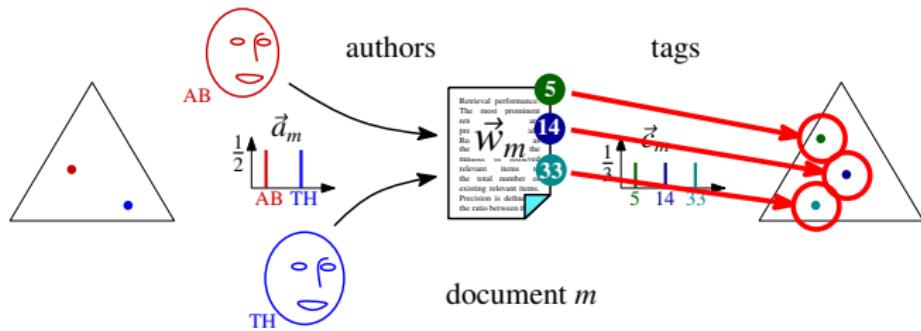
# Model assumptions



## (4) Model assumptions:

- (a) Expertise of an author is weighted with the portion of authorship
- (b) Semantics of expertise expressed by topics  $z$ . Each author has a single field of expertise (topic distribution).
- (c) Semantics of tags expressed by topics  $y$

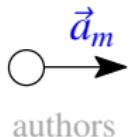
# Model assumptions



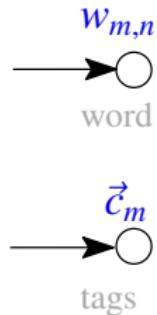
## (4) Model assumptions:

- (a) Expertise of an author is weighted with the portion of authorship
- (b) Semantics of expertise expressed by topics  $z$ . Each author has a single field of expertise (topic distribution).
- (c) Semantics of tags expressed by topics  $y$

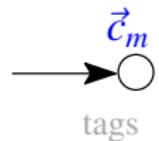
# Model construction



authors



word

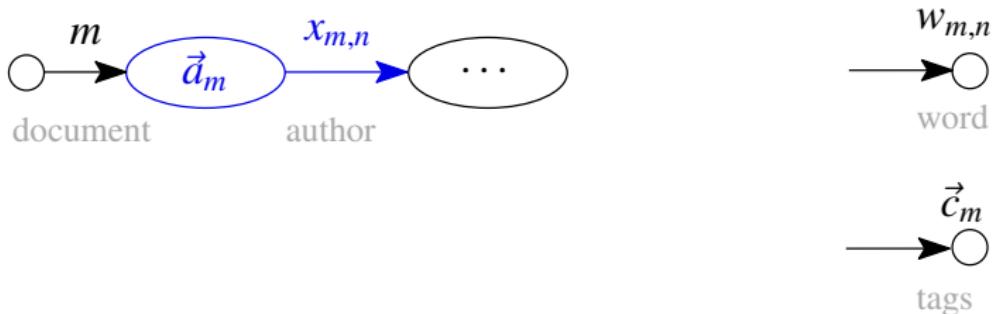


tags

$$p(\dots | \vec{d}, \vec{w}, \vec{c}) \propto \dots$$

(5) Model construction: (a) Start with terminal nodes (from step 3)

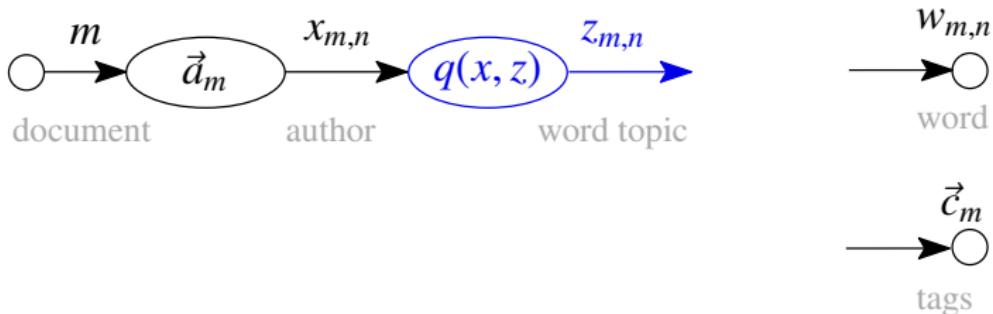
# Model construction



$$p(\mathbf{x}, \dots | \cdot) \propto a_{m,x} q(\mathbf{x}, \dots) \dots$$

(b) Authorship  $\vec{d}_m$  given as observed distribution  
→ node samples author  $x$  of a word

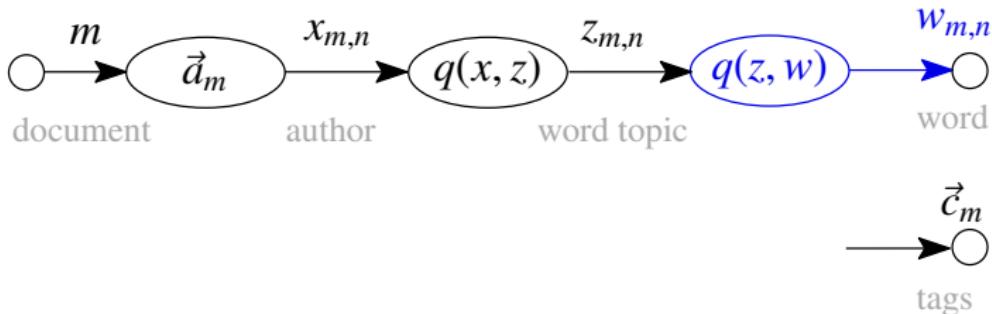
# Model construction



$$p(x, z, \dots | \cdot) \propto a_{m,x} q(x, z) \dots$$

- (c) Each author has only a single field of expertise (topic distribution)  
→  $q(x, z)$  associates (word-)topics with sampled authors  $x$  (cf. ATM)

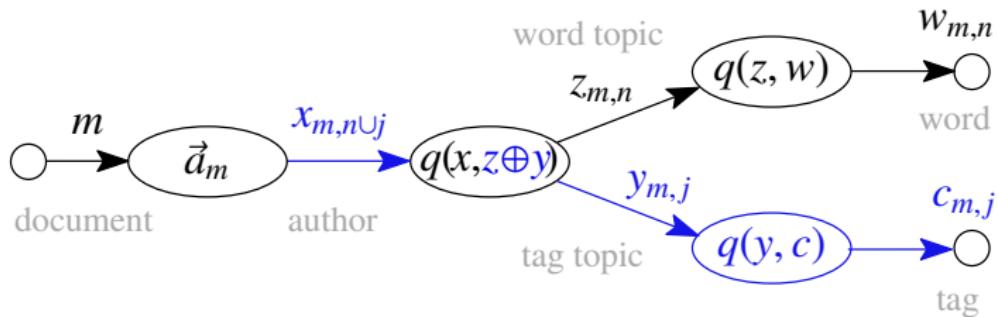
# Model construction



$$p(x, z, \dots | \cdot) \propto a_{m,x} q(x, z) \textcolor{blue}{q(z, w)} \dots$$

(d) Topic distribution over terms  
→ connect  $z$  and  $w$  via  $q(z, w)$

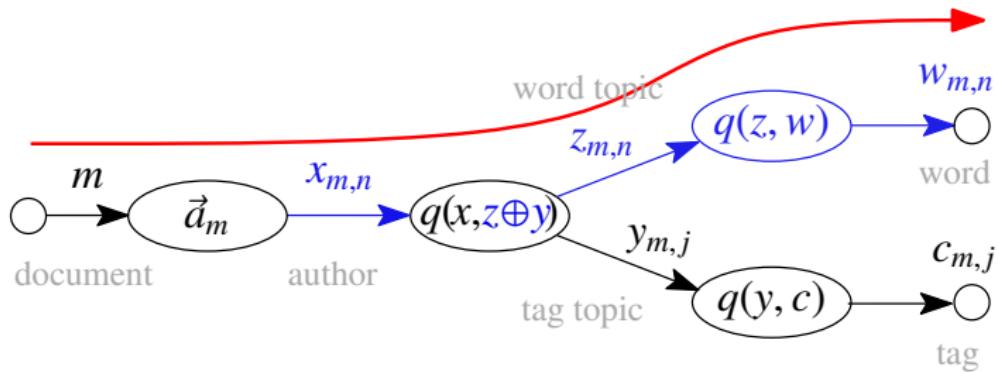
# Model construction



$$p(x, z, y | \cdot) \propto a_{m,x} q(x, z \oplus y) q(z, w) q(y, c)$$

- (e) Introduce tag topics  $y_{m,j}$  for  $c_{m,j}$  as distributions over tags  
 $q(x, z \oplus y)$  overlays values for  $z$  and  $y$

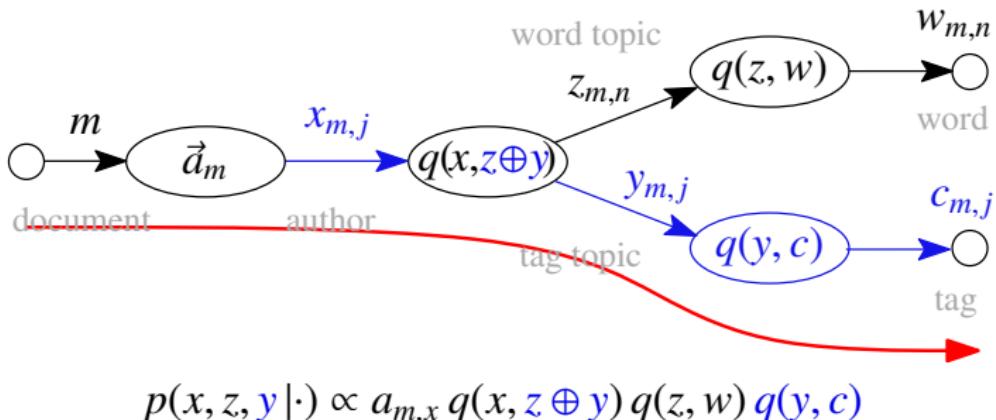
# Model construction



$$p(x, z, y | \cdot) \propto a_{m,x} q(x, z \oplus y) q(z, w) q(y, c)$$

- (e) Introduce tag topics  $y_{m,j}$  for  $c_{m,j}$  as distributions over tags  
 $q(x, z \oplus y)$  overlays values for  $z$  and  $y$

# Model construction

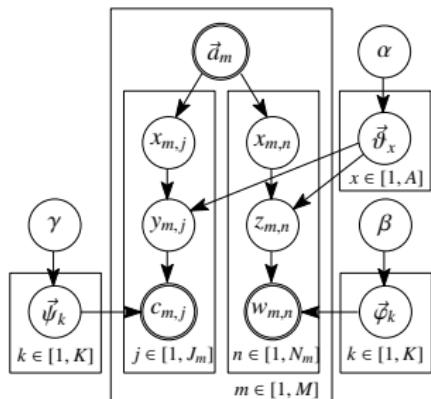


(e) Introduce tag topics  $y_{m,j}$  for  $c_{m,j}$  as distributions over tags  
 $q(x, z \oplus y)$  overlays values for  $z$  and  $y$

# Model construction – ordinary approach

## Expert–tag–topic model

(Heinrich 2011)



$$\begin{aligned}
p(\vec{w}, \vec{c}, \vec{d}, \vec{x}, \vec{z}, \vec{\theta}, \vec{\varphi}, \vec{\psi}, \vec{y}, \beta, \gamma) &= p(\vec{w}|\vec{c}, \vec{\theta})p(\vec{d}|\vec{\theta}) \cdot p(\vec{x}|\vec{c}, \vec{y})p(\vec{z}|\vec{y}) \\
&\quad \cdot p(\vec{\theta}|\alpha) \cdot p(\vec{\varphi}|\alpha) \cdot p(\vec{y}|\beta) \quad (\text{E.1}) \\
&= \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{z_{m,n}}) p(z_{m,n} | \vec{\theta}_{z_{m,n}}) a_{m,n,z_{m,n}} \right. \\
&\quad \left. \cdot \prod_{j=1}^{J_m} p(c_{m,j} | \vec{v}_{y_{m,j}}) p(y_{m,j} | \vec{\theta}_{y_{m,j}}) a_{m,j,y_{m,j}} \right) \\
&\quad \cdot p(\vec{\theta}|\alpha) \cdot p(\vec{\varphi}|\beta) \cdot p(\vec{y}|\gamma). \quad (\text{E.2})
\end{aligned}$$

$$\begin{aligned}
p(\vec{w}, \vec{c}, \vec{d}, \vec{x}, \vec{z}, \vec{\theta}, \vec{\varphi}, \vec{y}, \beta, \gamma) &= \int \int \int \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{z_{m,n}}) p(z_{m,n} | \vec{\theta}_{z_{m,n}}) a_{m,n,z_{m,n}} \right. \\
&\quad \left. \cdot \prod_{j=1}^{J_m} p(c_{m,j} | \vec{v}_{y_{m,j}}) p(y_{m,j} | \vec{\theta}_{y_{m,j}}) a_{m,j,y_{m,j}} \right) \\
&\quad \cdot d\mu(\vec{\theta}|\alpha) \cdot d\mu(\vec{\varphi}|\beta) \cdot d\mu(\vec{y}|\gamma) \quad (\text{E.3})
\end{aligned}$$

$$\begin{aligned}
&= \int \prod_{m=1}^M \prod_{n=1}^{N_m} p(w_{m,n} | \vec{v}_{z_{m,n}}) \prod_{k=1}^K p(\vec{\varphi}_k | \vec{\theta}) d\vec{\varphi}_k \\
&\quad \cdot \int \prod_{m=1}^M \prod_{n=1}^{N_m} p(c_{m,n} | \vec{v}_{y_{m,n}}) \prod_{k=1}^K p(\vec{y}_k | \vec{\theta}) d\vec{y}_k \\
&\quad \cdot \int \prod_{m=1}^M p(\vec{\theta}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n} | \vec{\theta}_{z_{m,n}}) a_{m,n,z_{m,n}} \prod_{j=1}^{J_m} p(y_{m,j} | \vec{\theta}_{y_{m,j}}) a_{m,j,y_{m,j}} d\vec{\theta}_m \quad (\text{E.4})
\end{aligned}$$

