Multiscale Topic Tomography

Ramesh Nallapati
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
nmramesh@cs.cmu.edu

William Cohen
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
wcohen@cs.cmu.edu

Susan Ditmore
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
sditmore@andrew.cmu.edu

John Lafferty
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
lafferty@cs.cmu.edu

Kin Ung
Networking and Computing Services
Johnson and Johnson group
1003 Route 202
Raritan, NJ 08869
kinvung@gmail.com

ABSTRACT

Modeling the evolution of topics with time is of great value in automatic summarization and analysis of large document collections. In this work, we propose a new probabilistic graphical model to address this issue. The new model, which we call the *Multiscale Topic Tomography Model* (MTTM), employs non-homogeneous Poisson processes to model generation of word-counts. The evolution of topics is modeled through a multi-scale analysis using Haar wavelets. One of the new features of the model is its modeling the evolution of topics at various time-scales of resolution, allowing the user to zoom in and out of the time-scales. Our experiments on *Science* data using the new model uncovers some interesting patterns in topics. The new model is also comparable to LDA in predicting unseen data as demonstrated by our perplexity experiments.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*data mining*

General Terms

Algorithms, Experimentation

Keywords

Topic modeling, Temporal evolution, time-scale, Poisson, Probabilistic graphical models, wavelets

KDD'07, August 12–15, 2007, San Jose, California, USA.

1. INTRODUCTION

Explosive growth of electronic document collections in the recent past has rendered their analysis by human experts extremely tedious and expensive. As a result, an increasing need is felt for automatic algorithms that analyze and summarize the topics contained in such large document collections.

Several probabilistic graphical models have been proposed recently, to address this problem. One of the first probabilistic and truly generative models among them is Latent Dirichlet Allocation (LDA) [2]. LDA models a topic as a multinomial distribution over the vocabulary. Given a document collection, the LDA learns its underlying topics in an unsupervised fashion. In the recent past, several extensions to this model have been proposed such as the Hierarchical Dirichlet Processes [12] model that automatically discovers the number of topics, Hidden Markov Model-LDA [5] that integrates topic modeling with syntax, Correlated topic models [1] that model pairwise correlations between topics, *etc.*

All the aforementioned models ignore an important factor that reveals a huge amount of information contained in large document collections - *time*. Some of the large corpora such as a collection of scientific journals or patent databases span several decades. Hence modeling the evolution and popularity of topics with time can reveal tremendous amount of hidden information in those collections.

Several models have been proposed in the recent past to address this issue. One of the models, called *Topics Over Time* (ToT) [13] associates a beta distribution over time to each topic that represents the occurrence probability of that topic at any given time. The model learns the parameters of this distribution for each topic based on the time-stamps of documents associated with that topic in the collection. This permits us to analyze the popularity of various topics as a function of time.

Another proposed model called the *Dynamic Topic Models* (DTM) [3] takes a slightly different approach. The DTM explicitly models the evolution of topics with time by estimating the topic distribution at various epochs. Thus the DTM allows us to predict what words are 'in vogue' in a

particular topic at different points in time. To model the evolution of a topic with time, the authors assume that the *natural* parameters corresponding to the topic multinomial at each epoch are conditionally distributed by a normal distribution with mean equal to the natural parameters at the previous epoch. However, since the normal distribution is not a conjugate to the multinomial distribution, the model does not yield a simple solution to the problems of inference and estimation.

In this work, we present an alternative to the DTM, that is more natural to sequential modeling of counts data. The new model uses conjugate priors on the topic parameters to model evolution of topics, thereby resulting in simpler solutions. In addition, our new model, which we refer to henceforth as the *Multiscale Topic Tomography Model* (MTTM), allows us to analyze the evolution of topics at various resolutions of time scale. Its expressiveness provides the user with additional flexibility to zoom-in and zoom-out on the time scale and study the evolution of topics at a chosen time scale. Thus, we believe that the MTTM brings us a step closer to the ultimate goal of effective and fully automatic analysis of document collections.

The rest of the paper is organized as follows. In section 2, we discuss past work related to the new model. In section 3, we describe the MTTM in detail including its generative process, the multi-scale analysis and the variational methods used for learning and inference. Section 4 presents some of the experiments we performed using the model. Section 5 concludes the paper with some analysis and directions for future work.

2. PAST WORK

The Poisson distribution, being a natural model for counts-data, has been considered as a potential candidate to model text in the past. One of the earliest models is the *2-Poisson* model for information retrieval [6], which generates words from a mixture of two classes called *elite* and *non-elite* classes. This model did not achieve empirical success, mainly owing to the lack of good estimation techniques, but inspired a heuristic model called *BM25* [11]. The latter is considered a strong IR baseline till date.

In the area of text modeling, the *GaP* model [4] proposed by Canny uses a combination of Gamma and Poisson distributions to discover latent topics or *themes* in document collections. The Gamma distribution is used to generate the topic weights vector $\mathbf{x}$ in each document, which the author calls *theme lengths*. The Poisson distribution is used to generate the vector of observed word counts $\mathbf{f}$ from expected counts $\mathbf{y}$. The expected counts $\mathbf{y}$ are related to the topic weights $\mathbf{x}$ through a matrix Λ, given by $\mathbf{y} = \Lambda\mathbf{x}$, where each column of Λ represents the probability distribution of words in a topic. Canny developed an EM algorithm to estimate the topic weights $\mathbf{x}$ for each document and the global matrix Λ. Furthermore, it is also shown that the model achieves a lower perplexity on test data compared to LDA while also outperforming baseline models on the task of text retrieval. However, the modeling scheme for *GaP* proposed by Canny optimizes likelihood of the complete data (i.e., the data with the maximum likelihood values used for the unobserved variables), whereas a pure generative model should optimize the likelihood of the observed data only. The model presented in this paper is very similar to the *Gap* model, except that the theme-weights in our case are distributed by a *Dirichlet distribution over documents* instead of a Gamma. This particular definition of Dirichlet means that the topic-weights are normalized over all the documents for each topic and not over all the topics per document as in LDA. Also, in our parameter-estimation, we are able to optimize a variational lower-bound on the *observed data* log-likelihood by marginalizing the theme weights in the complete-data log-likelihood. In addition, we extend this model to sequential data by performing multi-scale analysis.

