Gibbs Sampling Algorithms for some Bayesian Network Models

XU Shuo
Information Technology Supporting Center
Institute of Scientific and Technical Information of China (ISTIC)
No. 15 Fuxing Rd., Haidian District, Beijing 100038, China

June 15, 2012

Contents

1 LDA Model

Latent Dirichlet Allocation (LDA) by Blei et al. [1, 2] is a probabilistic generative model that can be used to estimate the properties of multinomial observations by unsupervised learning.

1.1 Notation & Generative Process

The notation is summarized in Table 1, and the graphical model representations of the LDA model is shown in Figure 1.

Table 1: Notation used in the LDA model.

Symbol	Description
K	number of topics / mixture components (const scalar)
M	number of documents (const scalar)
V	number of unique words (const scalar)
N_m	number of word tokens in document m (const scalar)
$\vec{\vartheta}_m$	the multinomial distribution of topics specific to the document m. One proportion for each document, $\Theta = \{\vec{\vartheta}_m\}_{m=1}^{M}$ ($M \times K$ matrix)
$\vec{\varphi}_k$	the multinomial distribution of words specific to the topic k. One component for each topic, $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$ ($K \times V$ matrix)
$z_{m,n}$	the topic associated with the n-th token in the document m
$w_{m,n}$	the n-th token in document m
$\vec{\alpha}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\vartheta}$ (K-vector or scalar if symmetric)
$\vec{\beta}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\varphi}$ (V-vector or scalar if symmetric)

The LDA can be viewed a generative process, which can be described as follows.

1. For each topic $k \in [1, K]$:
   (a) Draw a multinomial $\vec{\varphi}_k$ from a Dirichlet prior $\vec{\beta}$;
2. For each document $m \in [1, M]$:
   (a) Draw a multinomial $\vec{\vartheta}_m$ from a Dirichlet prior $\vec{\alpha}$;
   (b) For each word $n \in [1, N_m]$ in document m:
      i. Draw a topic $z_{m,n}$ from multinomial $\vec{\vartheta}_m$;
      ii. Draw a word $w_{m,n}$ from multinomial $\vec{\varphi}_{z_{m,n}}$;

As shown in the above process, the posterior distribution of topics depends on the information from the text. The parameterization of the LDA model is

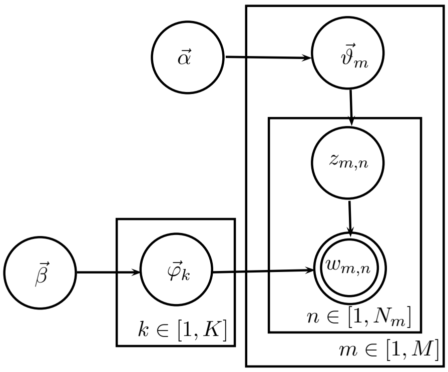

Figure 1: The Graphical Model Representation of the LDA Model

$$\begin{aligned}
\vec{\vartheta}_m|\vec{\alpha} &\sim \text{Dirichlet}(\vec{\alpha}) \\
\vec{\varphi}_k|\vec{\beta} &\sim \text{Dirichlet}(\vec{\beta}) \\
z_{m,n}|\vec{\vartheta}_m &\sim \text{Multinomial}(\vec{\vartheta}_m) \\
w_{m,n}|\vec{\varphi}_{z_{m,n}} &\sim \text{Multinomial}(\vec{\varphi}_{z_{m,n}})
\end{aligned}$$

1.2 Gibbs Sampling Derivation

Although LDA is still a relatively simple model, exact inference is generally intractable. Gibbs sampling can be employed to perform approximate inference [4, 3, 5]. In the Gibbs sampling procedure above, one needs to calculate the conditional distribution $P(z_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})$, where $\vec{z}_{\neg(m,n)}$ represents the topic assignments for all tokens except $w_{m,n}$. We begin with the joint distribution $P(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta})$. One can take advantage of conjugate priors to simplify the integrals.

$$\begin{aligned}
& P(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) \\
= \; & P(\vec{w}|\vec{z}, \vec{\beta}) P(\vec{z}|\vec{\alpha}) \\
= \; & \int P(\vec{w}|\Phi, \vec{z}) p(\Phi|\vec{\beta}) d\Phi \times \int P(\vec{z}|\Theta) p(\Theta|\vec{\alpha}) d\Theta \\
= \; & \int \prod_{m=1}^{M} \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \prod_{k=1}^{K} p(\vec{\varphi}_k|\vec{\beta}) d\Phi \\
& \times \int \prod_{m=1}^{M} \left(\prod_{n=1}^{N_m} P(z_{m,n}|\vec{\vartheta}_m) p(\vec{\vartheta}_m|\vec{\alpha}) \right) d\Theta \\
= \; & \int \prod_{k=1}^{K} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}} \prod_{k=1}^{K} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v - 1} \right) d\Phi
\end{aligned}$$

$$
\begin{aligned}
& \times \int \prod_{m=1}^{M} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}} \prod_{m=1}^{M} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma_k(\alpha_k)} \prod_{k=1}^{K} \vartheta_{m,k}^{\alpha_k - 1} \right) d\Theta \\
= & \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^M \\
& \times \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(n_k^{(v)} + \beta_v)}{\Gamma(\sum_{v=1}^{V} (n_k^{(v)} + \beta_v))} \prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma(n_m^{(k)} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k))} \qquad (1)
\end{aligned}
$$

where $n_k^{(v)}$ is the number of tokens of word v are assigned to topic k, $n_m^{(k)}$ represent the number of tokens in document m are assigned to topic k.