$$\begin{aligned}
&= \int \prod_{k=1}^K \frac{1}{\Delta_V(\vec{\theta})} \prod_{l=1}^V \varphi_{k,l}^{a_{k,l} + \beta - 1} d\vec{\varphi}_k \cdot \int \prod_{k=1}^K \frac{1}{\Delta_C(\vec{y})} \prod_{c=1}^C \varphi_{k,c}^{a_{k,c} + \gamma - 1} d\vec{y}_k \\
&\quad \cdot \int \prod_{m=1}^M \prod_{n=1}^{N_m} \frac{1}{\Delta_K(\alpha)} \prod_{k,k'}^{K-k} \theta_{k,k'}^{a_{k,k'} + a_{k,k'}^{(z)} + \alpha - 1} d\vec{\theta}_m \cdot \prod_{m=1}^M \prod_{n=1}^{N_m} \theta_{m,n}^{a_{m,n}^{(z)} + a_{m,n}^{(y)}} \quad (\text{E.5})
\end{aligned}$$

$$= \prod_{k=1}^K \frac{\Delta(\vec{h}_k^{(z)} + \beta)}{\Delta_V(\vec{\theta})} \cdot \frac{\Delta(\vec{h}_k^{(y)} + \gamma)}{\Delta_C(\vec{y})} \prod_{m=1}^M \frac{\Delta(\vec{h}_m^{(z)} + \vec{h}_m^{(y)} + \alpha)}{\Delta_K(\alpha)} \prod_{m=1}^M a_{m,n}^{a_{m,n}^{(z)} + a_{m,n}^{(y)}}. \quad (\text{E.6})$$

$$\begin{aligned}
p(z_i=k, x_i=i | w_i=i, \vec{z}_{-i}, \vec{y}_{-i}, \vec{x}_{-i}, \vec{w}_{-i}, \vec{d}, \vec{c}) &= \frac{p(\vec{w} | \vec{z}, \vec{y}, \vec{d})}{p(\vec{w}, \vec{z}_{-i}, \vec{y}_{-i}, \vec{d}_{-i})} = \frac{p(\vec{w} | \vec{z}, \vec{y})}{p(\vec{w}, \vec{z}_{-i}, \vec{y}) p(w_i)} \cdot \frac{p(\vec{z} | \vec{x})}{p(\vec{z}_{-i} | \vec{x}_{-i})} \cdot \frac{p(\vec{d})}{p(\vec{x}_{-i})} \quad (\text{E.7})
\end{aligned}$$

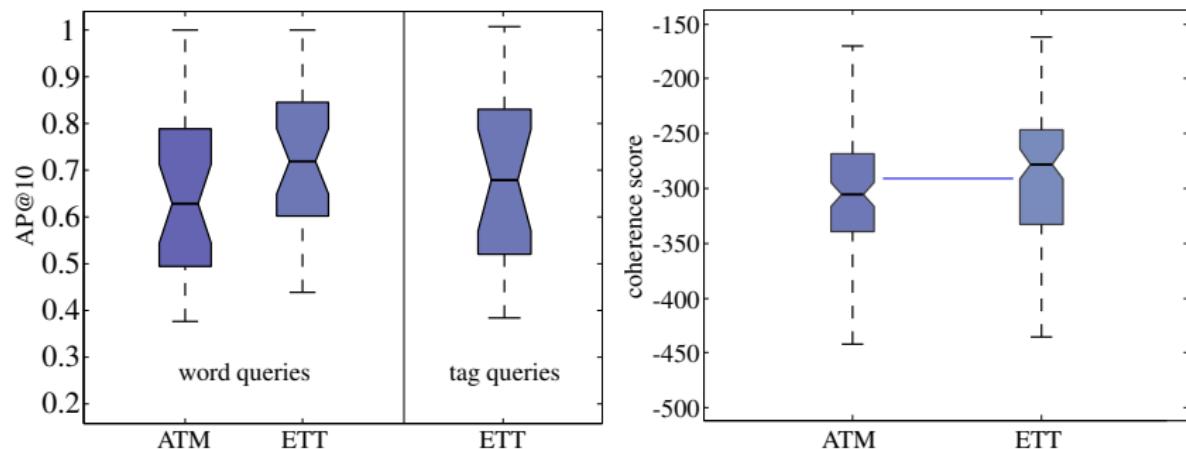
$$\begin{aligned}
&\propto \frac{\Delta(\vec{h}_k^{(z)} + \beta)}{\Delta(\vec{h}_{k,-i}^{(z)} + \beta)} \cdot \frac{\Delta(\vec{h}_i + \alpha)}{\Delta(\vec{h}_{k,-i}^{(z)} + \beta)} \cdot a_{m,i} \quad (\text{E.8})
\end{aligned}$$

$$= \frac{\Gamma(n_{k,i} + \beta)}{\Gamma(n_{k,-i} + \beta)} \frac{\Gamma(n_{k,i} + V\beta)}{\Gamma(n_{k,-i} + V\beta)} \cdot \frac{\Gamma(n_{i,-i}^{(z)} + \alpha)}{\Gamma(n_{k,-i}^{(z)} + \alpha)} \frac{\Gamma(n_{i,-i}^{(z)} + K\alpha)}{\Gamma(n_i^{(z)} + K\alpha)} \cdot a_{m,i} \quad (\text{E.9})$$

$$= \frac{n_{k,i,-i} + \beta}{n_{k,-i} + V\beta} \cdot \frac{n_{i,-i,-i}^{(z)} + \alpha}{n_{k,-i}^{(z)} + K\alpha} \cdot a_{m,i} \quad (\text{E.10})$$

$$\begin{aligned}
p(y_j=k, x_i=i | c_j=c, \vec{z}_{-i}, \vec{y}_{-i}, \vec{x}_{-i}, \vec{w}_{-i}, \vec{d}, \vec{c}_{-i}) &\propto \frac{n_{k,x,-i}^{(y)} + \gamma}{n_{k,-i} + V\gamma} \cdot \frac{n_{i,x,-i}^{(y)} + \alpha}{n_{k,-i}^{(y)} + K\alpha} \cdot a_{m,x} \quad (\text{E.12})
\end{aligned}$$

# Expert–tag–topic model: Evaluation



- NIPS Corpus: 2.3 million words, 2037 authors, 165 tags
- Retrieval: Average Precision @10:
  - Term queries: ETT > ATM
  - Tag queries: Similarly good AP values
- Topic coherence (Mimno et al. 2011): ETT > ATM
- Semi-supervised learning: Tag queries retrieve items without tags

# Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- **Conclusions and outlook**

## Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
  - Inference algorithms: Generic Gibbs sampler
    - Fast sampling methods (serial, parallel, independent)
    - Implementation in Gibbs meta-sampler
  - Design process based on typology of NoMM substructures
  - Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
  - Inference algorithms: Generic Gibbs sampler
    - Fast sampling methods (serial, parallel, independent)
    - Implementation in Gibbs meta-sampler
    - + [Variational Inference for NoMMs \(Heinrich and Goesele 2009\)](#)
  - Design process based on typology of NoMM substructures
  - Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
- Inference algorithms: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
  - + [Variational Inference for NoMMs \(Heinrich and Goesele 2009\)](#)
- Design process based on typology of NoMM substructures
  - + [AMQ model: Meta-model for virtual communities as formal basis for scenario modelling \(Heinrich 2010\)](#)
- Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
  - Inference algorithms: Generic Gibbs sampler
    - Fast sampling methods (serial, parallel, independent)
    - Implementation in Gibbs meta-sampler
    - + [Variational Inference for NoMMs \(Heinrich and Goesele 2009\)](#)
  - Design process based on typology of NoMM substructures
    - + [AMQ model: Meta-model for virtual communities as formal basis for scenario modelling \(Heinrich 2010\)](#)
  - Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
    - + [Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches \(Heinrich 2011b\)](#)
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
  - Inference algorithms: Generic Gibbs sampler
    - Fast sampling methods (serial, parallel, independent)
    - Implementation in Gibbs meta-sampler
    - + [Variational Inference for NoMMs \(Heinrich and Goesele 2009\)](#)
  - Design process based on typology of NoMM substructures
    - + [AMQ model: Meta-model for virtual communities as formal basis for scenario modelling \(Heinrich 2010\)](#)
  - Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
    - + [Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches \(Heinrich 2011b\)](#)
    - + [Graph-based expert search using ETT models: Integration of explorative search/browsing and distributions from topic models](#)
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Conclusions: Research contributions

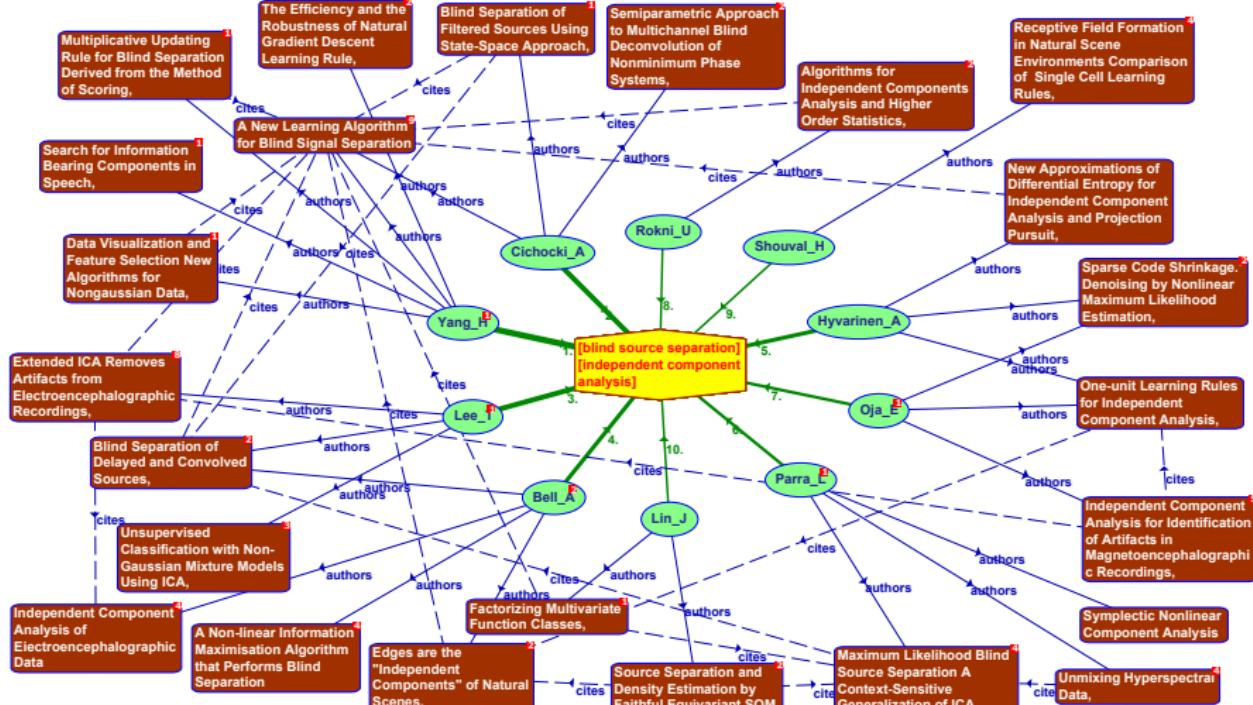


Figure: ETT1: Expert search in community browser

# Conclusions: Research contributions

- *Networks of Mixed Membership*: Generic model and domain-specific compact representation of topic models
  - Inference algorithms: Generic Gibbs sampler
    - Fast sampling methods (serial, parallel, independent)
    - Implementation in Gibbs meta-sampler
    - + [Variational Inference for NoMMs \(Heinrich and Goesele 2009\)](#)
  - Design process based on typology of NoMM substructures
    - + [AMQ model: Meta-model for virtual communities as formal basis for scenario modelling \(Heinrich 2010\)](#)
  - Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
    - + [Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches \(Heinrich 2011b\)](#)
    - + [Graph-based expert search using ETT models: Integration of explorative search/browsing and distributions from topic models](#)
- Σ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios

# Outlook

- New applications and NoMM structures, e.g., time as variable
- Alternative inference methods:
  - Generic Collapsed Variational Bayes (Teh et al. 2007): Structure similar to Collapsed Gibbs-Sampler
  - Non-parametric methods: Learning model dimensions using Dirichlet or Pitman–Yor process priors (Teh et al. 2004; Buntine and Hutter 2010), NoMM polymorphism (Heinrich 2011a)
- Improved support in design process:
  - Data-driven design: Search over model structures to obtain best model for data set
  - Architecture-specific Gibbs meta-sampler, e.g., massively-parallel or FPGA, cf. (Heinrich et al. 2011)
- Integration with interactive user interfaces: Models can be created on the fly, e.g., for visual analytics

Thank you!

Q + A

# References I

## References

- Barnard, K., P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan (2003, August).  
Matching words and pictures.  
*JMLR – Special Issue on Machine Learning Methods for Text and Images* 3(6), 1107–1136.
- Bellegarda, J. (2000, August).  
Exploiting latent semantic information in statistical language modeling.  
*Proc. IEEE* 88(8), 1279–1296.
- Blei, D., A. Ng, and M. Jordan (2003, January).  
Latent Dirichlet allocation.  
*Journal of Machine Learning Research* 3, 993–1022.
- Buntine, W. and M. Hutter (2010).  
A Bayesian review of the Poisson-Dirichlet process.  
*arXiv:1007.0296v1 [math.ST]*.
- Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei (2009).  
Reading tea leaves: How humans interpret topic models.  
In *Proc. Neural Information Processing Systems (NIPS)*.

## References II

Dietz, L., S. Bickel, and T. Scheffer (2007, June).

Unsupervised prediction of citation influences.

In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, Oregon, USA.