In our work, we use the Poisson distribution to model word counts not only because it is a natural choice for counts-data, but also because it is amenable to sequence modeling through Bayesian multiscale analysis. Bayesian multiscale models for Poisson processes were first introduced by Kolaczyk [8] and were applied to model physical phenomena such as gamma ray bursts. Nowak extended multiscale analysis to build multiscale hidden Markov models and applied it to the problem of image segmentation [9]. Nowak and Kolaczyk also presented multiscale analysis for the Poisson inverse problem [10], which is the problem of estimating latent Poisson means based on observed Poisson data, whose means are related to the latent Poisson means by a known linear function. In this paper, we cast the problem of topic discovery in document collections as a Poisson inverse problem. Unlike in the work of Nowak and Kolaczyk [10], we do not assume that the linear relationship between the latent Poisson parameters and observed Poissons is known, which makes the problem slightly more complex. Hence, we use variational approximations to estimate the parameters of the model. We also extend the analysis to multi-scale representation of the Poisson parameters, thereby allowing us to model temporal evolution of topics at various time-scales.

3. MULTISCALE TOPIC TOMOGRAPHY MODEL

3.1 Assumptions and Notations

Following standard notation, we use bold faced letters to represent vectors and matrices and regular font to indicate scalars.

We assume that our document collection is sorted in the ascending order of the publication dates of the documents. We also assume that the sorted collection is divided into 2^S equal-sized chunks (where S is an integer) of size M each, with the chunks $\{C_0, \cdots, C_{2^S-1}\}$ indexed in the ascending order of time. Thus, each chunk C_t represents an *epoch* of time t ranging from the publication date of its earliest published document $d = 1$ to the publication date of its latest document $d = M$. Henceforth, we will use the term *epoch* to also denote the chunk of documents C_t that it corresponds to. We represent each document d in an epoch t by a vector of term counts $\mathbf{n}_{td} = \{n_{td1}, \cdots, n_{tdV}\}$ where n_{tdw} is the count of word w in document d from epoch t, and V is the vocabulary size.

We use non-homogeneous Poisson processes to model evolution of topics with time. Accordingly, each epoch t is associated with its unique word generating Poisson parameters given by $\boldsymbol{\mu}_t = \{\boldsymbol{\mu}_{t1}, \cdots, \boldsymbol{\mu}_{tK}\}$ corresponding to K topics. Again each $\boldsymbol{\mu}_{tk}$ is a a vector of Poisson means over the vocabulary given by $\{\mu_{tk1}, \cdots, \mu_{tkV}\}$. Thus, the parameter μ_{tkw} represents the expected number of counts of word w from topic k during the epoch t. Unlike in LDA where

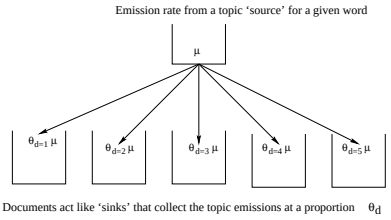

Figure 1: The intuitive idea of Topic Tomography

topics are represented as multinomial distributions over the vocabulary, we represent topics as vectors of Poisson means over the vocabulary μ_{tk}. The variation in the values of the Poisson means of a particular topic as a function of t will provide us information on the evolution of the topic content with time.

We use the terms Poisson rate, Poisson mean and Poisson parameter interchangeably in the rest of the paper.

3.2 Generative process

3.2.1 Data generation

Given the Poisson parameters for each epoch, we generate the data as follows. For each epoch t and topic k, we first generate the topic-weights vector $\boldsymbol{\theta}_{tk}$ from a Dirichlet distribution, where $\boldsymbol{\theta}_{tk} = \{\theta_{tk1}, \cdots, \theta_{tkM}\}$ is a multinomial distribution over the documents in the corresponding chunk C_t. Each component of the multinomial, θ_{tkd}, represents the degree to which the document d 'captures' the topic k. Then, for each document d in the chunk and for each word w, we generate the counts n_{tdw} using a Poisson distribution whose mean is given by $\sum_k \theta_{tkd}\mu_{tkw}$, a weighted combination of the Poisson means of all the topics corresponding to that word. The generative process is presented more precisely below.

1. For each epoch $t = 0, \cdots, 2^S - 1$
2. For each topic $k = 1, \cdots, K$
3. Generate $\boldsymbol{\theta}_{tk} \sim \text{Dir}(\cdot \mid \alpha)$
4. For each document $d = 1, \cdots, M$
5. For each word $w = 1, \cdots, V$
6. Generate $n_{tdw} \sim \text{Poiss}(\cdot \mid \sum_k \theta_{tkd}\mu_{tkw})$

Note that the topic-weights in the linear combination do not sum to 1 since θ_{tkd} represents $P(d|k)$, the probability that the topic k appears in document d and not $P(k|d)$, the probability that the document discusses the topic k, as defined in LDA. An intuitive way to understand the new model would be to think of each topic as an emission from a source, and the documents as sinks that share the topic emissions amongst themselves. Thus the new model captures how the topic is *sectioned* among the documents in a given epoch, hence the name *Topic Tomography*. This idea is illustrated in figure 1.