Using the chain rule, the conditional probability (full conditional) can be obtained conveniently,

$$
\begin{aligned}
& P(z_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}) \\
= & \frac{P(z_{m,n}, w_{m,n}|\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})}{P(w_{m,n}|\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})} \\
= & \frac{P(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta})}{P(\vec{w}, \vec{z}_{\neg(m,n)}|\vec{\alpha}, \vec{\beta})} \\
= & \frac{P(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta})}{P(\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}|\vec{\alpha}, \vec{\beta}) P(w_{m,n}|\vec{\alpha}, \vec{\beta})} \\
\propto & \frac{P(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta})}{P(\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}|\vec{\alpha}, \vec{\beta})} \\
\propto & \frac{n_{z_{m,n}}^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V} (n_{z_{m,n}}^{(v)} + \beta_v) - 1} \times (n_m^{(z_{m,n})} + \alpha_{z_{m,n}} - 1) \qquad (2)
\end{aligned}
$$

Finally, we need to obtain the multinomial parameter sets Θ and Φ. According to their definition as multinomial distributions with Dirichlet prior, applying Bayes' rule yields:

$$
\begin{aligned}
P(\vec{\vartheta}_m|\vec{z}_{m,\cdot}, \vec{\alpha}) & = \frac{P(\vec{\vartheta}_m, \vec{z}_{m,\cdot}|\vec{\alpha})}{P(\vec{z}_{m,\cdot}|\vec{\alpha})} \\
& = \frac{P(\vec{z}_{m,\cdot}|\vec{\vartheta}_m) p(\vec{\vartheta}_m|\vec{\alpha})}{P(\vec{z}_{m,\cdot}|\vec{\alpha})} \\
& = \frac{\prod_{n=1}^{N_m} P(z_{m,n}|\vec{\vartheta}_m) \times p(\vec{\vartheta}_m|\vec{\alpha})}{P(\vec{z}_{m,\cdot}|\vec{\alpha})} \\
& = \frac{1}{P(\vec{z}_m|\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}} \times \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \vartheta_{m,k}^{\alpha_k - 1}
\end{aligned}
$$

$$\begin{aligned} &= \frac{1}{Z_{\vec{\vartheta}_m}} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1} \\ &= \text{Dirichlet}(\vec{\vartheta}_m | \vec{n}_m + \vec{\alpha}) \end{aligned} \tag{3}$$

$$\begin{aligned} P(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) &= \frac{P(\vec{\varphi}_k,\vec{w}|\vec{z},\vec{\beta})}{P(\vec{w}|\vec{z},\vec{\beta})} \\ &= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{(m,n):z_{m,n}=k} P(w_{m,n}|\vec{\varphi}_k)p(\vec{\varphi}_k|\vec{\beta}) \\ &= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}} \times \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v-1} \\ &= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}+\beta_v-1} \\ &= \text{Dirichlet}(\vec{\varphi}_k | \vec{n}_k + \vec{\beta}) \end{aligned} \tag{4}$$

where $\vec{z}_{m,\cdot}$ represents the topic assignments for all tokens in document m, that is, $\vec{z}_{m,\cdot} = \{z_{m,n}\}_{n=1}^{N_m}$, $\vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}$ is the vector of topic observation counts for document m and $\vec{n}_k = \{n_k^{(v)}\}_{v=1}^{V}$ that of word observation counts for topic k. Using the expectation of the Dirichlet distribution on Eqs. (3) and (4) yields:

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V}(n_k^{(v)} + \beta_v)} \tag{5}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_m^{(k)} + \alpha_k)} \tag{6}$$

Using Eqs. (2), (5) and (6), the Gibbs sampling procedure in Figure 2 can be run. The procedure itself uses only five larger data structures, the count variables $n_m^{(k)}$ and $n_k^{(v)}$, which have dimension $M \times K$ and $K \times V$ respectively, their row sums n_m and n_k with dimension M and K, as well as the state variable $z_{m,n}$ with dimension $W = \sum_{m=1}^{M} N_m$. The Gibbs sampling algorithm runs over the three periods: initialization, burn-in and sampling. However, to determine the required lengths of the burn-in is one of the drawbacks with MCMC (Markov-Chain Monte Carlo) approaches.

To obtain the resulting model parameters from a Gibbs sampler, several approaches exist. One is to just use only one read out, another is to average a number of samples, and often it is desirable to leave an interval of L iteration between subsequent read-outs to obtain decorrelated states of the Markov chain. This interval is often called ”thinning interval” or sampling lag.

```
**Algorithm** LDAGibbs($\{\vec{w}\}, \vec{\alpha}, \vec{\beta}, K$)
**Input**: word vectors $\{\vec{w}\}$, hyperparameters $\vec{\alpha}, \vec{\beta}$, topic number $K$
**Global data**: count statistics $\{n_m^{(k)}\}, \{n_k^{(v)}\}$ and their sums $\{n_m\}, \{n_k\}$, memory
for full conditional array $P(z_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})$
**Output**: topic associations $\{\vec{z}\}$, multinomial parameters $\Phi$ and $\Theta$, hyperparameter
estimates $\vec{\alpha}, \vec{\beta}$
// initialization
zero all count variables: $n_m^{(k)}, n_m, n_k^{(w_{m,n})}, n_k$
**for** all documents $m \in [1, M]$ **do**
    **for** all words $n \in [1, N_m]$ in document $m$ **do**
        sample topic index $z_{m,n} = k \sim$ Multinomial($1/K$)
        increment document-topic count: $n_m^{(k)}$ += 1
        increment document-topic sum: $n_m$ += 1
        increment topic-word count: $n_k^{(w_{m,n})}$ += 1
        increment topic-word sum: $n_k$ += 1
// Gibbs sampling over burn-in period and sampling period
**while** not finished **do**
    **for** all documents $m \in [1, M]$ **do**
        **for** all words $n \in [1, N_m]$ in documents $m$ **do**
            // for the current assignment of $k$ to the word token $w_{m,n}$:
            decrement counts and sums: $n_m^{(k)}$ -= 1; $n_m$ -= 1; $n_k^{(w_{m,n})}$ -= 1; $n_k$ -= 1
            // multinomial sampling according to Eq. (2):
            sample topic index $\tilde{k} \sim P(z_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})$
            // for the new assignment of $z_{m,n}$ to the word token $w_{m,n}$:
            increment counts and sums: $n_m^{(\tilde{k})}$ += 1; $n_m$ += 1; $n_{\tilde{k}}^{(w_{m,n})}$ += 1; $n_{\tilde{k}}$ += 1
    **if** converged and $L$ sampling iterations since last read out **then**
        // the different parameters read outs are averaged.
        read out parameter set $\Phi$ according to Eq. (5)
        read out parameter set $\Theta$ according to Eq. (6)
```

Figure 2: Gibbs sampling algorithm for LDA Model

2 ToT Model

Wang & McCallum [19] presents an LDA-style topic model that captures not only the low-dimensional structure of data, but also how the structure changes over time. Unlike other recent work that relies on Markov assumptions or discretization of time, here each topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. Thus, the meaning of a particular topic can be relied upon as constant, but the topics'occurrence and correlations change significantly over time.

2.1 Notation & Generative Process

The notation is summarized in Table 2, and the graphical model representations of the Topics over Time (ToT) model is shown in Figure 3.