Heinrich, G. (2009).

A generic approach to topic models.

In *Proc. European Conf. on Mach. Learn. / Principles and Pract. of Know. Discov. in Databases (ECML/PKDD)*, Part 1, pp. 517–532.

Heinrich, G. (2010).

Actors—media—qualities: a generic model for information retrieval in virtual communities.

In *Proc. 7th International Workshop on Innovative Internet Community Systems (I2CS 2007), part of I2CS Jubilee proceedings, Lecture Notes in Informatics, GI*.

Heinrich, G. (2011a, March).

“Infinite LDA” – Implementing the HDP with minimum code complexity.

Technical note TN2011/1, arbylon.net.

Heinrich, G. (2011b).

Typology of mixed-membership models: Towards a design method.

In *Proc. European Conf. on Mach. Learn. / Principles and Pract. of Know. Discov. in Databases (ECML/PKDD)*.

# References III

- Heinrich, G. and M. Goesele (2009).  
Variational Bayes for generic topic models.  
In *Proc. 32nd Annual German Conference on Artificial Intelligence (KI2009)*.
- Heinrich, G., J. Kindermann, C. Lauth, G. Paaß, and J. Sanchez-Monzon (2005).  
Investigating word correlation at different scopes – a latent concept approach.  
In *Workshop Lexical Ontology Learning at Int. Conf. Mach. Learning*.
- Heinrich, G., F. Logemann, V. Hahn, C. Jung, G. Figueiredo, and W. Luk (2011).  
*HW/SW co-design for heterogeneous multi-core platforms: The hArtes toolchain*, Chapter Audio  
array processing for telepresence, pp. 173–207.  
Springer.
- Li, W., D. Blei, and A. McCallum (2007).  
Mixtures of hierarchical topics with pachinko allocation.  
In *International Conference on Machine Learning*.
- Li, W. and A. McCallum (2006).  
Pachinko allocation: DAG-structured mixture models of topic correlations.  
In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, New York,  
NY, USA, pp. 577–584. ACM.

# References IV

- Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011, July). Optimizing semantic coherence in topic models.  
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK*, pp. 262272.
- Newman, D., A. Asuncion, P. Smyth, and M. Welling (2009, August). Distributed algorithms for topic models.  
*JMLR* 10, 1801–1828.
- Porteous, I., D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation.  
In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 569–577. ACM.
- Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth (2004). The author-topic model for authors and documents.  
In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Teh, Y., M. Jordan, M. Beal, and D. Blei (2004). Hierarchical Dirichlet processes.  
Technical Report 653, Department of Statistics, University of California at Berkeley.
- Teh, Y. W., D. Newman, and M. Welling (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.  
In *Advances in Neural Information Processing Systems*, Volume 19.

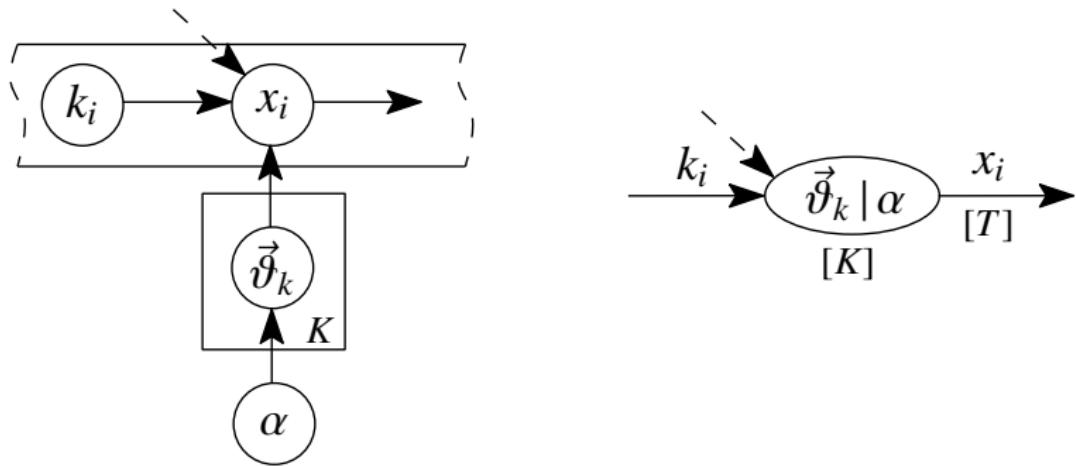
## Appendix

# Example: Text mining for semantic clusters

Topic label	Dominant terms according to $\varphi_{k,t} = p(\text{term} \text{topic})$
Bundesliga	FC SC München Borussia SV VfL Kickers SpVgg Uhr Köln Bochum Freiburg VfB Eintracht Bayern Hamburger Bayern+München
Polizei / Unfall	Polizei verletzt schwer Auto Unfall Fahrer Angaben schwer+verletzt Menschen Wagen Verletzungen Lawine Mann vier Meter Straße
Tschetschenien	Rebellen russischen Grosny russische Tschetschenien Truppen Kaukasus Moskau Angaben Interfax tschetschenischen Agentur
Politik / Hessen	FDP Koch Hessen CDU Koalition Gerhardt Wagner Liberalen hessischen Westerwelle Wolfgang Roland+Koch Wolfgang+Gerhardt
Wetter	Grad Temperaturen Regen Schnee Süden Norden Sonne Wetter Wolken Deutschland zwischen Nacht Wetterdienst Wind
Politik / Kroatien	Parlament Partei Stimmen Mehrheit Wahlen Wahl Opposition Kroatien Präsident Parlamentswahlen Mesic Abstimmung HDZ
Die Grünen	Grünen Parteitag Atomausstieg Trittin Grüne Partei Trennung Mandat Aussieg Amt Roestel Jahren Müller Radcke Koalition
Russische Politik	Russland Putin Moskau russischen russische Jelzin Wladimir Tschetschenien Russlands Wladimir+Putin Kreml Boris Präsidenten
Polizei / Schulen	Polizei Schulen Schüler Täter Polizisten Schule Tat Lehrer erschossen Beamten Mann Polizist Beamte verletzt Waffe

Bigram-LDA: Topics from 18400 dpa news messages, Jan. 2000 (Heinrich et al. 2005)

# Notation: Bayesian network vs. NoMM levels

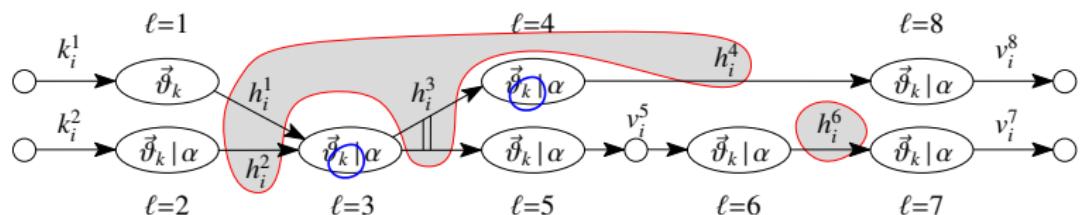
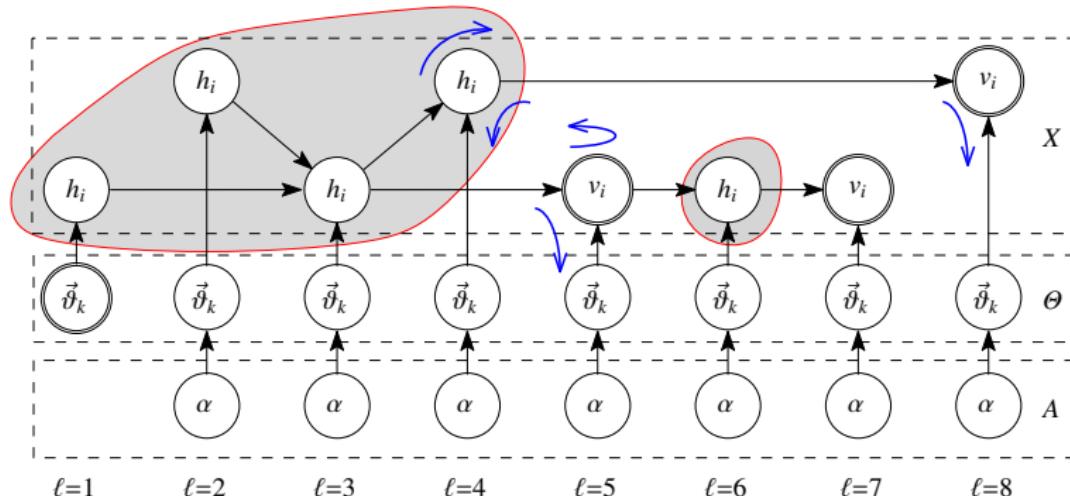


parameters  $\vartheta$  + hyperparameters  $\alpha \Leftrightarrow$  nodes  $(\vartheta | \alpha)$

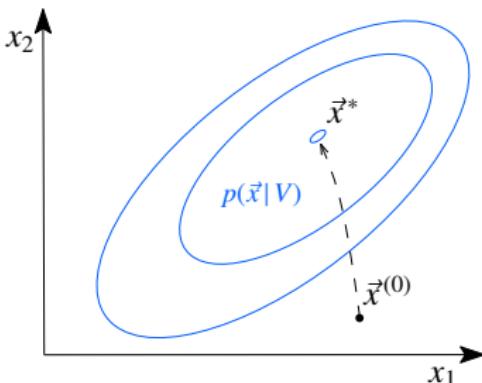
variables  $k_i, x_i \Leftrightarrow$  edges  $k_i, x_i$

plates (i.i.d. repetitions)  $i, k \Leftrightarrow$  indexes  $i$  + dimensions  $k$

# NoMM representation: Variable dependencies

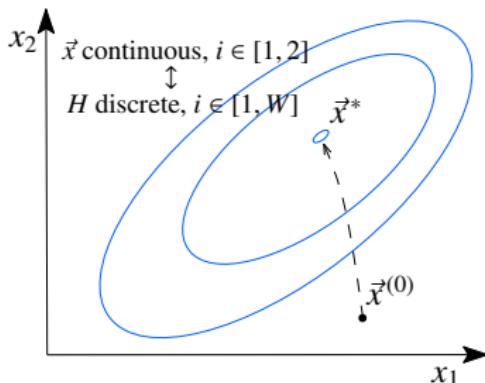


# Collapsed Gibbs sampler



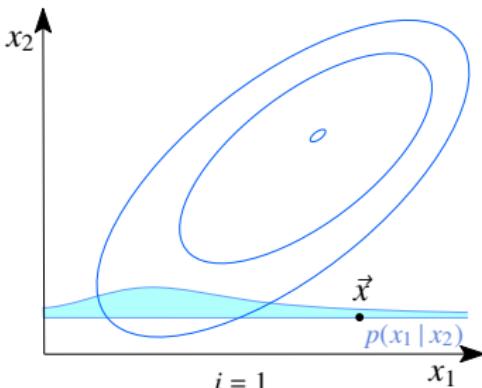
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



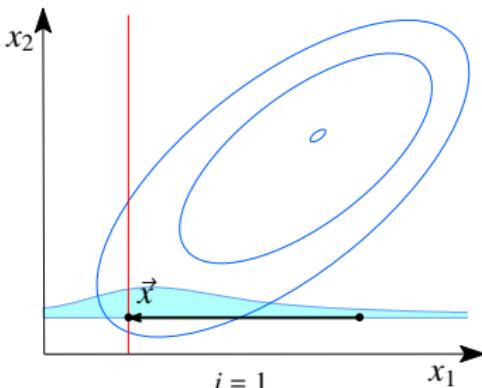
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



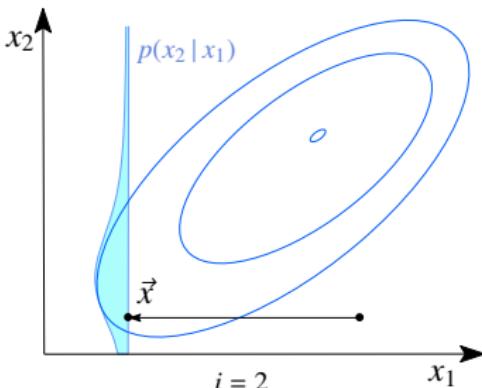
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A) . \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



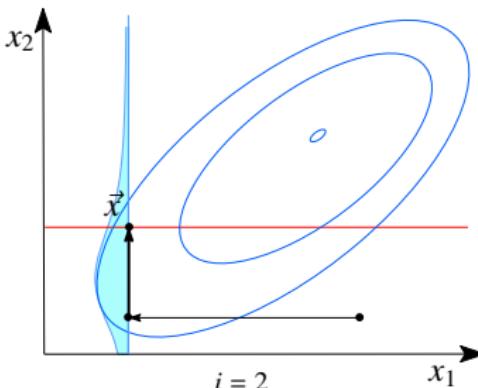
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



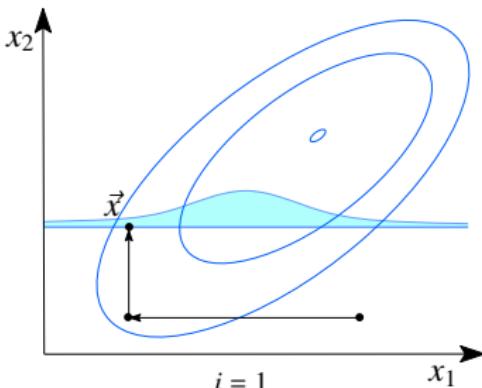
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



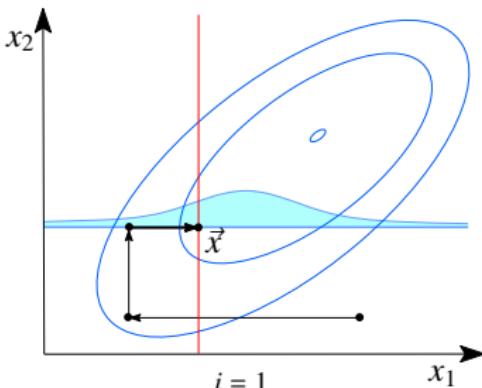
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