The generative process of the observed data is graphically represented in figure 2. Accordingly, the data likelihood

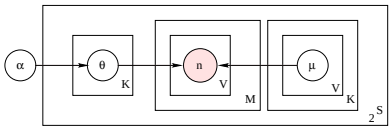

Figure 2: Graphical representation of data generation

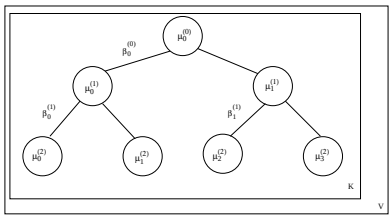

Figure 3: Binary tree representation of multiscale parameters for the case S=2

given the Poisson parameters is given by:

$$P(\mathbf{n}|\boldsymbol{\theta}, \boldsymbol{\mu}, \alpha) = \prod_{t=0}^{2^S-1} \{\prod_{k=1}^{K} \text{Dir}(\boldsymbol{\theta}_{tk}|\alpha) \times \prod_{d=1}^{M}\prod_{w=1}^{V} \text{Poiss}(n_{tdw}|\sum_k \theta_{tkd}\mu_{tkw})\} \quad (1)$$

3.2.2 Parameter generation

The process described above is already a complete generative process for a document collection. However, we have not yet defined how the Poisson parameters of different epochs are related to each other. In this section, we define a multiscale generative process for the Poisson parameters that allows us to model temporal evolution of topics. First, we define multiscale wavelet parameters given by a binary tree representation as shown below:

$$\boldsymbol{\mu}_t^{(S)} = \boldsymbol{\mu}_t \text{ for } t = 0, \cdots, 2^S - 1 \quad (2)$$

$$\boldsymbol{\mu}_t^{(s)} = \boldsymbol{\mu}_{(2t)}^{(s+1)} + \boldsymbol{\mu}_{(2t+1)}^{(s+1)} \text{ for } s = 0, .., S-1 \text{ and } t = 0, .., 2^s - 1 \quad (3)$$

where the index s is called the scale and corresponds to the depth of the tree. The highest scale of resolution given by S corresponds to the leaves of the binary tree, where each leaf node represents an epoch. The multiscale Poisson parameters $\boldsymbol{\mu}_t^{(S)}$ at each leaf node $t \in \{0, \cdots, 2^S - 1\}$ are set equal to the Poisson parameters corresponding to the respective epoch t. At any lower scale of resolution ($0 \leq s \leq S-1$), the Poisson parameter at node $t \in \{0, \cdots, 2^s - 1\}$, given by $\boldsymbol{\mu}_t^{(s)}$, is set equal to sum of the corresponding parameters at its two children. The parameters $\boldsymbol{\mu}_t^{(s)}$ defined this way are known as the unnormalized Haar wavelet scaling coefficients of $\boldsymbol{\mu}_t$ [9]. The multi-scale Poisson parameters are pictorially represented in figure 3.

While each leaf node in the tree corresponds to an epoch, any non-leaf node at scale $0 \leq s \leq S-1$ corresponds to a larger epoch of time whose span ranges the epochs of the leaf nodes to which it is an ancestor. At scale $s = 0$, we have only the root node whose epoch spans the time-period of the entire collection and the Poisson parameters at this scale correspond to the average topic representation for the

whole corpus. As we descend down the tree to a higher scale of resolution s, we have 2^s nodes at that scale with shorter epochs for each node and a breadth-wise traversal from left to right gives us the evolution of topics at that scale.

Now, we also define the *canonical* multiscale parameters $\boldsymbol{\beta}_t^{(s)}$ as follows.

$$\boldsymbol{\beta}_t^{(s)} = \frac{\boldsymbol{\mu}_{(2t)}^{(s+1)}}{\boldsymbol{\mu}_t^{(s)}} \quad \text{for } s = 0, .., S-1 \text{ and } t = 0, \cdots, 2^s - 1 \quad (4)$$

In other words, at each scale s (except $s = S$) and for each node t at that scale, $\boldsymbol{\beta}_t^{(s)}$ represents the ratio of the Poisson parameter at the left child and that at the node under consideration. The canonical parameters are also called splitting factors since they govern how the multiscale parameter $\mu_t^{(s)}$ is 'split' between its children. We can also invert the relation in Eq. (4) to obtain

$$\boldsymbol{\mu}_{(2t)}^{(s)} = \boldsymbol{\beta}_t^{(s-1)} \times \boldsymbol{\mu}_t^{(s-1)} \quad (5)$$

$$\boldsymbol{\mu}_{(2t+1)}^{(s)} = \boldsymbol{\mu}_t^{(s-1)} - \boldsymbol{\mu}_{(2t)}^{(s)} \quad (6)$$

$$\text{for } s = 1, \cdots, S \text{ and } t = 0, \cdots, 2^{s-1} - 1$$

where we obtained Eq. (6) from Eq. (3). The canonical parameters are represented at the edges of the binary tree in figure 3 to indicate that they are a function of the Poisson parameters at the two nodes that share the respective edges. We will later show that one can factor the joint likelihood of the observed data under the independent Poissons as a multi-scale likelihood using the canonical parameters.

Note that setting $\boldsymbol{\beta}_t^{(s)} = 0.5$ is equivalent to the relation $\boldsymbol{\mu}_{(2t)}^{(s+1)} = \boldsymbol{\mu}_{(2t+1)}^{(s+1)}$ meaning the multiscale Poissons that share the same parent are equal. This relation should immediately delight a Bayesian statistician, since we can conveniently encode our prior information that the Poisson parameters of a given topic are expected to be more or less the same over various epochs (in other words, topics do not change too drastically with time), by imposing a symmetric, conjugate prior on $\boldsymbol{\beta}_t^{(s)}$.