Table 2: Notation used in the ToT model.

Symbol	Description
K	number of topics / mixture components (const scalar)
M	number of documents (const scalar)
V	number of unique words (const scalar)
N_m	number of word tokens in document m (const scalar)
$\vec{\vartheta}_m$	the multinomial distribution of topics specific to the document m. One proportion for each document, $\Theta = \{\vec{\vartheta}_m\}_{m=1}^{M}$ ($M \times K$ matrix)
$\vec{\varphi}_k$	the multinomial distribution of words specific to the topic k. One component for each topic, $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$ ($K \times V$ matrix)
$\vec{\psi}_k$	the beta distribution of time specific to the topic k. One component for each topic, $\Psi = \{\vec{\psi}_k\}_{k=1}^{K}$ ($K \times 2$ matrix)
$z_{m,n}$	the topic associated with the n-th token in the document m
$w_{m,n}$	the n-th token in document m
$t_{m,n}$	the timestamp associated with the n-th token in the document m
$\vec{\alpha}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\vartheta}$ (K-vector or scalar if symmetric)
$\vec{\beta}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\varphi}$ (V-vector or scalar if symmetric)

The ToT is a generative model of timestamps and the words in the timestamped documents. The generative process can be described as follows.

1. For each topic $k \in [1, K]$:

   (a) Draw a multinomial $\vec{\varphi}_k$ from a Dirichlet prior $\vec{\beta}$;

2. For each document $m \in [1, M]$:

(a) Draw a multinomial $\vec{\vartheta}_m$ from a Dirichlet prior $\vec{\alpha}$;

(b) For each word $n \in [1, N_m]$ in document m:

   i. Draw a topic $z_{m,n}$ from multinomial $\vec{\vartheta}_m$;

   ii. Draw a word $w_{m,n}$ from multinomial $\vec{\varphi}_{z_{m,n}}$;

   iii. Draw a timestamp $t_{m,n}$ from Beta $\vec{\psi}_{z_{m,n}}$;

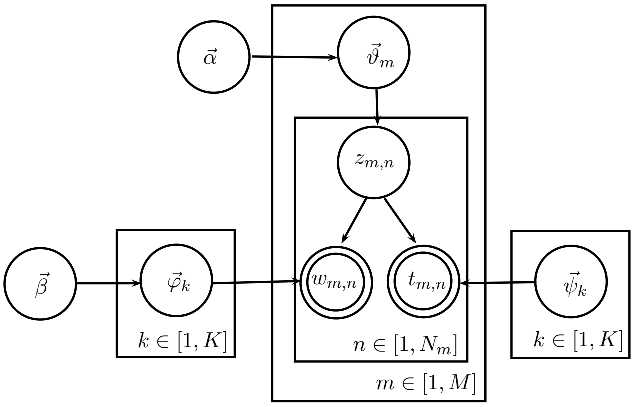

Figure 3: The Graphical Model Representation of the ToT Model

As shown in the above process, the posterior distribution of topics depends on the information from two modalities—both text and time. The parameterization of the ToT model is

$$\begin{aligned}
\vec{\vartheta}_m|\vec{\alpha} &\sim \text{Dirichlet}(\vec{\alpha}) \\
\vec{\varphi}_k|\vec{\beta} &\sim \text{Dirichlet}(\vec{\beta}) \\
z_{m,n}|\vec{\vartheta}_m &\sim \text{Multinomial}(\vec{\vartheta}_m) \\
w_{m,n}|\vec{\varphi}_{z_{m,n}} &\sim \text{Multinomial}(\vec{\varphi}_{z_{m,n}}) \\
t_{m,n}|\vec{\psi}_{z_{m,n}} &\sim \text{Beta}(\vec{\psi}_{z_{m,n}})
\end{aligned}$$

2.2 Gibbs Sampling Derivation

Inference can not be done exactly in this model. We employ Gibbs sampling to perform approximate inference. In the Gibbs sampling procedure above, we need to calculate the conditional distribution $P(z_{m,n}|\vec{w}, \vec{t}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}, \Psi)$, where $\vec{z}_{\neg(m,n)}$ represents the topic assignments for all tokens except $w_{m,n}$. We begin with the joint distribution $P(\vec{w}, \vec{t}, \vec{z}|\vec{\alpha}, \vec{\beta}, \Psi)$. We can take advantage of conjugate priors to simplify the integrals.

$$P(\vec{w}, \vec{t}, \vec{z}|\vec{\alpha}, \vec{\beta}, \Psi)$$

$$
\begin{aligned}
&= P(\vec{w}|\vec{z},\vec{\beta})p(\vec{t}|\Psi,\vec{z})P(\vec{z}|\vec{\alpha}) \\
&= \int P(\vec{w}|\Phi,\vec{z})p(\Phi|\vec{\beta})d\Phi \times p(\vec{t}|\Psi,\vec{z}) \times \int P(\vec{z}|\Theta)p(\Theta|\vec{\alpha})d\Theta \\
&= \int \prod_{m=1}^{M}\prod_{n=1}^{N_m} P(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \prod_{k=1}^{K} p(\vec{\varphi}_k|\vec{\beta})d\Phi \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(t_{m,n}|\vec{\psi}_{z_{m,n}}) \\
&\quad \times \int \prod_{m=1}^{M}\left(\prod_{n=1}^{N_m} P(z_{m,n}|\vec{\vartheta}_m)p(\vec{\vartheta}_m|\vec{\alpha})\right)d\Theta \\
&= \int \prod_{k=1}^{K}\prod_{v=1}^{V}\varphi_{k,v}^{n_k^{(v)}} \prod_{k=1}^{K}\left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}\varphi_{k,v}^{\beta_v-1}\right)d\Phi \\
&\quad \times \int \prod_{m=1}^{M}\prod_{k=1}^{K}\vartheta_{m,k}^{n_m^{(k)}} \prod_{m=1}^{M}\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma_k(\alpha_k)}\prod_{k=1}^{K}\vartheta_{m,k}^{\alpha_k-1}\right)d\Theta \\
&\quad \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(t_{m,n}|\vec{\psi}_{z_{m,n}}) \\
&= \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\right)^{K}\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{M}\prod_{m=1}^{M}\prod_{n=1}^{N_m} p(t_{m,n}|\vec{\psi}_{z_{m,n}}) \\
&\quad \times \prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(n_k^{(v)}+\beta_v)}{\Gamma(\sum_{v=1}^{V}(n_k^{(v)}+\beta_v))}\prod_{m=1}^{M}\frac{\prod_{k=1}^{K}\Gamma(n_m^{(k)}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}(n_m^{(k)}+\alpha_k))} \qquad (7)
\end{aligned}
$$

where $n_k^{(v)}$ is the number of tokens of word v are assigned to topic k, $n_m^{(k)}$ represent the number of tokens in document m are assigned to topic k.