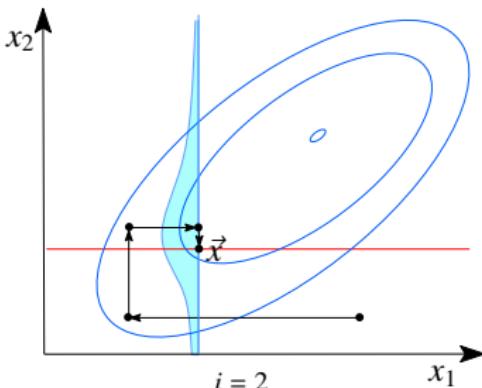
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



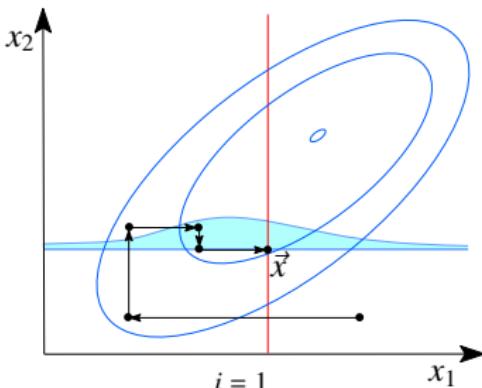
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



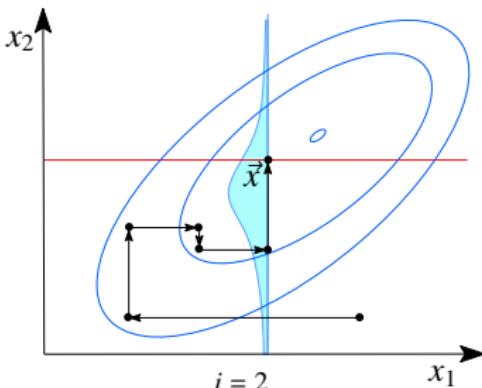
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



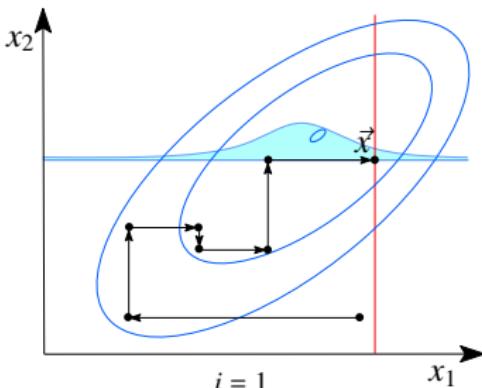
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



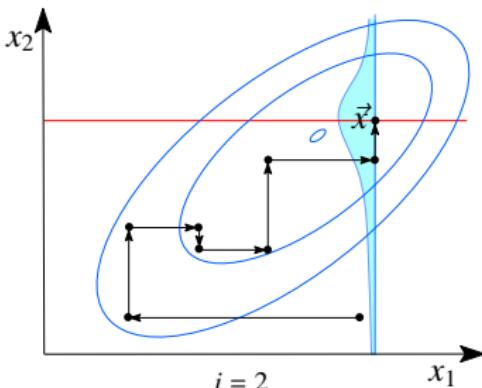
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



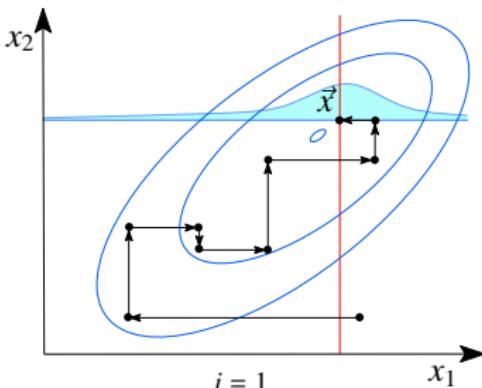
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



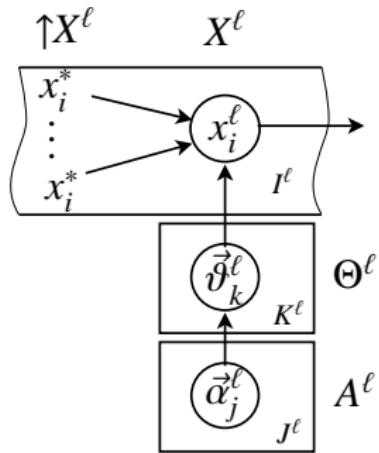
- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- $$H_i \sim p(H_i | H_{\neg i}, V, A). \quad (1)$$
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Collapsed Gibbs sampler



- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters  $\Theta$  correlate with  $H \rightarrow$  marginalise
  - For each data point,  $i$ : draw latent variables,  $H_i = (y_i, z_i, \dots)$ , given all other data, latent,  $H_{\neg i}$ , and observed,  $V$ :
- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)

# Generic topic models: Generative process



Generative process on level  $\ell$ :

$$x_i \sim \text{Mult}(x_i | \vec{\vartheta}_k) \quad \vec{\vartheta}_k \sim \text{Dir}(\vec{\vartheta}_k | \vec{a}_j) \quad (2)$$

$$k = f_k(\text{parents}(x_i), i) \quad j = f_j(\text{known\_parents}(x_i), i) . \quad (3)$$

# Generic topic models: Complete-data likelihood

Likelihood of all hidden and visible data  $X = \{H, V\}$  and parameters  $\Theta$ :

$$\begin{aligned} p(X, \Theta | A) &= \prod_{\ell \in L} \left[ \underbrace{\prod_i p(x_{i,\text{out}} | \vec{\vartheta}_{x_{i,\text{in}}})}_{\text{data items } \sim \text{Discrete } \boxplus} \cdot \underbrace{\prod_k p(\vec{\vartheta}_k | \vec{\alpha})}_{\text{components } \sim \text{Dirichlet } \Delta} \right]^{[\ell]} \\ &= \prod_{\ell \in L} \left[ \prod_k f(\vec{\vartheta}_k, \vec{n}_{\textcolor{blue}{k}}, \vec{\alpha}) \right]^{[\ell]} \quad \vec{n}_{\textcolor{blue}{k}} = (n_{k,1}, n_{k,2}, \dots) \end{aligned} \quad (4)$$

- Product dependent on co-occurrences  $n_{k,t}$  between **input** and **output** values,  $x_{i,\text{in}}=k$  and  $x_{i,\text{out}}=t$ , on each level  $\ell$
- There are variants to component selection  $x_{i,\text{in}}=k$
- There are mixture node variants, e.g., observed components

## Generic topic models: Complete-data likelihood

The conjugacy between the multinomial and Dirichlet distributions of model levels leads to a simple complete-data likelihood:

$$p(X, \Theta | A) = \prod_{\ell} \prod_i \text{Mult}(x_i^{\ell} | \Theta^{\ell}, k_i^{\ell}) \prod_k \text{Dir}(\vec{\vartheta}_k^{\ell} | \vec{\alpha}_j^{\ell}) \quad (5)$$

$$= \prod_{\ell} \left[ \prod_i \vartheta_{k_i, x_i} \prod_k \frac{1}{B(\vec{\alpha}_j)} \prod_t \vartheta_{k, x_i}^{\alpha_j - 1} \right]^{\ell} \quad (6)$$

$$= \prod_{\ell} \left[ \prod_k \frac{1}{B(\vec{\alpha}_j)} \prod_t \vartheta_{k, x_i}^{\alpha_j + n_{k,t} - 1} \right]^{\ell} \quad (7)$$

$$= \prod_{\ell} \left[ \prod_k \frac{B(\vec{n}_k + \vec{\alpha}_j)}{B(\vec{\alpha}_j)} \text{Dir}(\vec{\vartheta}_k | \vec{n}_k + \vec{\alpha}_j) \right]^{\ell} \quad (8)$$

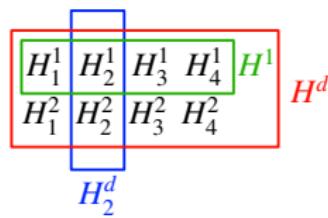
where brackets  $[\cdot]^{\ell}$  enclose a particular level  $\ell$ .  
 $n_{k,t}$  is how often  $k$  and  $t$  co-occur.

# Inference: Generic full conditionals

Gibbs full conditionals are derived for groups of dependent hidden edges,  $H_i^d \in H^d \subset X$  and “surrounding” edges  $S_i^d \in S^d$  considered observed. All tokens co-located with a particular observation:  $X_i^d = \{H_i^d, S_i^d\}$ .  
Full conditional via chain rule applied to (8) with  $\Theta$  integrated out:

$$p(H_i^d | X \setminus H_i^d, A) = \frac{p(H_i^d, S_i^d | X \setminus \{H_i^d, S_i^d\}, A)}{p(S_i^d | X \setminus \{H_i^d, S_i^d\}, A)} \quad (9)$$

$$\propto p(X_i^d | X \setminus X_i^d, A) = \frac{p(X | A)}{p(X \setminus X_i^d | A)} \quad (10)$$



$$= \prod_{\ell} \left[ \prod_k \frac{B(\vec{n}_k + \vec{\alpha}_j)}{B(\vec{n}_k \setminus X_i^d + \vec{\alpha}_j)} \right]^{\ell} \quad (11)$$

$$\propto \prod_{\ell \in \{H^d, S^d\}} \left[ \frac{B(\vec{n}_k + \vec{\alpha}_j)}{B(\vec{n}_k \setminus X_i^d + \vec{\alpha}_j)} \right]^{\ell} \quad (12)$$

# Inference: $q$ -functions

$$q(\textcolor{blue}{k}, \textcolor{red}{t}) = \frac{\text{B}(\vec{n}_{\textcolor{blue}{k}} + \vec{\alpha}_j)}{\text{B}(\vec{n}_{\textcolor{blue}{k}} \setminus x_i^d + \vec{\alpha}_j)} \stackrel{|x_i^d|=1}{=} \frac{n_{\textcolor{blue}{k}, \textcolor{red}{t}}^{\neg i} + \alpha}{\sum_t n_{\textcolor{blue}{k}, \textcolor{red}{t}}^{\neg i} + \alpha}$$
$$\stackrel{|x_i^d|=2}{=} \frac{n_{\textcolor{blue}{k}, \textcolor{red}{t}} \setminus x_{i,1}^d + \alpha}{\sum_t n_{\textcolor{blue}{k}, \textcolor{red}{t}} \setminus x_{i,1}^d + \alpha} \cdot \frac{n_{\textcolor{blue}{k}, \textcolor{red}{t}} \setminus x_{i,2}^d + \alpha + \delta(x_{i,1}^d - x_{i,2}^d)}{\sum_t n_{\textcolor{blue}{k}, \textcolor{red}{t}} \setminus x_{i,2}^d + \alpha + 1}$$

...

# $q$ -functions: Pólya urn and sampling weights

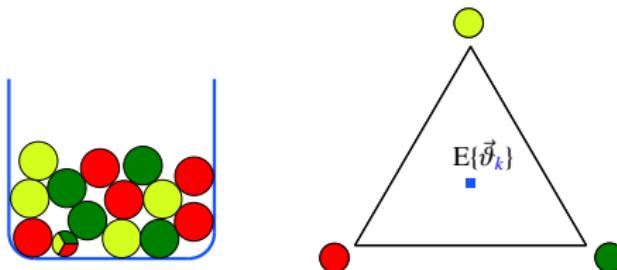


Figure: Pólya urn: sampling with over-replacement.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{\text{B}(\vec{n}_{\mathbf{k}} + \alpha)}{\text{B}(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \mathbf{t}}^{\neg t_i} + \alpha}{n_{\mathbf{k}}^{\neg t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u, v\}}{=} \frac{n_{\mathbf{k}, \mathbf{u}}^{\neg u_i} + \alpha}{n_{\mathbf{k}}^{\neg u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, \mathbf{v}}^{\neg v_i} + \alpha + \delta(\mathbf{u} - \mathbf{v})}{n_{\mathbf{k}}^{\neg v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v})$$

...

# $q$ -functions: Pólya urn and sampling weights

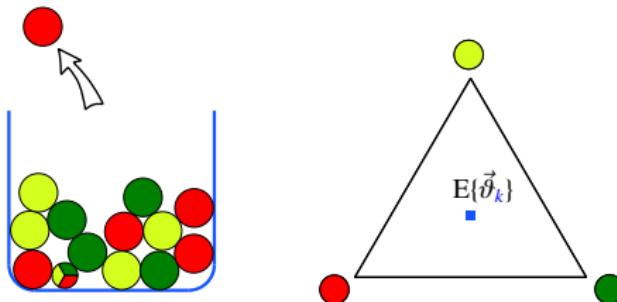


Figure: Pólya urn: sampling with over-replacement.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, t}^{\neg t_i} + \alpha}{n_{\mathbf{k}}^{\neg t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u, v\}}{=} \frac{n_{\mathbf{k}, u}^{\neg u_i} + \alpha}{n_{\mathbf{k}}^{\neg u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, v}^{\neg v_i} + \alpha + \delta(u - v)}{n_{\mathbf{k}}^{\neg v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v})$$

...

# $q$ -functions: Pólya urn and sampling weights

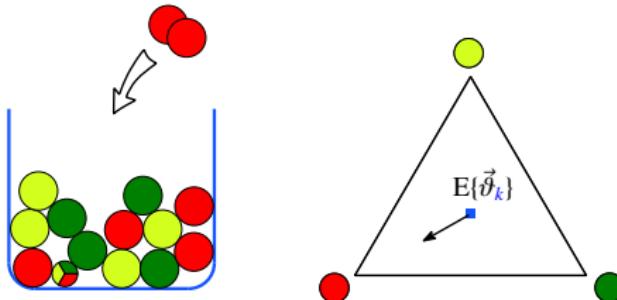


Figure: Pólya urn: sampling with over-replacement.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{\text{B}(\vec{n}_{\mathbf{k}} + \alpha)}{\text{B}(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \mathbf{t}}^{\neg t_i} + \alpha}{n_{\mathbf{k}}^{\neg t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u,v\}}{=} \frac{n_{\mathbf{k}, \mathbf{u}}^{\neg u_i} + \alpha}{n_{\mathbf{k}}^{\neg u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, \mathbf{v}}^{\neg v_i} + \alpha + \delta(\mathbf{u} - \mathbf{v})}{n_{\mathbf{k}}^{\neg v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v})$$

...