Given this background, the generative process for the Poisson parameters is as shown below.

1. For each topic $k = 0, \cdots, K$
2. For each word $w = 1, \cdots, V$
3. Generate $\mu_{0kw}^{(0)} \sim \text{Gamma}(\cdot|\lambda_\mu, \delta_\mu)$
4. For each scale $s = 0, \cdots, S-1$
5. For each epoch $t = 0, \cdots, 2^s - 1$
6. Generate $\beta_{tkw}^{(s)} \sim \text{Beta}(\cdot|\delta_\beta, \delta_\beta)$

We used the Gamma distribution to generate the Poisson parameters since it is a conjugate prior to the Poisson. We will later show that the observed data log-likelihood can be factored into a multiscale log-likelihood in which the canonical parameters act as binomial parameters. Hence we used the Beta distribution, their natural conjugate prior, to generate them. In particular, the symmetric Beta aligns topics in one epoch to topics in the adjacent epoch and also ensures that their evolution remains smooth.

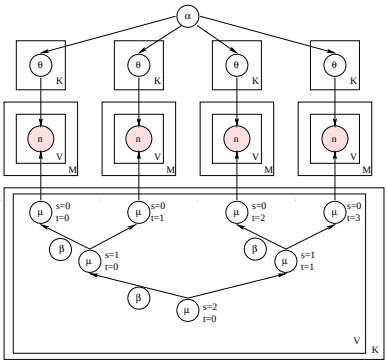

Figure 4: Graphical representation of MTTM for S=2: we purposely omitted the hyper-parameters in the figure for clarity.

The prior probability of the model parameters given the hyperparameters $\boldsymbol{\delta} = \{\lambda_\mu, \delta_\mu, \delta_\beta\}$ is then given as follows:

$$P(\boldsymbol{\mu}|\boldsymbol{\delta}) = \prod_{k=1}^{K}\prod_{w=1}^{V}\{\text{Gamma}(\mu_{0kw}^{(0)}|\lambda_\mu, \delta_\mu) \times \prod_{s=0}^{S-2}\prod_{t=0}^{2^s-1}\text{Beta}(\beta_{tkw}^{(s)}|\delta_\beta, \delta_\beta)\} \quad (7)$$

The generative process of the data as well as the model parameters together is represented graphically in figure 4. Combining Eq. (1) and Eq. (7), one can compute the marginal likelihood of the observed data given the topic parameters and the hyper-parameters of the priors as follows.

$$P(\mathbf{n}|\boldsymbol{\mu}, \alpha, \boldsymbol{\delta}) = \int_{\boldsymbol{\theta}}\{P(\mathbf{n}|\boldsymbol{\theta}, \boldsymbol{\mu}, \alpha)d\boldsymbol{\theta}\}\, P(\boldsymbol{\mu}|\boldsymbol{\delta}) = P(\mathbf{n}|\boldsymbol{\mu}, \alpha)P(\boldsymbol{\mu}|\boldsymbol{\delta}) \quad (8)$$

3.3 Variational EM

Since estimating the parameters of the model is intractable, we use variational EM to estimate the parameters of the model [7]. We only summarize the results below but the interested reader may refer to appendix A for more details.

3.3.1 Variational E-step

We introduce variational parameters given by γ_{tkd} and ϕ_{tdwk}, to approximate the observed data log-likelihood given in Eq. (8). One can think of the γ_{tkd} as proportional to the posterior probability that document d of epoch t captures topic k. ϕ_{tdwk} can be interpreted as the posterior probability that word w in document d of epoch t came from topic k. We estimate them by maximizing a variational lower-bound of Eq. (8) with respect to the variational parameters. We summarize the results in Eq. (9) and Eq. (10) below, with details in appendix A.

$$\phi_{tdwk} \propto \mu_{tkw}^{(S-1)} \exp(\psi(\gamma_{tkd}) - \psi(\sum_{d=1}^{M}\gamma_{tkd})) \quad (9)$$

$$\gamma_{tkd} = \alpha + \sum_{w=1}^{V} n_{tdw}\phi_{tdwk} \quad (10)$$

3.3.2 Variational M-step

In the M-step, we estimate the model parameters, namely $\boldsymbol{\mu}$. Instead of directly estimating the parameters $\mu^{(S)}_{tkw}$ by maximizing the variational lower-bound of Eq. (8) with respect to these parameters, we express the likelihood in a slightly different form so as to be able to estimate the multi-scale parameters $\mu^{(s)}_{tkw}$ for $0 \leq s \leq S$. Since, we are only interested in estimating $\boldsymbol{\mu}$ in the M-step, we collect all the terms in the variational lower-bound of Eq. (8), given by Eq. (19) in the appendix, that contain $\mu^{(S)}_{tkw}$ and call the expression $L_{[\boldsymbol{\mu}]}$ as shown below.

$$L_{[\boldsymbol{\mu}]} = \sum_{t=0}^{2^S-1}\sum_{w=1}^{V}\sum_{k=1}^{K}\{-\mu^{(S)}_{tkw} + \log \mu^{(S)}_{tkw} \sum_{d=1}^{M} n_{tdw}\phi_{tdwk}\}$$

$$= \sum_{t=0}^{2^S-1}\sum_{w=1}^{V}\sum_{k=1}^{K}\{-\mu^{(S)}_{tkw} + z_{twk}\log \mu^{(S)}_{tkw}\} \quad (11)$$

$$\overset{\boldsymbol{\mu}}{\equiv} \sum_{t=0}^{2^S-1}\sum_{w=1}^{V}\sum_{k=1}^{K}\log \text{Poiss}(z_{twk}|\mu^{(S)}_{tkw}) \quad (12)$$

where z_{twk} is the latent count of the word w in topic k in the entire chunk of documents corresponding to epoch t (corresponding to scale S) and is given by $z_{twk} = \sum_{d=1}^{M} n_{tdw}\phi_{tdwk}$. Although we showed that $z_{twk} \sim \text{Poiss}(\cdot|\mu^{(S)}_{tkw})$ by simple algebraic manipulation, it is also possible to prove it theoretically. This proof is presented in appendix B.