Using the chain rule, we can obtain the conditional probability conveniently,

$$
\begin{aligned}
& P(z_{m,n}|\vec{w},\vec{t},\vec{z}_{\neg(m,n)},\vec{\alpha},\vec{\beta},\Psi) \\
= \; & \frac{P(z_{m,n},w_{m,n},t_{m,n}|\vec{w}_{\neg(m,n)},\vec{t}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{\alpha},\vec{\beta},\Psi)}{P(w_{m,n},t_{m,n}|\vec{w}_{\neg(m,n)},\vec{t}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{\alpha},\vec{\beta},\Psi)} \\
= \; & \frac{P(\vec{w},\vec{t},\vec{z}|\vec{\alpha},\vec{\beta},\Psi)}{P(\vec{w},\vec{t},\vec{z}_{\neg(m,n)}|\vec{\alpha},\vec{\beta},\Psi)} \\
= \; & \frac{P(\vec{w},\vec{t},\vec{z}|\vec{\alpha},\vec{\beta},\Psi)}{P(\vec{w}_{\neg(m,n)},\vec{t}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{\alpha},\vec{\beta},\Psi)P(w_{m,n},t_{m,n}|\vec{\alpha},\vec{\beta},\Psi)} \\
\propto \; & \frac{P(\vec{w},\vec{t},\vec{z}|\vec{\alpha},\vec{\beta},\Psi)}{P(\vec{w}_{\neg(m,n)},\vec{t}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{\alpha},\vec{\beta},\Psi)} \\
\propto \; & \frac{n_{z_{m,n}}^{(w_{m,n})}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n_{z_{m,n}}^{(v)}+\beta_v)-1} \times (n_m^{(z_{m,n})}+\alpha_{z_{m,n}}-1) \times p(t_{m,n}|\vec{\psi}_{z_{m,n}})
\end{aligned}
$$

$$\begin{aligned}
&\propto \frac{n_{z_{m,n}}^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V} (n_{z_{m,n}}^{(v)} + \beta_v) - 1} \times (n_m^{(z_{m,n})} + \alpha_{z_{m,n}} - 1) \\
&\times \frac{\Gamma(\psi_{z_{m,n},1} + \psi_{z_{m,n},2})}{\Gamma(\psi_{z_{m,n},1})\Gamma(\psi_{z_{m,n},2})} t_{m,n}^{\psi_{z_{m,n},1}-1} (1 - t_{m,n})^{\psi_{z_{m,n},2}-1} \qquad (8)
\end{aligned}$$

Finally, we need to obtain the multinomial parameter sets Θ and Φ. According to their definition as multinomial distributions with Dirichlet prior, applying Bayes' rule yields:

$$\begin{aligned}
P(\vec{\vartheta}_m | \vec{z}_{m,\cdot}, \vec{\alpha}) &= \frac{P(\vec{\vartheta}_m, \vec{z}_{m,\cdot} | \vec{\alpha})}{P(\vec{z}_{m,\cdot} | \vec{\alpha})} \\
&= \frac{P(\vec{z}_{m,\cdot} | \vec{\vartheta}_m) p(\vec{\vartheta}_m | \vec{\alpha})}{P(\vec{z}_{m,\cdot} | \vec{\alpha})} \\
&= \frac{\prod_{n=1}^{N_m} P(z_{m,n} | \vec{\vartheta}_m) \times p(\vec{\vartheta}_m | \vec{\alpha})}{P(\vec{z}_{m,\cdot} | \vec{\alpha})} \\
&= \frac{1}{P(\vec{z}_m | \vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}} \times \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \vartheta_{m,k}^{\alpha_k - 1} \\
&= \frac{1}{Z_{\vec{\vartheta}_m}} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)} + \alpha_k - 1} \\
&= \text{Dirichlet}(\vec{\vartheta}_m | \vec{n}_m + \vec{\alpha}) \qquad (9)
\end{aligned}$$

$$\begin{aligned}
P(\vec{\varphi}_k | \vec{z}, \vec{w}, \vec{\beta}) &= \frac{P(\vec{\varphi}_k, \vec{w} | \vec{z}, \vec{\beta})}{P(\vec{w} | \vec{z}, \vec{\beta})} \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{(m,n): z_{m,n} = k} P(w_{m,n} | \vec{\varphi}_k) p(\vec{\varphi}_k | \vec{\beta}) \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}} \times \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v - 1} \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)} + \beta_v - 1} \\
&= \text{Dirichlet}(\vec{\varphi}_k | \vec{n}_k + \vec{\beta}) \qquad (10)
\end{aligned}$$

where $\vec{z}_{m,\cdot}$ represents the topic assignments for all tokens in document m, that is, $\vec{z}_{m,\cdot} = \{z_{m,n}\}_{n=1}^{N_m}$, $\vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}$ is the vector of topic observation counts for document m and $\vec{n}_k = \{n_k^{(v)}\}_{v=1}^{V}$ that of word observation counts for topic k. Using the expectation of the Dirichlet distribution on Eqs. (9) and (10) yields:

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V} (n_k^{(v)} + \beta_v)} \tag{11}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)} \tag{12}$$