# $q$ -functions: Pólya urn and sampling weights

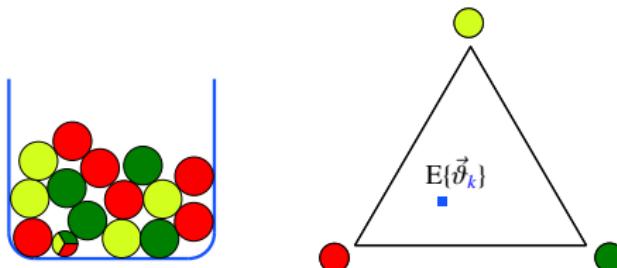


Figure: Pólya urn: sampling with over-replacement.

$$\begin{aligned} q(\mathbf{k}, \mathbf{t}) &\triangleq \frac{\text{B}(\vec{n}_{\mathbf{k}} + \alpha)}{\text{B}(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \mathbf{t}}^{\neg t_i} + \alpha}{n_{\mathbf{k}}^{\neg t_i} + T\alpha} = \text{smoothed ratio of occurrences} \\ &\stackrel{t=\{u, v\}}{=} \frac{n_{\mathbf{k}, \mathbf{u}}^{\neg u_i} + \alpha}{n_{\mathbf{k}}^{\neg u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, \mathbf{v}}^{\neg v_i} + \alpha + \delta(\mathbf{u} - \mathbf{v})}{n_{\mathbf{k}}^{\neg v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v}) \end{aligned}$$

...

# $q$ -functions: Pólya urn and sampling weights

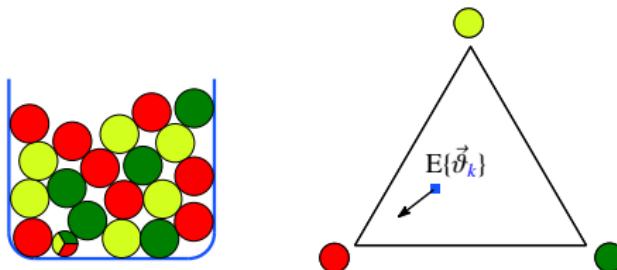


Figure: Pólya urn and discrete parameters.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \mathbf{t}}^{-t_i} + \alpha}{n_{\mathbf{k}}^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u, v\}}{=} \frac{n_{\mathbf{k}, \mathbf{u}}^{-u_i} + \alpha}{n_{\mathbf{k}}^{-u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, \mathbf{v}}^{-v_i} + \alpha + \delta(\mathbf{u} - \mathbf{v})}{n_{\mathbf{k}}^{-v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v})$$

...

# $q$ -functions: Pólya urn and sampling weights

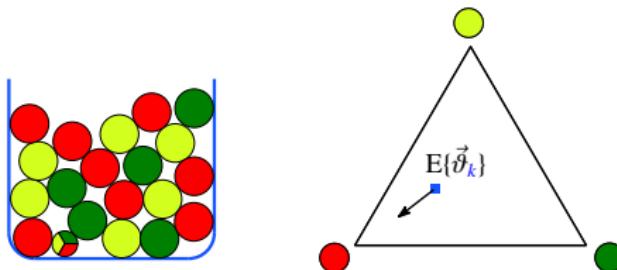


Figure: Pólya urn and discrete parameters.

$$q(\vec{k}, \vec{t}) \triangleq \underbrace{\frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k^{-i} + \alpha)}}_{B\left(\begin{array}{c|ccc} \times 8 & \times 6 & \times 5 & +\alpha \\ \hline \text{red} & \text{yellow} & \text{green} & \text{red} \end{array}\right)} \underset{|t|=1}{=} \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(\vec{k}, \vec{u} \oplus \vec{v})$$

$t_i = \text{green}$

# $q$ -functions: Pólya urn and sampling weights

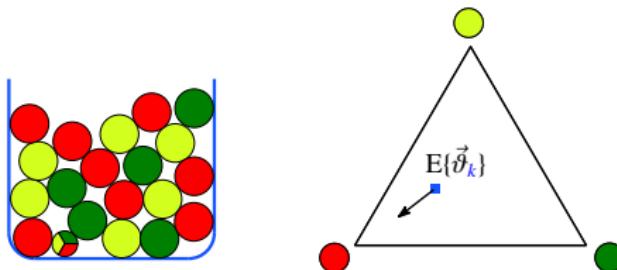


Figure: Pólya urn and discrete parameters.

$$q(\vec{k}, \vec{t}) \triangleq \frac{B(\vec{n}_{\vec{k}} + \alpha)}{B(\vec{n}_{\vec{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \underbrace{\frac{n_{\vec{k}, t}^{-t_i} + \alpha}{n_{\vec{k}}^{-t_i} + T\alpha}}_{\text{smoothed ratio of occurrences}} = \text{smoothed ratio of occurrences}$$

Below this equation is a diagram of a Pólya urn. The urn is represented by a green-bordered box containing four colored balls: red, yellow, green, and blue. A blue curved line labeled  $\times 8 + \alpha$  connects the red ball to the bottom of the box. Another blue curved line labeled  $\times 8 \times 6 \times 4 + \alpha$  connects the red, yellow, green, and blue balls to the bottom of the box. To the right of the urn, there is a fraction:

$$\frac{n_{\vec{k}}^{-v_i} + \alpha + \delta(u - v)}{n_{\vec{k}}^{-v_i} + T\alpha + 1} \triangleq q(\vec{k}, u \oplus v)$$

# $q$ -functions: Pólya urn and sampling weights

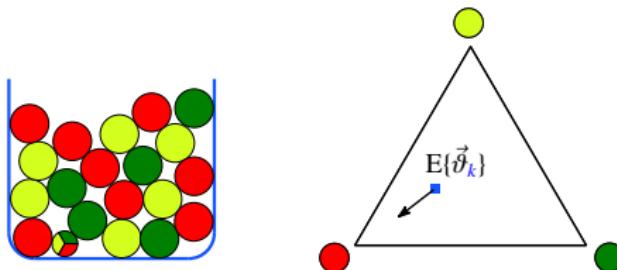


Figure: Pólya urn and discrete parameters.

$$q(\vec{k}, \vec{t}) \triangleq \underbrace{\frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k^{-i} + \alpha)}}_{B\left(\begin{array}{c|ccc} \times 8 & \times 6 & \times 5 & + \alpha \\ \hline \textcolor{red}{\bullet} & \textcolor{yellow}{\bullet} & \textcolor{green}{\bullet} & \textcolor{red}{\bullet} \end{array}\right)} \stackrel{|\vec{t}|=1}{=} \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(\vec{k}, \vec{u} \oplus \vec{v})$$

$t_i = \{u_i, v_i\}$

# $q$ -functions: Pólya urn and sampling weights

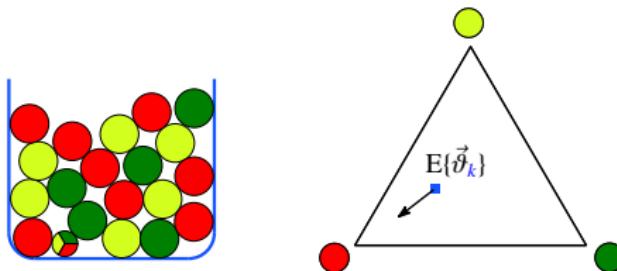


Figure: Pólya urn and discrete parameters.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \mathbf{t}}^{-t_i} + \alpha}{n_{\mathbf{k}}^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u,v\}}{=} \underbrace{\frac{n_{\mathbf{k}, \mathbf{u}}^{-u_i} + \alpha}{n_{\mathbf{k}}^{-u_i} + T\alpha}}_{\dots} \cdot \frac{n_{\mathbf{k}, \mathbf{v}}^{-v_i} + \alpha + \delta(u - v)}{n_{\mathbf{k}}^{-v_i} + T\alpha + 1} \triangleq q(\mathbf{k}, \mathbf{u} \oplus \mathbf{v})$$

A magnifying glass is focused on a portion of the Pólya urn. Inside the magnified view, there are two red balls labeled  $\neg u_i$ . Below the magnifying glass, the entire urn is shown with four colored balls: red, yellow, green, and blue. The red ball is labeled  $\neg u_i$ .

# $q$ -functions: Pólya urn and sampling weights

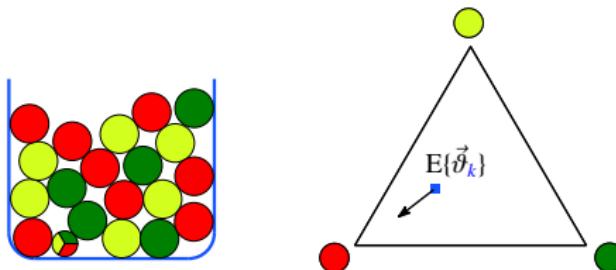


Figure: Pólya urn and discrete parameters.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \textcolor{red}{t}}^{-t_i} + \alpha}{n_{\mathbf{k}}^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u,v\}}{=} \underbrace{\frac{n_{\mathbf{k}, \textcolor{red}{u}}^{-u_i} + \alpha}{n_{\mathbf{k}}^{-u_i} + T\alpha} \cdot \frac{n_{\mathbf{k}, \textcolor{green}{v}}^{-v_i} + \alpha + \delta(u-v)}{n_{\mathbf{k}}^{-v_i} + T\alpha + 1}}_{\dots} \triangleq q(\mathbf{k}, \textcolor{red}{u} \oplus \textcolor{green}{v})$$

The diagram shows a Pólya urn process. At the bottom left, a urn contains balls labeled  $\neg u_i$ ,  $\neg v_i$ , and  $1$ . An arrow points from this state to a state where a ball labeled  $\neg u_i$  has been drawn. At the bottom right, another urn contains balls labeled  $\neg v_i$ ,  $\neg u_i$ , and  $1$ . An arrow points from this state to a state where a ball labeled  $\neg v_i$  has been drawn. This illustrates the discrete nature of the parameters  $u$  and  $v$ .

# $q$ -functions: Pólya urn and sampling weights

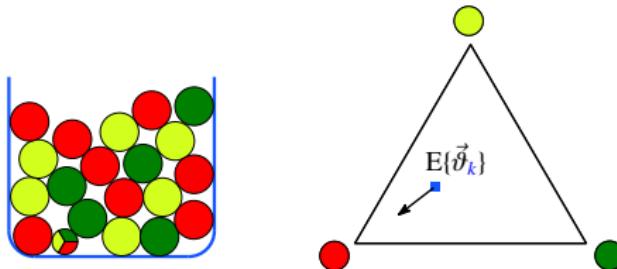


Figure: Pólya urn and discrete parameters.

$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \textcolor{red}{t}}^{-t_i} + \alpha}{n_{\mathbf{k}}^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u,v\}}{=} \underbrace{\frac{n_{\mathbf{k}, \textcolor{red}{u}}^{-u_i} + \alpha}{n_{\mathbf{k}}^{-u_i} + T\alpha}}_{\dots} \cdot \underbrace{\frac{n_{\mathbf{k}, \textcolor{green}{v}}^{-v_i} + \alpha + \delta(u-v)}{n_{\mathbf{k}}^{-v_i} + T\alpha + 1}}_{\stackrel{\textcolor{blue}{u} \neq \textcolor{green}{v}}{= q(\mathbf{k}, \textcolor{red}{u} \oplus \textcolor{green}{v})}}$$

The figure shows two urns at the bottom. The left urn contains balls labeled  $\neg u_i$ ,  $\neg v_i$ , and  $u=v$ . The right urn contains balls labeled  $\neg u_i$ ,  $\neg v_i$ , and  $1$ . Above the urns, arrows point from the corresponding labels to the respective colored vertices (red, green, black) in the triangular parameter space diagram. This illustrates how the observed counts in the urn map to the discrete parameters in the simplex.

# $q$ -functions: Pólya urn and sampling weights

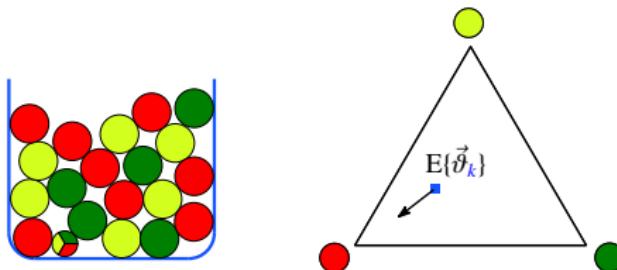


Figure: Pólya urn and discrete parameters.