In Eq. (12), the notation $\overset{\boldsymbol{\mu}}{\equiv}$ indicates that its left-hand-side is equal to its right-hand-side as far as the terms containing $\boldsymbol{\mu}$ are concerned. Note that Eq. (11) and Eq. (12) differ by the factor $\sum_t \sum_w \sum_k \log(z_{twk}!)$ but since it doesn't contain $\boldsymbol{\mu}$, it does not affect our estimation. Also note that one may round-off z_{twk} to the nearest integer to account for the fact that the Poisson generates only integers, but this plays no major role in terms of estimating $\boldsymbol{\mu}$.

We now define a new multi-scale variable $z^{(s)}_{twk}$ on the same lines as the multiscale parameters as follows. We will show shortly that $L_{[\boldsymbol{\mu}]}$ can be expressed in terms of this variable.

$$z^{(S)}_{twk} = z_{twk} \text{ for } t = 0, \cdots, 2^S - 1 \quad (13)$$

$$z^{(s)}_{twk} = z^{(s+1)}_{(2t)wk} + z^{(s+1)}_{(2t+1)wk} \quad (14)$$
$$\text{for } s = 0, .., S-1 \text{ and } t = 0, .., 2^s - 1$$

The simplified version of $L_{[\boldsymbol{\mu}]}$ in (12) can be equivalently expressed in terms of the multiscale parameters as shown below.

$$L_{[\boldsymbol{\mu}]} = \sum_{t=0}^{2^S-1}\sum_{w=1}^{V}\sum_{k=1}^{K}\log \text{Poiss}(z_{twk}|\mu^{(S)}_{tkw})$$

$$= \sum_{s=0}^{S-1}\sum_{t=0}^{2^s-1}\sum_{w=1}^{V}\sum_{k=1}^{K}\log \text{Bin}(z^{(s+1)}_{(2t)wk}|\beta^{(s)}_{tkw}, z^{(s)}_{twk})$$

$$+ \sum_{w=1}^{V}\sum_{k=1}^{K}\log \text{Poiss}(z^{(0)}_{0wk}|\mu^{(0)}_{0kw}) \quad (15)$$

The proof for the above transformation is as follows: we first note the result that the joint probability of two independent Poisson variables x_1 and x_2 can be equivalently expressed as a product of a binomial and a Poisson as follows.

$$\begin{aligned}
& \text{Poiss}(x_1|\mu_1)\text{Poiss}(x_2|\mu_2) \\
= \quad & \frac{\exp(-(\mu_1+\mu_2))\mu_1^{x_1}\mu_1^{x_1}}{x_1!x_2!} \\
= \quad & \frac{(x_1+x_2)!}{x_1!x_2!}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{x_1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{x_2} \\
\times \quad & \frac{\exp(-(\mu_1+\mu_2))(\mu_1+\mu_2)^{x_1+x_2}}{(x_1+x_2)!} \\
= \quad & \text{Bin}(x_1|x_1+x_2, \frac{\mu_1}{\mu_1+\mu_2}) \\
\times \quad & \text{Poiss}(x_1+x_2|\mu_1+\mu_2)
\end{aligned}$$

Applying this result to the Poisson likelihood terms in left-hand-side of Eq. (15) recursively results in its right-hand-side.

We do a MAP estimate of the multiscale parameters using Eq. (15) and the priors defined in Eq. (7) to obtain the following relations.

$$\beta^{(s)}_{tkw} = \frac{z^{(s+1)}_{(2t)wk} + \delta_\beta - 1}{z^{(s)}_{twk} + 2(\delta_\beta - 1)} \quad (16)$$

$$\mu^{(0)}_{tkw} = \frac{z^{(0)}_{0wk} + \lambda_\mu - 1}{1 + \delta_\mu} \quad (17)$$

4. EXPERIMENTS

4.1 Analysis of Science

We analyzed a subset of 30,000 articles from *Science*, 250 from each of the 120 years between 1883 and 2002. This is essentially the same data used by Blei and Lafferty in their experiments with the DTM [3]. We divided the data into 16 chunks, each consisting of 1875 documents. Each of these chunks represents a 15 year epoch. We then trained a 5-scale topic tomography model with number of topics $K = 50$ on this data set with the following values for the hyper-parameters : $(\lambda_\mu = 1.0001; \delta_\mu = 1; \delta_\beta = 50; \alpha = 0.8)$. The large value of δ_β ensures that the Poisson parameters of adjacent epochs are nearly equal, resulting in a smooth evolution of topics.