For simplicity and speed we update Ψ after each Gibbs sample by the method of moments, detailed as follows:

$$\hat{\psi}_{k,1} = \bar{t}_k \left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1 \right) \tag{13}$$

$$\hat{\psi}_{k,2} = (1-\bar{t}_k) \left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1 \right) \tag{14}$$

where $\bar{t}_k$ and s_k^2 indicate the sample mean and biased sample variance of the timestamps belonging to topic k, respectively, detailed as follows:

$$\begin{aligned}
\bar{t}_k &= \frac{1}{|\{(m,n)|z_{m,n}=k|} \sum_{m=1}^{M} \sum_{n:z_{m,n}=k} t_{m,n} \\
&= \frac{\sum_{m=1}^{M} \left(n_m^{(k)} \times t_m \right)}{\sum_{m=1}^{M} n_m^{(k)}} \\
&= \frac{\sum_{m=1}^{M} \left(n_m^{(k)} \times t_m \right)}{\sum_{v=1}^{V} n_k^{(v)}} && (15) \\
s_k^2 &= \frac{1}{|\{(m,n)|z_{m,n}=k\}|} \sum_{m=1}^{M} \sum_{n:z_{m,n}=k} (t_{m,n} - \bar{t}_k)^2 \\
&= \frac{1}{\sum_{m=1}^{M} n_m^{(k)}} \sum_{m=1}^{M} \left(n_m^{(k)} \times (t_m - \bar{t}_k)^2 \right) \\
&= \frac{\sum_{m=1}^{M} \left(n_m^{(k)} \times t_m^2 \right)}{\sum_{m=1}^{M} n_m^{(k)}} - \bar{t}_k^2 \\
&= \frac{\sum_{m=1}^{M} \left(n_m^{(k)} \times t_m^2 \right)}{\sum_{v=1}^{V} n_k^{(v)}} - \bar{t}_k^2 && (16)
\end{aligned}$$

where $n_m^{(k)}$ represents the number of tokens in document m assigned to topic k, and t_m is the timestamp associated with the document m, usually $t_{m,1} = t_{m,2} = \cdots\ t_{m,N_m} = t_m$.

It is also interesting to consider obtaining a distribution over topics, conditioned on a timestamp. This allows us to see the topic occurrence patterns over time. By Bayes rule, $\mathrm{E}(\vec{\varphi}_k|t) = P(z = k|t) \propto p(t|z = k)P(z = k)$, where $P(z = k)$ can be estimated from data or simply assumed as uniform.

Using Eqs. (8), (11), (12), (13) and (14), the Gibbs sampling procedure in Figure 4 can be run. The procedure itself uses only six larger data structures, the count variables $n_m^{(k)}$ and $n_k^{(v)}$, which have dimension $M \times K$ and $K \times V$ respectively, their row sums n_m and n_k with dimension M and K, Beta parameters Ψ with dimension $K \times 2$, as well as the state variable $z_{m,n}$ with dimension $W = \sum_{m=1}^{M} N_m$. The Gibbs sampling algorithm runs over the three periods: initialization, burn-in and sampling. However, to determine the required lengths of the burn-in is one of the drawbacks with MCMC (Markov-Chain Monte Carlo) approaches.

To obtain the resulting model parameters from a Gibbs sampler, several approaches exist. One is to just use only one read out, another is to average a number of samples, and often it is desirable to leave an interval of L iteration between subsequent read-outs to obtain decorrelated states of the Markov chain. This interval is often called ”thinning interval” or sampling lag.

```
**Algorithm** ToTGibbs($\{\vec{w}\}, \{\vec{t}\}, \vec{\alpha}, \vec{\beta}, \Psi, K$)
**Input**: word vectors $\{\vec{w}\}$, time vector $\{\vec{t}\}$, hyperparameters $\vec{\alpha}, \vec{\beta}$, Beta parameters $\Psi$,
topic number $K$
**Global data**: count statistics $\{n_m^{(k)}\}, \{n_k^{(v)}\}$ and their sums $\{n_m\}, \{n_k\}$, memory for
full conditional array $P(z_{m,n}|\vec{w}, \vec{t}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}, \Psi)$
**Output**: topic associations $\{\vec{z}\}$, multinomial parameters $\Phi$ and $\Theta$, Beta parameter
estimates $\Psi$, hyperparameter estimates $\vec{\alpha}, \vec{\beta}$
// initialization
zero all count variables: $n_m^{(k)}, n_m, n_k^{(w_{m,n})}, n_k$
**for** all documents $m \in [1, M]$ **do**
    **for** all words $n \in [1, N_m]$ in document $m$ **do**
        sample topic index $z_{m,n} = k \sim$ Multinomial($1/K$)
        increment document-topic count: $n_m^{(k)}$ += 1
        increment document-topic sum: $n_m$ += 1
        increment topic-word count: $n_k^{(w_{m,n})}$ += 1
        increment topic-word sum: $n_k$ += 1
// Gibbs sampling over burn-in period and sampling period
**while** not finished **do**
    **for** all documents $m \in [1, M]$ **do**
        **for** all words $n \in [1, N_m]$ in documents $m$ **do**
            // for the current assignment of $k$ to the word token $w_{m,n}$:
            decrement counts and sums: $n_m^{(k)}$ −= 1; $n_m$ −= 1; $n_k^{(w_{m,n})}$ −= 1; $n_k$ −= 1
            // multinomial sampling according to Eq. (8):
            sample topic index $\tilde{k} \sim P(z_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta})$
            // for the new assignment of $z_{m,n}$ to the word token $w_{m,n}$:
            increment counts and sums: $n_m^{(\tilde{k})}$ += 1; $n_m$ += 1; $n_{\tilde{k}}^{(w_{m,n})}$ += 1; $n_{\tilde{k}}$ += 1
            // update $\Psi$ after each Gibbs sample by the method of moments
            update $\Psi$ according to Eqs. (13)–(14)
    **if** converged and $L$ sampling iterations since last read out **then**
        // the different parameters read outs are averaged.
        read out parameter set $\Phi$ according to Eq. (11)
        read out parameter set $\Theta$ according to Eq. (12)
```

Figure 4: Gibbs sampling algorithm for ToT Model

3 AT Model

Rosen-Zvi et al. [13, 14, 12] propose an Author-Topic (AT) model for extracting information about authors and topics from large text collections. Rosen-Zvi et al. model documents as if they were generated by a two-stage stochastic process. An author is represented by a probability distribution over topics, and each topic is represented as a probability distribution over words. The probability distribution over topics in a multi-author paper is a mixture of the distributions associated with the authors.

3.1 Notation & Generative Process

The notation is summarized in Table 3, and the graphical model representations of the AT model is shown in Figure 5.

Table 3: Notation used in the AT model.

Symbol	Description
K	number of topics / mixture components (const scalar)
M	number of documents (const scalar)
V	number of unique words (const scalar)
A	number of unique authors (const scalar)
N_m	number of word tokens in document m (const scalar)
A_m	number of authors in document m (const scalar)
$\vec{a}_m$	authors in document m (A_m-vector)
$\vec{\vartheta}_a$	the multinomial distribution of topics specific to the author a. One proportion for each author, $\Theta = \{\vec{\vartheta}_a\}_{a=1}^{A}$ ($A \times K$ matrix)
$\vec{\varphi}_k$	the multinomial distribution of words specific to the topic k. One component for each topic, $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$ ($K \times V$ matrix)
$z_{m,n}$	the topic associated with the n-th token in the document m
$w_{m,n}$	the n-th token in document m
$x_{m,n}$	the chosen author associated with the word token $w_{m,n}$
$\vec{\alpha}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\vartheta}$ (K-vector or scalar if symmetric)
$\vec{\beta}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\vec{\varphi}$ (V-vector or scalar if symmetric)

The AT model can be viewed as a generative process, which can be described as follows.