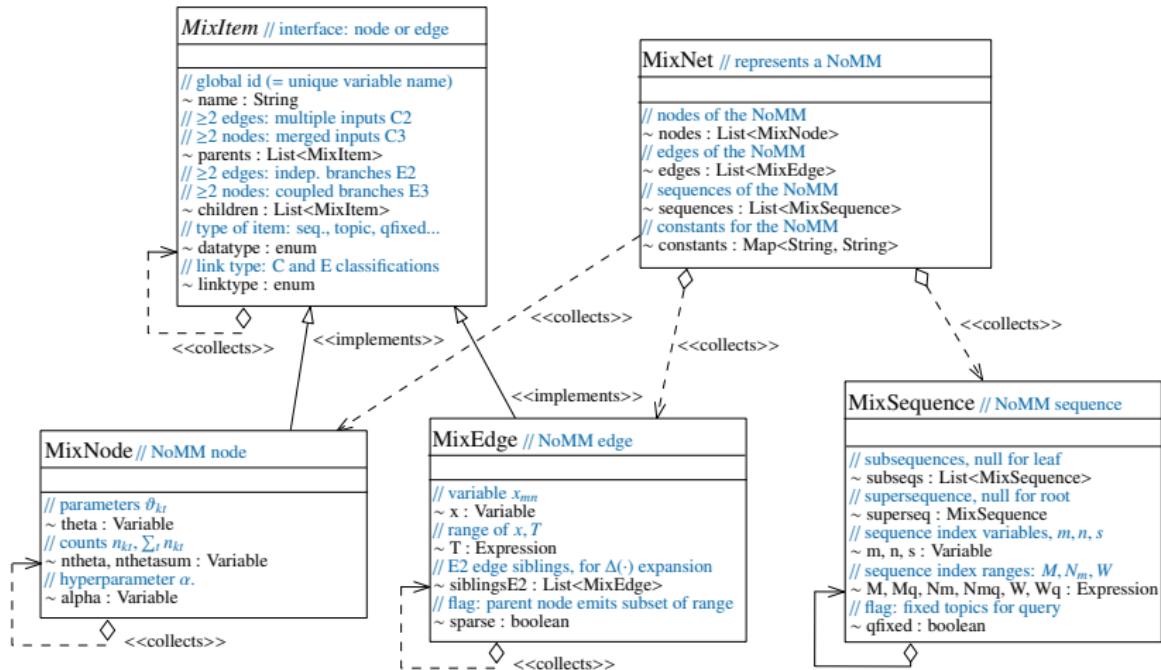
$$q(\mathbf{k}, \mathbf{t}) \triangleq \frac{B(\vec{n}_{\mathbf{k}} + \alpha)}{B(\vec{n}_{\mathbf{k}}^{-i} + \alpha)} \stackrel{|t|=1}{=} \frac{n_{\mathbf{k}, \textcolor{red}{t}}^{-t_i} + \alpha}{n_{\mathbf{k}}^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$
$$\stackrel{t=\{u,v\}}{=} \underbrace{\frac{n_{\mathbf{k}, \textcolor{red}{u}}^{-u_i} + \alpha}{n_{\mathbf{k}}^{-u_i} + T\alpha}}_{\dots} \cdot \underbrace{\frac{n_{\mathbf{k}, \textcolor{green}{v}}^{-v_i} + \alpha + \delta(u-v)}{n_{\mathbf{k}}^{-v_i} + T\alpha + 1}}_{\stackrel{\mathbf{k} \in \{u,v\}}{=}} \triangleq q(\mathbf{k}, \textcolor{red}{u} \oplus \textcolor{green}{v})$$

Below the equations are two diagrams of a Pólya urn. The left diagram shows a urn with four balls: one red ( $\neg u_i$ ), one yellow ( $\neg u_i$ ), one green ( $\neg v_i$ ), and one blue ( $\neg v_i$ ). The right diagram shows a similar urn with five balls: one red ( $\neg u_i$ ), one yellow ( $\neg u_i$ ), one green ( $\neg v_i$ ), one blue ( $\neg v_i$ ), and one red ( $u=v$ ). Below the urns is the expression  $q(\mathbf{k}, \textcolor{red}{u} \oplus \textcolor{green}{v})$ .

# NoMM substructure library: Gibbs weights and likelihood

ID. Name	Structure diagram	Gibbs sampler weight $w$ , Likelihood $p$ for single token Modelled aspect, example models
N1.E1.C1, Dir-Mult nodes, unbranched		$w(z) = q(a, z)q(z, b)$ $p(b a) = \sum_i \theta_{a,i} \theta_{z,i}$ <i>Mixture/adjustmixture: LDA [Blei et al. 2003b], PAM [Li &amp; McCullum 2006]; LDCC [Shafiei &amp; Milios 2006] (E1S)</i>
N2. Observed parameters		$w(z \vec{\theta}_a) = \vec{\theta}_{a,z}$ $p(b a) = \sum_i \theta_{a,i} \theta_{z,i}$ <i>Label distribution: ATM [Rosen-Zvi et al. 2004]</i>
N3. Non- Dirichlet prior		$w(z \vec{\theta}_a) = p(z \vec{\theta}_a)q(z, b)$ $p(b a) = \sum_i \theta_{a,i} \theta_{z,i}$ <i>M-step: estimate <math>\vec{\theta}_a</math> [Blei &amp; Lafferty 2007]</i>
N4. Non- discrete output		$w(z \vec{\theta}_a) = q(a, z)p(z \vec{\theta}_a)$ ; M-step: estimate $\vec{\theta}_a$ $p(v a) = \sum_i \theta_{a,i} p(v_i \vec{\theta}_a)$ <i>Non-mutualinfo obser.: Corr-LDA [Barnard et al. 2003], GMM [McLachlan &amp; Peel 2000]; <math>p(v \theta) = \mathcal{N}(v \vec{\mu}, \frac{1}{\vec{\sigma}^2})</math></i>
N5+E4. Aggregation		$w(z \vec{\theta}_a) = q(a, z)q(z, w)\mathcal{N}(v_m \vec{\theta}_{z,w}, \sigma^2)$ ; M-step: estimate $\vec{\theta}_a$ $p(b a) = \vec{\theta}_{a,b}^T \vec{\theta}_a$ (for linear regression, NSB) <i>Prediction: <math>v_m = \vec{\theta}_a^T \vec{\theta}_m</math></i> <i>Regression/supervised learning: Supervised LDA [Blei &amp; McAuliffe 2007], Relational topic model [Chang &amp; Blei 2009]</i>
E2. Autonomous edges		$w(x,y) = q(a, x\oplus y)q(x, b)q(y, c)$ $p(b, c a) = \sum_i \theta_{a,i} \theta_{x,i} \theta_{y,i}$ <i>Common mixture of causes: Multimodal LDA [Ramage et al. 2009]</i>
E3. Coupled edges		$w(z) = q(a, z)q(z, b)q(z, c)$ $p(b, c a) = \sum_i \theta_{a,i} \theta_{z,i}$ <i>Common cause for observations: Hidden relational model (HRM) [Xu et al. 2006], Link-LDA [Erosheva et al. 2004]</i>
C2. Combined indices		$w(x, y) = q(a, x\oplus y)q(x, k)$ $p(c x, b) = \sum_i \theta_{a,i} \theta_{x,i} \theta_{y,i}$ , $k = g(x, y, i, j)$ <i>Different dependent causes, relation: hPAM [Li et al. 2007a], HRM [Xu et al. 2006], Multi-LDA [Porteous et al. 2008a]</i>
C3. Interleaved indices		$w(z, s) = q(a, z)(q(b, 1)q(z, c))^{(k-1)} \cdot (q(b, 2)q(z, d))^{(l-p-2)}$ $p(c, d a, b) = \sum_i \theta_{a,i} (\theta_{b,i}\theta_{c,i} + \theta_{b,i}\theta_{d,i})$ <i>Select complex submodels: Multi-grain LDA [Titov &amp; McDonald 2008], Entity-topic models [Newman et al. 2006a]</i>
C5. Node coupling	<p>Gregor Heinrich</p>	

# Gibbs meta-sampler: Java data structure



```

1  /** run the main Gibbs sampling kernel */
2  public void run(int niter) {
3
4  // iteration loop
5  for (int iter = 0; iter < niter; iter++) {
6
7  // major loop, sequence [m][n]
8  for (int m = 0; m < M; m++) {
9    // component selectors
10   int mxsel = -1;
11   int mxjsel = -1;
12   int ksel = -1;
13
14  // minor loop, sequence [m][n]
15  for (int n = 0; n < w[m].length; n++) {
16    double psum;
17    double u;
18    // decrement counts
19    nmx[m][x[m][n]]--;
20    mxsel = X * m + x[m][n];
21    nmxy[mxsel][y[m][n]]--;
22    nmxysum[mxsel]--;
23    if (x[m][n] == 0)
24      ksel = 0;
25    else if (y[m][n] == 0)
26      ksel = 1 + x[m][n];
27    else
28      ksel = 1 + X + y[m][n];
29    nkw[ksel][w[m][n]]--;
30    nkwsum[ksel]--;
31
32  // compute weights
33  /* p(x_{-m,n} \eq x, y_{-m,n} \eq y ... (LaTeX omitted) */
34  psum = 0;
35  int hx = -1;
36  int hy = -1;
37  // hidden edge
38  for (hx = 0; hx < X; hx++) {
39    // hidden edge
40    for (hy = 0; hy < Y; hy++) {
41      mxsel = X * m + hx;
42      mxjsel = hx;
43      if (hx == 0)
44        ksel = 0;
45      else if (hy == 0)
46        ksel = 1 + hx;
47      else
48        ksel = 1 + X + hy;

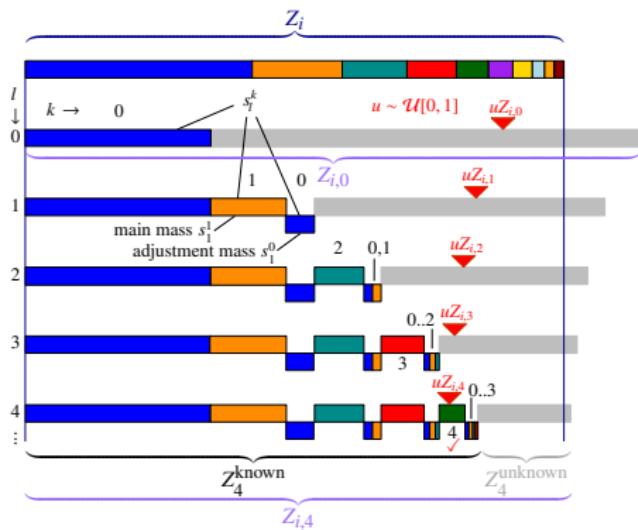
```

```

49
50  pp[hx][hy] = (nmx[m][hx] + alpha[hx])
51  * (nmxy[mxsel][hy] + alphax[mxjsel][hy])
52  / (nmxysum[mxsel] + alphaxsum[mxjsel])
53  * (nkw[ksel][w[m][n]] + beta)
54  / (nkwsum[ksel] + betasum);
55  psum += pp[hx][hy];
56  } // for h
57 } // for h
58
59 // sample topics
60 u = rand.nextDouble() * psum;
61 psum = 0;
62 SAMPLED:
63 // each edge value
64 for (hx = 0; hx < X; hx++) {
65   // each edge value
66   for (hy = 0; hy < Y; hy++) {
67     psum += pp[hx][hy];
68     if (u <= psum)
69       break SAMPLED;
70   } // h
71 } // h
72
73 // assign topics
74 x[m][n] = hx;
75 y[m][n] = hy;
76
77 // increment counts
78 nmx[m][x[m][n]]++;
79 mxsel = X * m + x[m][n];
80 nmxy[mxsel][y[m][n]]++;
81 nmxysum[mxsel]++;
82 if (x[m][n] == 0)
83   ksel = 0;
84 else if (y[m][n] == 0)
85   ksel = 1 + x[m][n];
86 else
87   ksel = 1 + X + y[m][n];
88 nkw[ksel][w[m][n]]++;
89 nkwsum[ksel]++;
90 } // for n
91 } // for m
92
93 // estimate hyperparameters
94 estAlpha();
95 } // for iter
96 } // run()

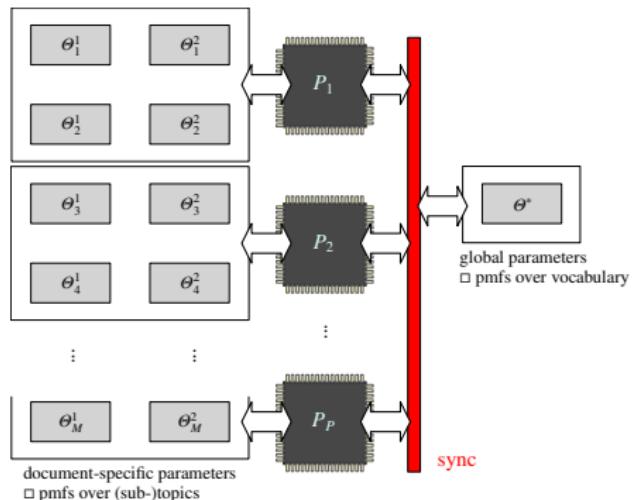
```

# Fast serial sampling: Using a normalisation bound

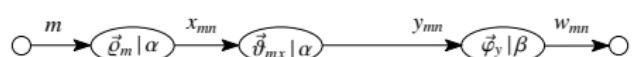


- Idea: Exploit saliency of few elements → compute only largest (=most likely) weights
- Approximate normalisation via vector norms (Porteous et al. 2008)
- Generalisation to multiple dependent variables: more expensive higher-order vector norms ↔ higher sparsity of sampling space

# Fast parallel sampling: Synchronisation methods



- Multi-processor parallelisation using shared memory (OpenMP)
- Main challenge: synchronisation and communication of global data
- Synchronisation methods (LDA + generic NoMMs):
  - a. Naïve synchronisation locks
  - b. Query read-only  $\varphi$  + MAP update step for  $\varphi$  ("split-state")
  - c. Local copies  $\varphi$  + reduction step (=AD-LDA (Newman et al. 2009))