Figure 5 shows the multi-scale representation of a topic which we manually labeled "particle physics"[1]. We only displayed the top 10 terms that had the highest Poisson means in that topic. The root node of the binary tree corresponds to $s = 0$, and it represents the summary of the topic over the entire 120 year span of the collection. At the highest scale ($s = 3$) displayed, each node presents a snap-shot summary of the topic in a 15 year period (Note that owing to space constraints, we did not display the highest scale of resolution $s = 4$). Thus, the user can choose one of the four scales of resolution depending on the desired granularity. Inspecting the topic snap-shots at the scale $s = 3$, one can easily gain an understanding of the evolution of the topic. The gradual transition in the topic from macro-matter to micro-matter is

[1]Note that this and the other topics we displayed emerged in our run with $K = 50$. If we use a different value of K, it is not guaranteed that the exact same topics will emerge again.

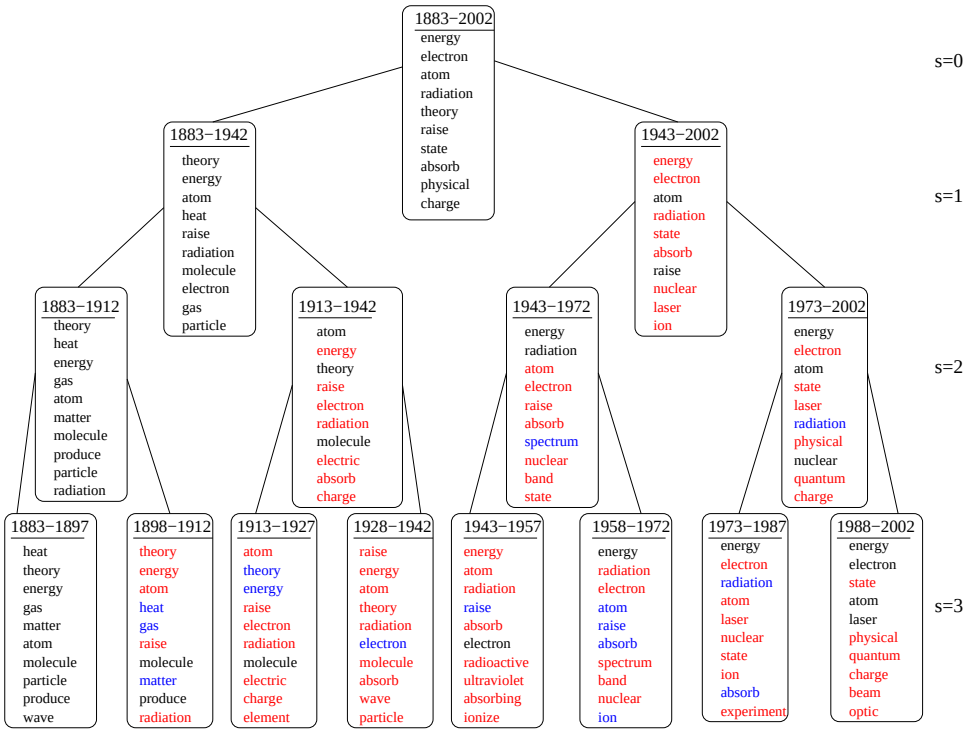

Figure 5: **A 4-scale representation of a topic which we manually labeled "Particle physics": best seen in color. Words colored red are those whose relative importance in the topic has gone up compared to previous epoch at the same scale. Words colored blue are those whose relative importance has gone down. Words not colored have retained their position compared to previous epoch.**

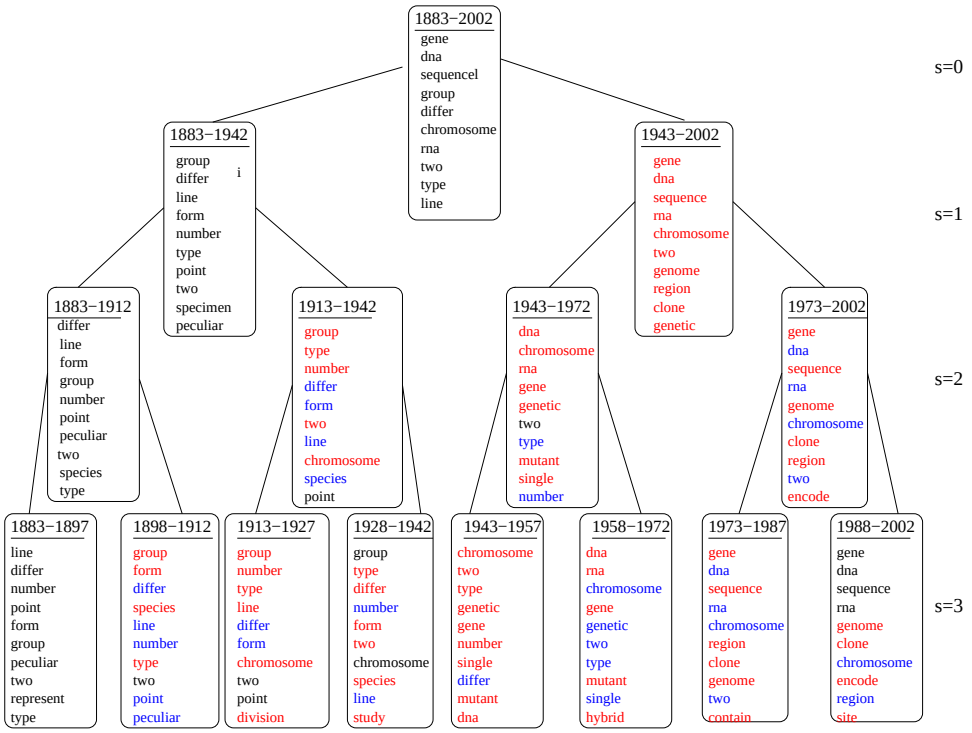

Figure 6: **A 4-scale representation of a topic we identified as "Genetics": best seen in color. The color code is same as in figure 5.**