1. For each topic $k \in [1, K]$:

   (a) Draw a multinomial $\vec{\varphi}_k$ from a Dirichlet prior $\vec{\beta}$;

2. For each author $a \in [1, A]$:

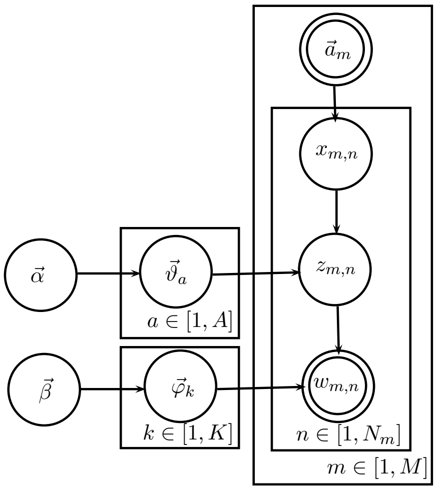

Figure 5: The Graphical Model Representation of the AT Model

   (a) Draw a multinomial $\vec{\vartheta}_a$ from a Dirichlet prior $\vec{\alpha}$;

3. For each document $m \in [1, M]$:

   (a) For each word $n \in [1, N_m]$ in document m:

      i. Draw an author $x_{m,n}$ uniformly from the group of authors $\vec{a}_m$;

      ii. Draw a topic assignment $z_{m,n}$ from per-author multinomial distribution over topics $\vec{\vartheta}_{x_{m,n}}$.

      iii. Draw a word $w_{m,n}$ from multinomial $\vec{\varphi}_{z_{m,n}}$;

As shown in the above process, the posterior distribution of topics depends on the information from the text and authors. The parameterization of the AT model is

$$\begin{aligned}
\vec{\vartheta}_a|\vec{\alpha} &\sim \text{Dirichlet}(\vec{\alpha}) \\
\vec{\varphi}_k|\vec{\beta} &\sim \text{Dirichlet}(\vec{\beta}) \\
z_{m,n}|\vec{\vartheta}_{x_{m,n}} &\sim \text{Multinomial}(\vec{\vartheta}_{x_{m,n}}) \\
w_{m,n}|\vec{\varphi}_{z_{m,n}} &\sim \text{Multinomial}(\vec{\varphi}_{z_{m,n}}) \\
x_{m,n}|A_m &\sim \text{Multinomial}(1/A_m)
\end{aligned}$$

3.2 Gibbs Sampling Derivation

The inference of the model is done by Gibbs Sampling. The biggest drawback of the model is that it loses the distribution over topics for documents. In [12], the authors proposed a heuristic solution to this problem: adding a fictitious author for each document. In the Gibbs sampling procedure above, we need to calculate the conditional distribution $P(z_{m,n}, x_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta})$, where $\vec{z}_{\neg(m,n)}$, $\vec{x}_{\neg(m,n)}$ represents the topic, author assignments for all tokens except $w_{m,n}$, respectively. We begin with the joint distribution $P(\vec{w}, \vec{z}, \vec{x}|\vec{a}, \vec{\alpha}, \vec{\beta})$. We can take advantage of conjugate priors to simplify the integrals.

$$\begin{aligned}
& P(\vec{w}, \vec{z}, \vec{x}|\vec{a}, \vec{\alpha}, \vec{\beta}) \\
= \; & P(\vec{w}|\vec{z}, \vec{\beta}) P(\vec{z}|\vec{x}, \vec{\alpha}) P(\vec{x}|\vec{a}) \\
= \; & \int P(\vec{w}|\vec{z}, \Phi) p(\Phi|\vec{\beta}) d\Phi \times \int P(\vec{z}|\vec{x}, \Theta) p(\Theta|\vec{\alpha}) d\Theta \times P(\vec{x}|\vec{a}) \\
= \; & \int \prod_{m=1}^{M} \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \prod_{k=1}^{K} p(\vec{\varphi}_k|\vec{\beta}) d\Phi \times \prod_{m=1}^{M} \prod_{n=1}^{N_m} P(x_{m,n}|\vec{a}_m) \\
& \times \int \prod_{m=1}^{M} \prod_{n=1}^{N_m} P(z_{m,n}|\vec{\vartheta}_{x_{m,n}}) \prod_{a=1}^{A} p(\vec{\vartheta}_a|\vec{\alpha}) d\Theta \\
= \; & \frac{1}{\prod_{m=1}^{M} A_m^{N_m}} \int \prod_{k=1}^{K} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}} \prod_{k=1}^{K} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v - 1} \right) d\Phi \\
& \times \int \prod_{a=1}^{A} \prod_{k=1}^{K} \vartheta_{a,k}^{n_a^{(k)}} \prod_{a=1}^{A} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \vartheta_{a,k}^{\alpha_k - 1} \right) d\Theta \\
= \; & \frac{1}{\prod_{m=1}^{M} A_m^{N_m}} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^A \\
& \times \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(n_k^{(v)} + \beta_v)}{\Gamma(\sum_{v=1}^{V} (n_k^{(v)} + \beta_v))} \prod_{a=1}^{A} \frac{\prod_{k=1}^{K} \Gamma(n_a^{(k)} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (n_a^{(k)} + \alpha_k))} \qquad (17)
\end{aligned}$$

where $n_k^{(v)}$ is the number of times tokens of word v is assigned to topic k, $n_a^{(k)}$ represent the number of times author a is assigned to topic k.