# Fast sampling: Serial × parallel

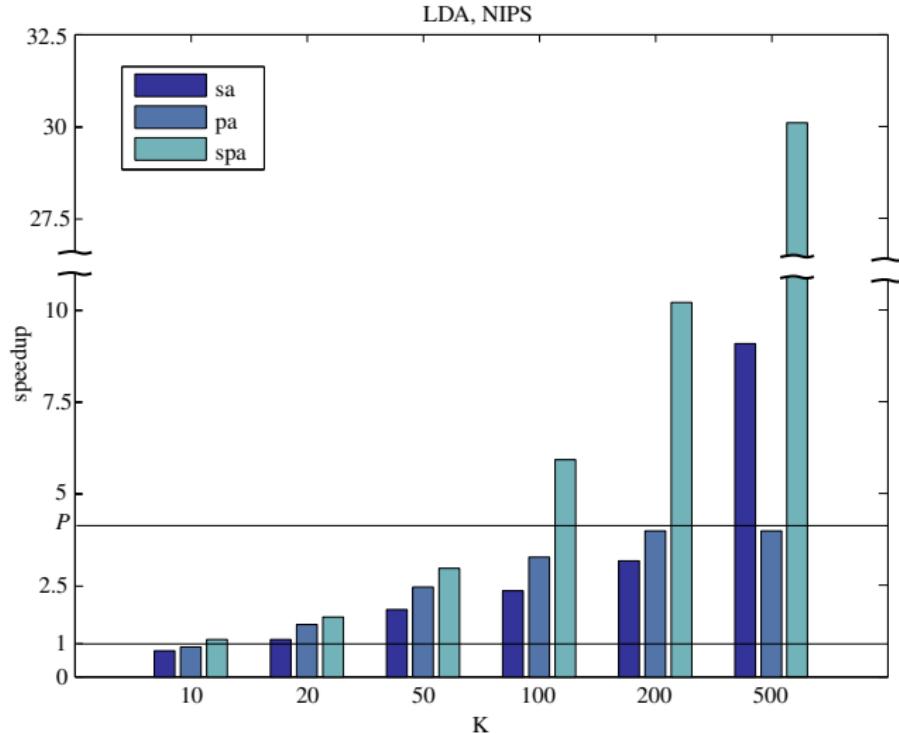
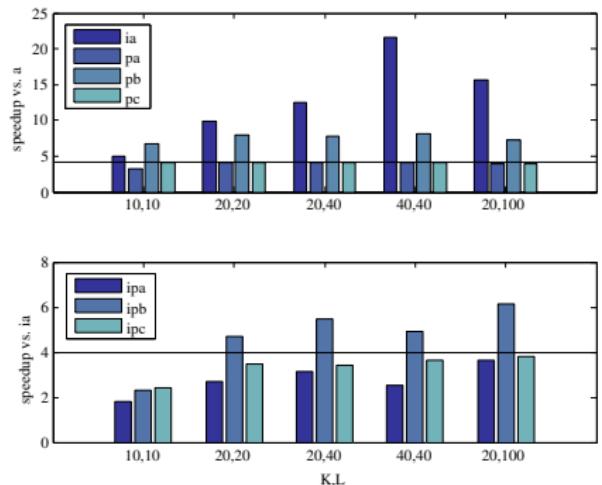
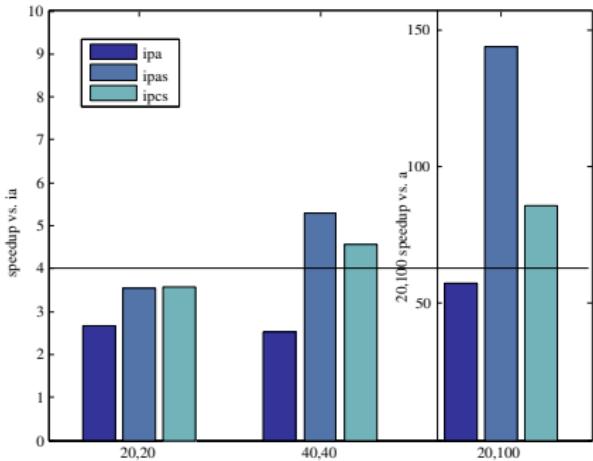


Figure: Speed-up for fast sampling methods: LDA.

# Fast sampling: Serial $\times$ parallel $\times$ independent



(a) Parallel, independent



(b) Parallel, serial, independent

Figure: Speed-up for combined fast samplers: PAM4 (2 dependent variables).

# Fast sampling: The impact of assumed independence

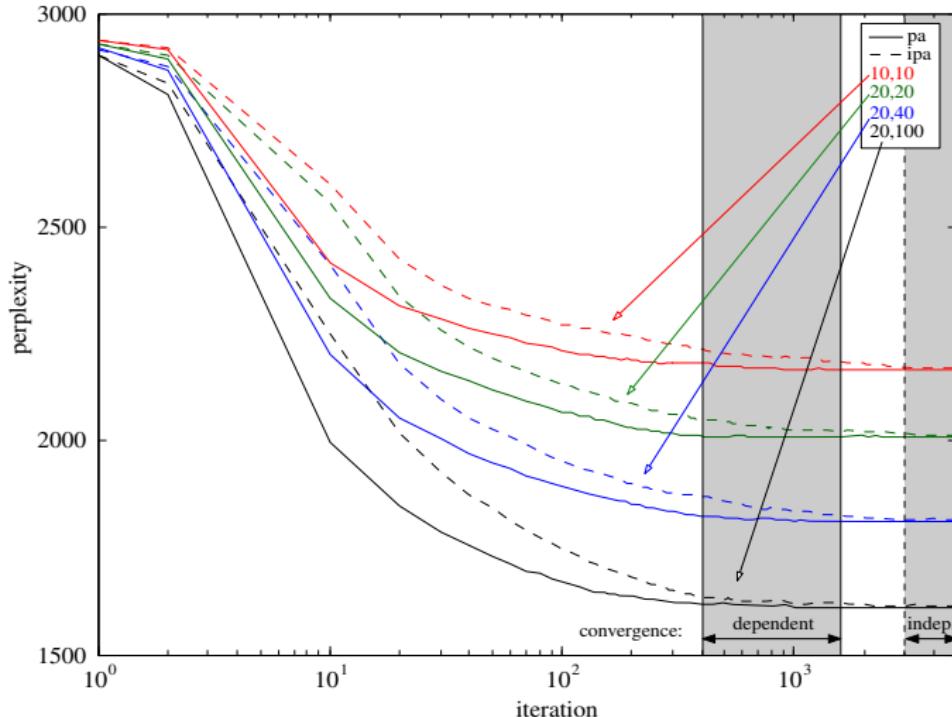
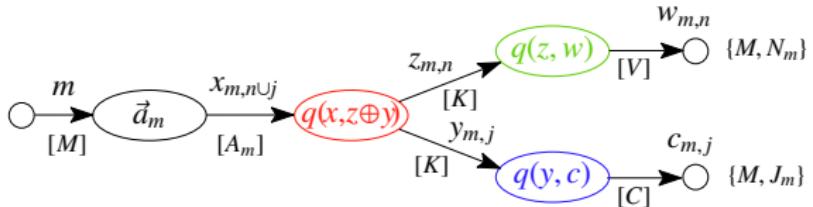


Figure: Perplexity over iterations. Example model: PAM4.

# ETT1 model: Derivation using NoMM structure



Lining up  $q$ -functions:

$$p(x, z, y | \cdot) \propto a_{m,x} q(x, z \oplus y) q(z, w) q(y, c) \quad (13)$$

Transforming to standard Gibbs *full conditionals*:

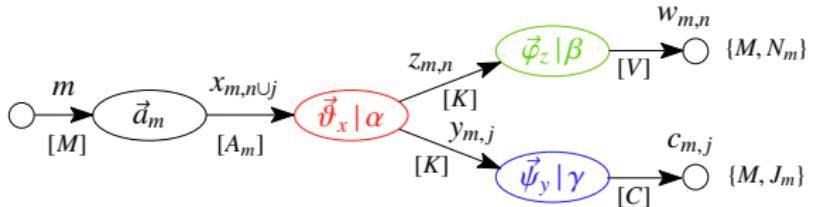
$$p(x_{m,n}=x, z_{m,n}=z | \cdot) \propto a_{m,x} \cdot \frac{n_{x,z}^{\neg\{x,z\}_{m,n}} + \alpha}{n_x^{\neg\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n_{z,w}^{\neg z_{m,n}} + \beta}{n_z^{\neg z_{m,n}} + V\beta} \quad (14)$$

$$p(x_{m,j}=x, y_{m,j}=y | \cdot) \propto a_{m,x} \cdot \frac{n_{x,y}^{\neg\{x,y\}_{m,j}} + \alpha}{n_y^{\neg\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n_{y,c}^{\neg y_{m,j}} + \gamma}{n_y^{\neg y_{m,j}} + C\gamma} \quad (15)$$

Retrieval über Anfrage-Likelihood-Modell:

$$p(\vec{w} | a) = \prod_{w \in \vec{w}} \sum_z \vartheta_{a,z} \varphi_{z,w} \quad p(\vec{c} | a) = \prod_{c \in \vec{c}} \sum_y \vartheta_{a,y} \psi_{y,c} . \quad (16)$$

# ETT1 model: Derivation using NoMM structure



Lining up  $q$ -functions:

$$p(x, z, y | \cdot) \propto a_{m,x} q(x, z \oplus y) q(z, w) q(y, c) \quad (13)$$

Transforming to standard Gibbs *full conditionals*:

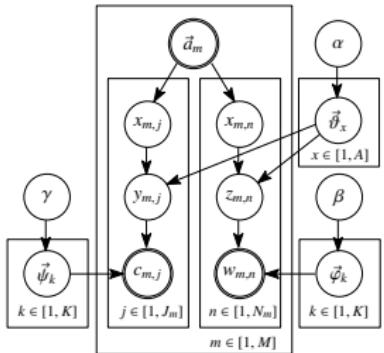
$$p(x_{m,n}=x, z_{m,n}=z | \cdot) \propto a_{m,x} \cdot \frac{n_{x,z}^{\neg\{x,z\}_{m,n}} + \alpha}{n_x^{\neg\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n_{z,w}^{\neg z_{m,n}} + \beta}{n_z^{\neg z_{m,n}} + V\beta} \quad (14)$$

$$p(x_{m,j}=x, y_{m,j}=y | \cdot) \propto a_{m,x} \cdot \frac{n_{x,y}^{\neg\{x,y\}_{m,j}} + \alpha}{n_y^{\neg\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n_{y,c}^{\neg y_{m,j}} + \gamma}{n_y^{\neg y_{m,j}} + C\gamma} \quad (15)$$

Retrieval über Anfrage-Likelihood-Modell:

$$p(\vec{w} | a) = \prod_{w \in \vec{w}} \sum_z \vartheta_{a,z} \varphi_{z,w} \quad p(\vec{c} | a) = \prod_{c \in \vec{c}} \sum_y \vartheta_{a,y} \psi_{y,c} . \quad (16)$$

# ETT1 model: Derivation using ordinary method (excerpt)



(a) Expert–tag–topic model (ETT)  
(Heinrich 2011b)

$$\begin{aligned}
 p(\vec{w}, \vec{c}, \vec{d}, \vec{x}, \vec{z}, \underline{\theta}, \underline{\Phi}, \underline{\Psi} | \alpha, \beta, \gamma) &= p(\vec{w}|\vec{z}, \underline{\Phi})p(\underline{\Phi}|\beta) \cdot p(\vec{c}|\vec{y}, \underline{\Psi})p(\underline{\Psi}|\gamma) \\
 &\quad \cdot p(\vec{y}|\vec{x}, \underline{\Theta})p(\vec{z}|\vec{x}, \underline{\Theta})p(\underline{\Theta}|\alpha) \cdot p(\vec{x}|\vec{d}) \\
 &= \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{z}_{m,n}) p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,x_{m,n}} \right. \\
 &\quad \cdot \left. \prod_{j=1}^{J_m} p(c_{m,j} | \vec{\psi}_{y_{m,j}}) p(y_{m,j} | \vec{\theta}_{x_{m,j}}) a_{m,x_{m,j}} \right) \\
 &\quad \cdot p(\underline{\Theta}|\alpha) \cdot p(\underline{\Phi}|\beta) \cdot p(\underline{\Psi}|\gamma). \tag{E.1}
 \end{aligned}$$

(E.2)

$$\begin{aligned}
 p(\vec{w}, \vec{c}, \vec{d}, \vec{x}, \vec{z} | \alpha, \beta, \gamma) &= \int \int \int \prod_{m=1}^M \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{z}_{m,n}) p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,x_{m,n}} \right. \\
 &\quad \cdot \left. \prod_{j=1}^{J_m} p(c_{m,j} | \vec{\psi}_{y_{m,j}}) p(y_{m,j} | \vec{\theta}_{x_{m,j}}) a_{m,x_{m,j}} \right) \\
 &\quad \cdot d\underline{\theta}|\alpha) \cdot d\underline{\Phi}|\beta) \cdot d\underline{\Psi}|\gamma) \tag{E.3}
 \end{aligned}$$

$$\begin{aligned}
 &= \int \prod_{m=1}^M \prod_{n=1}^{N_m} p(w_{m,n} | \vec{z}_{m,n}) \prod_{k=1}^K p(\vec{\varphi}_k | \beta) d\varphi_k \\
 &\quad \cdot \int \prod_{m=1}^M \prod_{j=1}^{J_m} p(c_{m,j} | \vec{\psi}_{y_{m,j}}) \prod_{k=1}^K p(\vec{\varphi}_k | \beta) d\vec{\varphi}_k \\
 &\quad \cdot \int \prod_{m=1}^M p(\underline{\Theta}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n} | \vec{\theta}_{x_{m,n}}) a_{m,x_{m,n}} \prod_{j=1}^{J_m} p(y_{m,j} | \vec{\theta}_{x_{m,j}}) a_{m,x_{m,j}} d\vec{\theta}_m \tag{E.4}
 \end{aligned}$$

$$\begin{aligned}
 &= \int \prod_{k=1}^K \frac{1}{\Delta_V(\beta)} \prod_{l=1}^V \varphi_{k,l}^{n_{k,l} + \beta - 1} d\varphi_{k,l} \cdot \int \prod_{k=1}^K \frac{1}{\Delta_C(\gamma)} \prod_{c=1}^C \psi_{k,c}^{n_{k,c} + \gamma - 1} d\vec{\varphi}_k \\
 &\quad \cdot \int \prod_{a=1}^A \frac{1}{\Delta_K(\alpha)} \prod_{k=1}^K \theta_{a,k}^{n_{a,k}^{(z)} + n_{a,k}^{(y)} + \alpha - 1} d\vec{\theta}_a \cdot \prod_{m=1}^M \prod_{a=1}^A \frac{n_{m,a}^{(z)} + n_{m,a}^{(y)}}{a_{m,a}} \tag{E.5}
 \end{aligned}$$