1890	"The Cause of Motion in the Radiometer"
1903	"Electricity at High Pressures"
1927	"Ionization by Positive Ions"
1936	"The Production of Cosmic Ray Showers"
1949	"Luminescent Solids (Phosphors)"
1964	"Mossbauer Effect in Chemistry and Solid-State Physics"
1978	"Analytical Chemistry: Using Lasers to Detect Less and Less"
1992	"Vibrational Modes and the Dynamic Solvent Effect in Electron and Proton Transfer"

Table 1: Documents in which the topic "Particle-physics" appears with the highest probability in each of the epochs

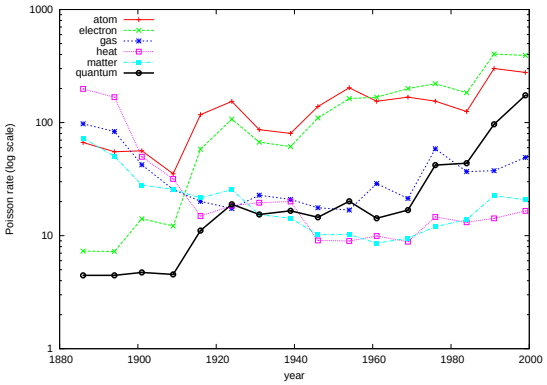

Figure 7: Evolution of content bearing words in the topic manually labeled as "Particle physics": the words "atom", "electron" and "quantum" gain prominence with time, while words such as "heat" and "gas" lose ground, indicating a paradigm shift in the field from macro-matter to micro-matter.

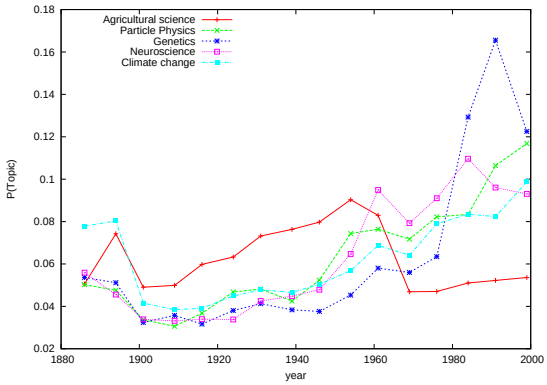

Figure 9: Occurrence probability of topics with time: we plotted the statistic γ_{tk} for the two topics we analyzed earlier, namely "particle physics" and "genetics" and for three other topics, which we identified as "agricultural science", "neuroscience" and "climate change". The plot reveals some interesting patterns. For example, while *agricultural science* remains more or less stable with time, we see an explosion of *genetics* in the 1990's. The topics of *climate change*, *atomic physics* and *neuroscience* also exhibit an increasing prominence in the late 20th century, consistent with the trends in the real-world.

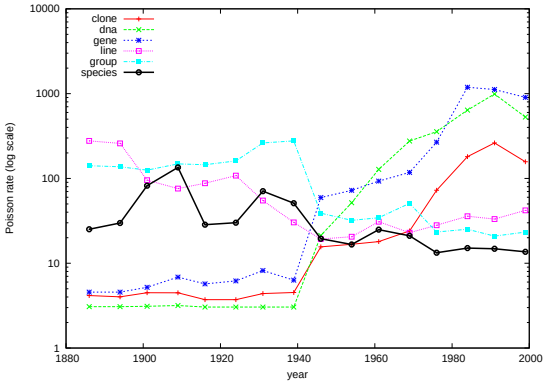

Figure 8: Evolution of content bearing words in the topic "Genetics": it is apparent from the figure that words related to modern genetics such as 'dna', 'clone' and 'gene' exhibit higher emission rates in the late 1990's while words related to evolutionary biology such as 'group' and 'species' taper off with time. Interestingly, the word 'dna' starts coming into prominence only in in the 1950's, just around the time when it was discovered.

more apparent from figure 7, which plots the Poisson rates of a few representative words as a function of time. Table 1 lists the titles of documents that have the highest value of the posterior Dirichlet parameter γ_{tkd} among all documents in each epoch. In other words, these are the documents in which the topic particle-physics appears with the highest probability in each epoch. The shift in the topic is also evident from these titles.

In figure 6, we displayed the multiscale representation of another topic, which we labeled "genetics". An examination of the topic snapshots at scale $s = 3$ clearly shows a gradual transition from evolutionary biology in the late 19th century to modern genetics in the early 21st century. Figure 8 plots the popularity of a few representative terms with time. Table 2, that displays the titles of documents in which the topic appears with the highest probability in each epoch, demonstrates a very similar pattern of topic evolution.

In figure 9, we plotted another interesting statistic, namely the sum of the posterior Dirichlet parameters of a topic k over all documents in each epoch t given by $\gamma_{tk} = \sum_d \gamma_{tkd}$, as a function of t. This statistic is proportional to the occurrence frequency of a topic in a given epoch. We normalized this statistic, so that one can interpret the plot as the probability of occurrence of a topic as a function of time, similar to the plots in [13].