Using the chain rule, the conditional probability can be obtained conveniently,

$$\begin{aligned}
& P(z_{m,n}, x_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta}) \\
= \; & \frac{P(w_{m,n}, z_{m,n}, x_{m,n}|\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta})}{P(w_{m,n}|\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta})} \\
= \; & \frac{P(\vec{w}, \vec{z}, \vec{x}|\vec{a}, \vec{\alpha}, \vec{\beta})}{P(\vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}|\vec{a}, \vec{\alpha}, \vec{\beta})} \\
= \; & \frac{P(\vec{w}, \vec{z}, \vec{x}|\vec{a}, \vec{\alpha}, \vec{\beta})}{P(\vec{w}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}|\vec{a}, \vec{\alpha}, \vec{\beta}) P(w_{m,n}|\vec{a}, \vec{\alpha}, \vec{\beta})}
\end{aligned}$$

$$\begin{aligned}
&\propto \frac{P(\vec{w},\vec{z},\vec{x}|\vec{a},\vec{\alpha},\vec{\beta})}{P(\vec{w}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{x}_{\neg(m,n)}|\vec{a},\vec{\alpha},\vec{\beta})} \\
&\propto \frac{n_{z_{m,n}}^{(w_{m,n})}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n_{z_{m,n}}^{(v)}+\beta_v)-1} \times \frac{n_{x_{m,n}}^{(z_{m,n})}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n_{x_{m,n}}^{(k)}+\alpha_k)-1} \times \frac{1}{A_m} \\
&\propto \frac{n_{z_{m,n}}^{(w_{m,n})}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n_{z_{m,n}}^{(v)}+\beta_v)-1} \times \frac{n_{x_{m,n}}^{(z_{m,n})}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n_{x_{m,n}}^{(k)}+\alpha_k)-1} \qquad (18)
\end{aligned}$$

If one further manipulates the above formula , one can turn them into separated update equations for the topic and author of each token, suitable for random or systematic scan updates:

$$\begin{aligned}
&P(x_{m,n}|\vec{x}_{\neg(m,n)},\vec{z},\vec{a},\vec{\alpha}) \\
&= \frac{P(z_{m,n},x_{m,n}|\vec{x}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{a},\vec{\alpha})}{P(z_{m,n}|\vec{x}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{a},\vec{\alpha})} \\
&= \frac{P(\vec{x},\vec{z}|\vec{a},\vec{\alpha})}{P(\vec{x}_{\neg(m,n)},\vec{z}|\vec{a},\vec{\alpha})} \\
&= \frac{P(\vec{x},\vec{z}|\vec{a},\vec{\alpha})}{P(\vec{x}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{a},\vec{\alpha})P(z_{m,n}|\vec{a},\vec{\alpha})} \\
&\propto \frac{P(\vec{x},\vec{z}|\vec{a},\vec{\alpha})}{P(\vec{x}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{a},\vec{\alpha})} \\
&\propto \frac{n_{x_{m,n}}^{(z_{m,n})}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n_{x_{m,n}}^{(k)}+\alpha_k)-1} \qquad (19)
\end{aligned}$$

$$\begin{aligned}
&P(z_{m,n}|\vec{w},\vec{z}_{\neg(m,n)},\vec{x},\vec{\alpha},\vec{\beta}) \\
&= \frac{P(w_{m,n},z_{m,n}|\vec{w}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{x},\vec{\alpha},\vec{\beta})}{P(w_{m,n}|\vec{w}_{\neg(m,n)},\vec{z}_{\neg(m,n)},\vec{x},\vec{\alpha},\vec{\beta})} \\
&= \frac{P(\vec{w},\vec{z}|\vec{x},\vec{\alpha},\vec{\beta})}{P(\vec{w},\vec{z}_{\neg(m,n)}|\vec{x},\vec{\alpha},\vec{\beta})} \\
&= \frac{P(\vec{w},\vec{z}|\vec{x},\vec{\alpha},\vec{\beta})}{P(\vec{w}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{x},\vec{\alpha},\vec{\beta})P(w_{m,n}|\vec{x},\vec{\alpha},\vec{\beta})} \\
&\propto \frac{P(\vec{w},\vec{z}|\vec{x},\vec{\alpha},\vec{\beta})}{P(\vec{w}_{\neg(m,n)},\vec{z}_{\neg(m,n)}|\vec{x},\vec{\alpha},\vec{\beta})} \\
&\propto \frac{n_{z_{m,n}}^{(w_{m,n})}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n_{z_{m,n}}^{(v)}+\beta_v)-1} \times \frac{n_{x_{m,n}}^{(z_{m,n})}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n_{x_{m,n}}^{(k)}+\alpha_k)-1} \qquad (20)
\end{aligned}$$

Finally, we need to obtain the multinomial parameter sets Θ and Φ.

According to their definition as multinomial distributions with Dirichlet prior, applying Bayes' rule yields:

$$\begin{aligned}
P(\vec{\vartheta}_a|\vec{z},\vec{x},\vec{\alpha}) &= \frac{P(\vec{\vartheta}_a,\vec{z}|\vec{x},\vec{\alpha})}{P(\vec{z}|\vec{x},\vec{\alpha})} \\
&= \frac{1}{Z_{\vec{\vartheta}_a}} \prod_{(m,n):x_{m,n}=a} P(z_{m,n}|\vec{\vartheta}_a)p(\vec{\vartheta}_a|\vec{\alpha}) \\
&= \frac{1}{Z_{\vec{\vartheta}_a}} \prod_{k=1}^{K} \vartheta_{a,k}^{n_a^{(k)}} \times \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \vartheta_{a,k}^{\alpha_k-1} \\
&= \frac{1}{Z_{\vec{\vartheta}_a}} \prod_{k=1}^{K} \vartheta_{a,k}^{n_a^{(k)}+\alpha_k-1} \\
&= \text{Dirichlet}(\vec{\vartheta}_a|\vec{n}_a+\vec{\alpha}) \qquad (21)
\end{aligned}$$

$$\begin{aligned}
P(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) &= \frac{P(\vec{\varphi}_k,\vec{w}|\vec{z},\vec{\beta})}{P(\vec{w}|\vec{z},\vec{\beta})} \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{(m,n):z_{m,n}=k} P(w_{m,n}|\vec{\varphi}_k)p(\vec{\varphi}_k|\vec{\beta}) \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}} \times \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v-1} \\
&= \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)}+\beta_v-1} \\
&= \text{Dirichlet}(\vec{\varphi}_k|\vec{n}_k+\vec{\beta}) \qquad (22)
\end{aligned}$$

where $\vec{n}_a = \{n_a^{(k)}\}_{k=1}^{K}$ is the vector of topic observation counts for the author a and $\vec{n}_k = \{n_k^{(v)}\}_{v=1}^{V}$ that of word observation counts for topic k. Using the expectation of the Dirichlet distribution on Eqs. (21) and (22) yields:

$$\varphi_{k,v} = \frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)} \qquad (23)$$

$$\vartheta_{a,k} = \frac{n_a^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_a^{(k)}+\alpha_k)} \qquad (24)$$

Using Eqs. (18)/(19)-(20), (23) and (24), the Gibbs sampling procedure in Figure 6 can be run. The procedure itself uses only six larger data structures, the count variables $n_a^{(k)}$ and $n_k^{(v)}$, which have dimension $A \times K$ and $K \times V$ respectively, their row sums n_a and n_k with dimension A and

K, as well as the state variable $z_{m,n}, x_{m,n}$ with dimension $W = \sum_{m=1}^{M} N_m$. The Gibbs sampling algorithm runs over the three periods: initialization, burn-in and sampling. However, to determine the required lengths of the burn-in is one of the drawbacks with MCMC (Markov-Chain Monte Carlo) approaches.