$$= \prod_{k=1}^K \frac{\Delta(\vec{n}_k^{(z)} + \beta)}{\Delta_V(\beta)} \cdot \frac{\Delta(\vec{n}_k^{(y)} + \gamma)}{\Delta_C(\gamma)} \prod_{a=1}^A \frac{\Delta(n_{a,k}^{(z)} + n_{a,k}^{(y)} + \alpha)}{\Delta_K(\alpha)} \prod_{m=1}^M \frac{n_{m,a}^{(z)} + n_{m,a}^{(y)}}{a_{m,a}}. \tag{E.6}$$

$$p(z_i=k, x_i=x | w_i=i, \vec{z}_{-i}, \vec{y}, \vec{x}_{-i}, \vec{w}_{-i}, \vec{d}, \vec{c})$$

$$= \frac{p(\vec{w}, \vec{z}, \vec{y}, \vec{x})}{p(\vec{w}, \vec{z}_{-i}, \vec{y}, \vec{x}_{-i})} = \frac{p(\vec{w}|\vec{z}, \vec{y})}{p(\vec{w}_{-i}|\vec{z}_{-i}, \vec{y})} \cdot \frac{p(\vec{z}|\vec{x})}{p(\vec{z}_{-i}|\vec{x}_{-i})} \cdot \frac{p(\vec{x})}{p(\vec{x}_{-i})} \tag{E.7}$$

$$\propto \frac{\Delta(\vec{n}_k^{(z)} + \beta)}{\Delta(\vec{n}_{k,-i}^{(z)} + \beta)} \cdot \frac{\Delta(\vec{n}_x + \alpha)}{\Delta(\vec{n}_{x,-i} + \alpha)} \cdot a_{m,x} \tag{E.8}$$

$$= \frac{\Gamma(n_{k,i} + \beta) \Gamma(n_{k,-i} + V\beta)}{\Gamma(n_{k,i,-i} + \beta) \Gamma(n_k + V\beta)} \cdot \frac{\Gamma(n_{x,k}^{(z)} + \alpha) \Gamma(n_{x,-i}^{(z)} + K\alpha)}{\Gamma(n_{x,k,-i}^{(z)} + \alpha) \Gamma(n_x^{(z)} + K\alpha)} \cdot a_{m,x} \tag{E.9}$$

$$= \frac{n_{k,i,-i} + \beta}{n_{k,-i} + V\beta} \cdot \frac{n_{x,k,-i}^{(z)} + \alpha}{n_{x,-i}^{(z)} + K\alpha} \cdot a_{m,x} \tag{E.10}$$

$$p(y_i=k, x_i=x | c_i=c, \vec{z}_{-i}, \vec{y}_{-i}, \vec{x}_{-i}, \vec{w}, \vec{d}, \vec{c}_{-i}) \propto \frac{n_{k,c,-i} + \gamma}{n_{k,-i} + V\gamma} \cdot \frac{n_{x,k,-i}^{(y)} + \alpha}{n_{x,-i}^{(y)} + K\alpha} \cdot a_{m,x} \tag{E.12}$$

# ETT1 evaluation: Truncated Average Precision

1.      2.      3.      4.      5.



$$AP@5 = \frac{1/2 + 2/4 + 3/5}{3} = 0.533$$



$$AP@5 = \frac{1/1 + 2/2 + 3/5}{3} = 0.867$$

Figure: Average Precision at 5 (assuming 3 relevant documents in corpus)

# ETT1 results: Term Retrieval

query: svm support vector machine | kernel classifier hyperplane regression

- 
- 1. Scholkopf\_B**, lik = -76.272, tokens = 2830, docs = 10: judged relevant  
✓ From Regularization Operators to *Support Vector Kernels* (9); Improving the Accuracy and Speed of *Support Vector Machines* (9); Shrinking the Tube: A New *Support Vector Regression* Algorithm (11) ...
- 
- 2. Smola\_A**, lik = -77.509, tokens = 2760, docs = 11: judged relevant  
✓ *Support Vector Regression* Machines (9); Prior Knowledge in *Support Vector Kernels* (10); *Support Vector* Method for Novelty Detection (12)  
The Entropy Regularization Information Criterion (12, *support vector machines, regularization*) ...
- 
- 3. Vapnik\_V**, lik = -77.525, tokens = 2332, docs = 10: judged relevant  
✓ *Support Vector Regression* Machines (9); Prior Knowledge in *Support Vector Kernels* (10); Prior Knowledge in *Support Vector Kernels* (10);  
*Support Vector* Method for Multivariate Density Estimation (12); ...
- 
- 4. Crisp\_D**, lik = -81.401, tokens = 699, docs = 2: judged relevant  
✓ A Geometric Interpretation of t-/SVM Classifiers (12); Uniqueness of the SVM Solution (12)
- 
- 5. Burges\_C**, lik = -81.630, tokens = 1309, docs = 5: judged relevant  
✓ Improving the Accuracy and Speed of *Support Vector Machines* (9); A Geometric Interpretation of t-/SVM Classifiers (12); Uniqueness of the SVM Solution (12) ...
- 
- 6. Laskov\_P**, lik = -84.275, tokens = 738, docs = 1: judged relevant  
✓ An Improved Decomposition Algorithm for *Regression Support Vector Machines* (12)
- 
- 7. Steinage\_V**, lik = -84.600, tokens = 438, docs = 1: judged irrelevant  
✗ Nonlinear Discriminant Analysis Using *Kernel Functions* (12)
- 
- 8. Bennett\_K**, lik = -86.754, tokens = 384, docs = 1: judged relevant  
✓ Semi-Supervised *Support Vector Machines* (11)
- 
- 9. Herbrich\_R**, lik = -86.754, tokens = 462, docs = 2: judged irrelevant  
✗ Classification on Pairwise Proximity Data (11); Bayesian Transduction (12, *classification*)
- 
- 10. Chapelle\_O**, lik = -87.431, tokens = 494, docs = 2: judged relevant  
✓ Model Selection for *Support Vector Machines* (12); Transductive Inference for Estimating Values of Functions (12, *regression, classification*)

# ETT1 results: Tag retrieval

query: *face recognition*

- 
- 1. Movellan.J**, lik = -4.680, tokens = 3153, docs = 8: judged relevant  
✓ Dyn. Features for Visual Speechreading: A System Comparison (9, no tags); Image Representation for Facial Expression Coding (12, tags: *face recognition, image, ICA*); Visual Speech Recognition with Stochastic Networks (7, tags: *HMM, speech recognition*) ...
- 
- 2. Bartlett.M**, lik = -4.951, tokens = 812, docs = 3: judged relevant  
✓ Viewpoint Invariant *Face Recognition* using ICA and Attractor Networks (9, tags: *face recognition, invariances, pattern recognition*); Image Representation for Facial Expression Coding (12, tags: *face recognition, image, ICA*) ...
- 
- 3. Dailey.M**, lik = -4.952, tokens = 903, docs = 2: judged relevant  
✓ Task and Spatial Frequency Effects on Face Specialization (10, tags: *face recognition*); Facial Memory Is Kernel Density Estimation (Almost) (11, no tags)
- 
- 4. Padgett.C**, lik = -4.974, tokens = 499, docs = 1: judged relevant  
✓ Representing Face Images for Emotion Classification (9, tags: *classification, face recognition, image*)
- 
- 5. Hager.J**, lik = -5.023, tokens = 377, docs = 2: judged relevant  
✓ Classifying Facial Action (8, tags: *classification*); Image Representation for Facial Expression Coding (12, tags: *face recognition, image, ICA*)
- 
- 6. Ekman.P**, lik = -5.027, tokens = 374, docs = 2: judged relevant  
✓ Image Representation for Facial Expression Coding (12, tags: *face recognition, image, ICA*); Classifying Facial Action (8, tags: *classification*)
- 
- 7. Phillips.P**, lik = -5.127, tokens = 795, docs = 1: judged relevant  
✓ Support Vector Machines Applied to Face Recognition (11, tags: *face recognition, SVM*)
- 
- 8. Gray.M**, lik = -5.159, tokens = 470, docs = 2: judged irrelevant  
✗ Dynamic Features for Visual Speechreading: A Systematic Comparison (9, text: dynamic visual features; no tags)
- 
- 9. Lawrence.D**, lik = -5.217, tokens = 265, docs = 1: judged relevant  
✓ SEXNET: A Neural Network Identifies Sex From Human Faces (3, tags: *neural networks, object recognition, pattern recognition*)

# ETT1 results: Tag query and expert topics

---

## tag: *face recognition* (ETT1/J20)

---

0.82702 face images faces image facial visual human video database detection  
0.09392 image images texture pixel resolution pyramid regions pixels region search  
0.02696 wavelet video view images tracking user camera image motion shape  
0.00117 eeg brain ica artifacts subjects activity subject erp signals scalp  
0.00100 image images visual vision optical pixel surface edge disparity receptive  
0.00094 orientation cortical dominance ocular cortex development lateral eye cells visual  
0.00089 chip neuron synapse digital pulse analog synaptic chips synapses murray  
0.00084 hinton object image energy cost images code visible zemel codes

---

## author: Movellan\_J (ETT1/J20)

---

0.53816: face images faces image facial visual human video database detection  
0.16216: image images texture pixel resolution pyramid regions pixels region search  
0.08954: speech speaker acoustic vowel phonetic phoneme utterances spoken formant  
0.06216: bayesian prior density posterior entropy evidence likelihood distributions  
0.03939: filter frequency signals phase channel amplitude frequencies temporal spectrum  
0.03508: activation boltzmann annealing temperature neuron stochastic schedule machine  
0.02770: cell firing cells neuron activity excitatory inhibitory synaptic potential membrane  
0.02154: convergence stochastic descent optimization batch density global update

---

## author: Cottrell\_G (ETT1/J20)

---

0.41865: recurrent nets correlation cascade activation connection epochs representations  
0.27523: face images faces image facial visual human video database detection  
0.17531: subjects human stimulus cue subject trials experiment perceptual psychophysical  
0.11287: tangent transformation image simard images invariant invariance euclidean  
0.07130: modules attractors cortex phase olfactory frequency bulb activity oscillatory eeg  
0.06143: word connectionist representations words activation production cognitive musical  
0.03695: node activation graph cycle nets message recurrence links connection child  
0.02049: visual attention contour search selective orientation iiiii region saliency segment

---

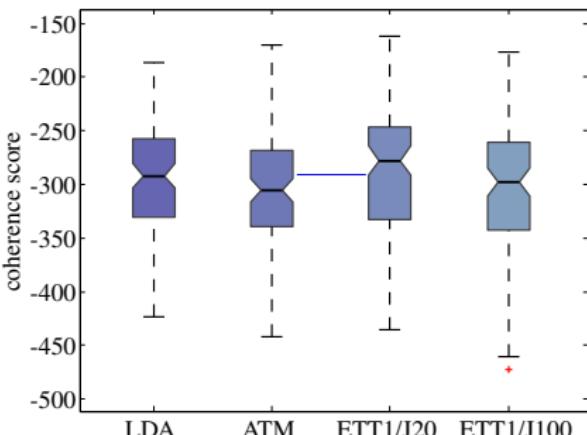
# ETT1 results: Topic coherence

Topic coherence (Mimno et al. 2011):

- $\approx$  How often do top-ranked topic terms co-occur in documents?
- Re-enacts human judgement in topic intrusion experiments (Chang et al. 2009; Heinrich 2011b)

Words in topic (choose worst match (A-F) in every group):					
1. A. orientation	2. A. likelihood	3. A. risk			
B. cortex	B. mixture	B. return			
C. visual	C. theorem	C. stock			
D. ocular	D. density	D. trading			
E. acoustic	E. em	E. processor			
F. eye	F. prior	F. prediction			
4. A. language	5. A. circuit	6. A. validation			
B. word	B. bayesian	B. set			
C. stress	C. analog	C. variance			
D. grammar	D. voltage	D. regression			
E. neural	E. vlsi	E. selection			
F. syllable	F. chip	F. bias			

(a) Topic intrusion experiment



(b) Coherence scores

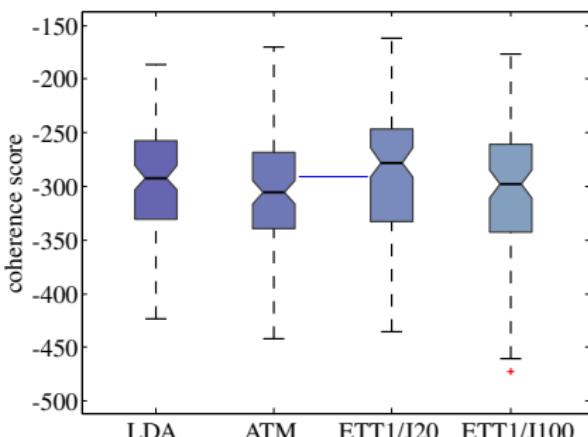
# ETT1 results: Topic coherence

Topic coherence (Mimno et al. 2011):

- $\approx$  How often do top-ranked topic terms co-occur in documents?
- Re-enacts human judgement in topic intrusion experiments (Chang et al. 2009; Heinrich 2011b)

Words in topic (choose worst match (A-F) in every group):					
1. A. orientation	2. A. likelihood	3. A. risk			
B. cortex	B. mixture	B. return			
C. visual	<b>C. theorem</b>	C. stock			
D. ocular	D. density	D. trading			
<b>E. acoustic</b>	E. em	<b>E. processor</b>			
F. eye	F. prior	F. prediction			
4. A. language	5. A. circuit	6. A. validation			
B. word	<b>B. bayesian</b>	<b>B. set</b>			
C. stress	C. analog	C. variance			
D. grammar	D. voltage	D. regression			
<b>E. neural</b>	E. vlsi	E. selection			
F. syllable	F. chip	F. bias			

(a) Topic intrusion experiment



(b) Coherence scores