To obtain the resulting model parameters from a Gibbs sampler, several approaches exist. One is to just use only one read out, another is to average a number of samples, and often it is desirable to leave an interval of L iteration between subsequent read-outs to obtain decorrelated states of the Markov chain. This interval is often called "thinning interval" or sampling lag.

```
**Algorithm** ATGibbs$(\{\vec{w}\}, \{\vec{a}\}, \vec{\alpha}, \vec{\beta}, K)$[1]
**Input**: word vectors $\{\vec{w}\}$, author vectors $\{\vec{a}\}$, hyperparameters $\vec{\alpha}, \vec{\beta}$, topic number $K$
**Global data**: count statistics $\{n_a^{(k)}\}, \{n_k^{(v)}\}$ and their sums $\{n_a\}, \{n_k\}$, memory
for full conditional array $P(z_{m,n}, x_{m,n} | \vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta})$
**Output**: topic associations $\{\vec{z}\}$, author associations $\{\vec{a}\}$, multinomial parameters $\Phi$ and
$\Theta$, hyperparameter estimates $\vec{\alpha}, \vec{\beta}$
// initialization
zero all count variables: $n_a^{(k)}, n_a, n_k^{(w_{m,n})}, n_k$
**for** all documents $m \in [1, M]$ **do**
    **for** all words $n \in [1, N_m]$ in document $m$ **do**
        sample topic index $z_{m,n} = k \sim \text{Multinomial}(1/K)$
        sample author index $x_{m,n} = a \sim \text{Multinomial}(\vec{p})$, where $p_a = \{^{1/A_m, a \in \vec{a}_m}_{0,\text{otherwise}}$
        increment author-topic count: $n_a^{(k)} \mathrel{+}= 1$
        increment author-topic sum: $n_a \mathrel{+}= 1$
        increment topic-word count: $n_k^{(w_{m,n})} \mathrel{+}= 1$
        increment topic-word sum: $n_k \mathrel{+}= 1$
// Gibbs sampling over burn-in period and sampling period
**while** not finished **do**
    **for** all documents $m \in [1, M]$ **do**
        **for** all words $n \in [1, N_m]$ in document $m$ **do**
            // for the current assignment of $k, a$ to the word token $w_{m,n}$:
            decrement counts and sums: $n_a^{(k)} \mathrel{-}= 1$; $n_a \mathrel{-}= 1$; $n_k^{(w_{m,n})} \mathrel{-}= 1$; $n_k \mathrel{-}= 1$
            // multinomial sampling according to Eq. (18):
            sample topic and author index $(\tilde{k}, \tilde{a}) \sim P(z_{m,n}, x_{m,n} | \vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{a}, \vec{\alpha}, \vec{\beta})$
            // multinomial sampling according to Eqs. (19)-(20):
            Sample author index $\tilde{a} \sim P(x_{m,n} | \vec{x}_{\neg(m,n)}, \vec{z}, \vec{a}, \vec{\alpha})$
            Sample topic index $\tilde{z} \sim P(z_{m,n} | \vec{x}, \vec{z}_{\neg(m,n)}, \vec{w}, \vec{\alpha}, \vec{\beta})$
            // for the new assignment of $z_{m,n}, x_{m,n}$ to the word token $w_{m,n}$:
            increment counts and sums: $n_{\tilde{a}}^{(\tilde{k})} \mathrel{+}= 1$; $n_{\tilde{a}} \mathrel{+}= 1$; $n_{\tilde{k}}^{(w_{m,n})} \mathrel{+}= 1$; $n_{\tilde{k}} \mathrel{+}= 1$
    **if** converged and $L$ sampling iterations since last read out **then**
        // the different parameters read outs are averaged.
        read out parameter set $\Phi$ according to Eq. (23)
        read out parameter set $\Theta$ according to Eq. (24)
```

[1] One should not run both red and blue lines, but either red or blue lines.

Figure 6: Gibbs sampling algorithm for AT Model

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, 2002.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[3] Thomas L. Griffiths and Mark Steyvers. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235, 2004.

[4] Tom Griffiths. Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation. Technical report, Stanford University, 2002.

[5] Gregor Heinrich. Parameter Estimation for Text Analysis. Technical report version 2.9, vsonix GmbH and University of Leipzig, 2009.

[6] Noriaki Kawamae. Author interest topic model. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 887–888, New York, NY, USA, 2010. ACM.

[7] Noriaki Kawamae. Latent interest-topic model: Finding the causal relationships behind dyadic data. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 649–658, New York, NY, USA, 2010. ACM.

[8] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Technical report um-cs-2004-096, Department of Computer Science, University of Massachusetts Amherst, 2004.

[9] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 786–791, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[10] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in socail networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.

[11] David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509, New York, NY, USA, 2007. ACM.

[12] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning Author-Topic Models from Text Corpora. *ACM Transactions on Information Systems*, 28(1):1–38, 2010.

[13] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic S-myth. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, Arlington, Virginia, USA, 2004. AUAI Press.

[14] Mark Steyvers, Padhraic Smyth, Michael Rosen-Zvi, and Thomas Griffiths. Probabilistic Author-Topic Models for Information Discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 306–315, New York, NY, USA, 2004. ACM.

[15] Jie Tang, Ruo ming Jin, and Jing Zhang. A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 1055–1060, Washington, DC, USA, 2008. IEEE Computer Society.

[16] Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. *Machine Learning*, 82(2):211–237, 2011.

[17] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, New York, NY, USA, 2008. ACM.

[18] Hanna M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984, 2006.

[19] Xuerui Wang and Andrew McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, New York, NY, USA, 2006. ACM.

[20] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining*, pages 697–702, Washington, DC, USA, 2007. IEEE Computer Society.