1893	"A Space-Relation of Numbers"
1911	"'Genotype" and "Pure Line""
1922	"Spermatogenesis of the Garter Snake"
1941	"The Artificial Synthesis of a 42-Chromosome Wheat"
1949	"Cytological Evidence Opposing the Theory of Brachymeiosis in the Ascomycetes"
1965	"Bipolarity of Information Transfer from the Salmonella typhimurium Chromosome"
1979	"Distribution of RNA Transcripts from Structural and Intervening Sequences of the Ovalbumin Gene"
2000	"DNA Replication Fork Pause Sites Dependent on Transcription"

Table 2: Documents in which the topic "Genetics" appears with the highest probability in each of the epochs

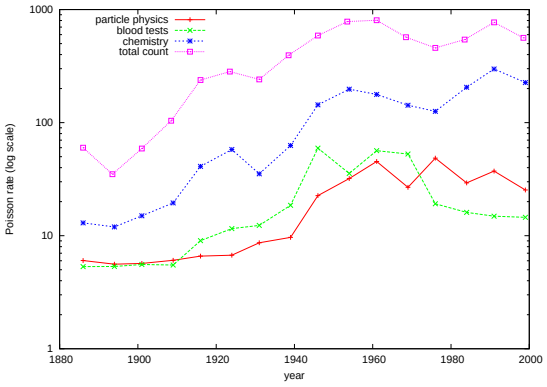

Figure 10: Occurrence rate of "reaction" in three different topics: the word 'reaction' could have several meanings depending on the context in which it is used. In this plot, we see that it is resolved into three topics - 'particle physics', 'blood tests' and 'chemistry'. While the topic 'chemistry' accounts for the majority of occurrences of the word 'reaction', it also occurs at a much lower rate in the context of 'blood tests' and 'particle physics', where it assumes different connotations. One would not be able to see this from a simple plot of overall occurrence counts of the word (which is also displayed above.)

Finally, in figure 10, we plot the Poisson rates of the word 'reaction' in three different topics and compared with the total counts in each epoch. The plot clearly demonstrates the utility of the topic model in disambiguating an ambiguous word based on its context.

4.2 Perplexity

Perplexity is a standard objective metric that measures the ability of a model to predict unseen data. Lower perplexity means better predictiveness and a better model. In case of documents, the average perplexity of a word in a test set D_{test} comprising M documents is defined as

$$\text{Perplexity}(D_{test}|\mathcal{M}) = 2^{\left(\frac{-\log P(\mathbf{n}_{test}|\mathcal{M})}{\sum_{d=1}^{M}\sum_{w=1}^{V} n_{dw}}\right)} \qquad (18)$$

where $\mathbf{n}_{test}$ is the entire vector of observed word counts in the test set and $\mathcal{M}$ is the model.

In this section, we compare the perplexity of the topic tomography model with that of LDA. Note that these two models generate completely different events: while the former models counts-data (*e.g.*: 2 a's and 3 b's), the latter models one particular instance of the counts vector (*e.g.*: 'aabbb'). In order to be able to make a fair comparison, we added the multinomial normalizing coefficient $(\sum_w n_{dw})!/(\prod_w (n_{dw}!))$ for each document in the expression for LDA likelihood. This term converts the probability of a string to the probability of the corresponding counts vector allowing us to directly compare the perplexities of both the models. Hence the perplexity numbers we show in the plots for LDA may not directly correspond to the values obtained by previous authors [2, 3].

For our experiments, we split the data time wise into 8 chunks each spanning 15 years and comprising 3750 documents as done in section 4.1. We further randomly split each chunk into equal halves to generate training and test sets. The train and test sets each have 8 chunks, each of which spans 15 years but consists of only 1875 documents.

We consider three variants of the topic tomography model in our experiments.

The first variant, which we call *basic TT*, is the closest counterpart to LDA. In this model, we completely ignore the multiscale analysis and assume that the entire training (or test) set represents a single epoch. We estimate one set of Poissons for the entire collection, using no prior distributions on the Poisson parameters. The learned model is used to estimate perplexity on the test set, which is done by running the E-step of the variational EM algorithm. The second variant, which we name *multiple TT* model, relaxes the assumption of the basic TT model and estimates topic Poissons for each of the 8 epochs in the training set separately. However, it still does not perform any multiscale analysis and uses no priors on the Poisson means. For each chunk in the test set, we predict the model's perplexity by running the E-step of the variational EM with respect to the model parameters corresponding to the same epoch in the training set. The last variant is the complete multiscale topic tomography model with multiscale analysis using beta priors on the multiscale binomials with hyper-parameters set at the same values used in section 4.1.

For LDA baseline, we used the standard version that estimates a single set of topic multinomials for the whole collection. For all the aforementioned models, we fixed the Dirichlet parameter at 0.8 to encourage sparsity of topics.

Figure 11 compares the perplexity of LDA with the three variants of the topic tomography model as a function of the number of topics used in the model. The figure shows that both LDA and basic TT are almost identical in performance. Also notice that multiscale TT has a consistently lower perplexity than the multiple TT model. This result justifies the intuition behind our definition of the priors in the multiscale-analysis. The priors allow information to propagate from one epoch to another and hence improve the ability of the model to predict unseen data in any given epoch. Finally, we notice that although the multiple TT and the multiscale TT models have many more parameters than the basic TT model, they produce a slightly higher perplexity on the test set compared to the latter. On inspection, we noticed that the performance comparison on the training data is the ex-

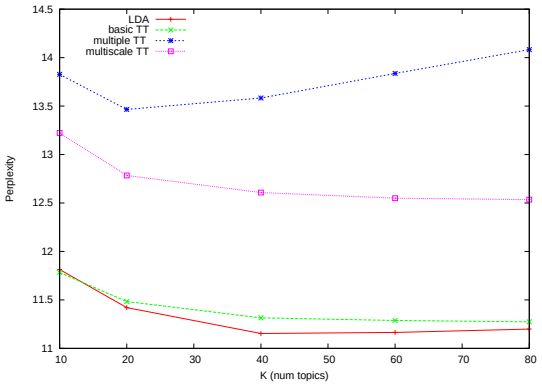

Figure 11: Comparison of Perplexities of the models: lower is better.

act reverse. This is a clear case of over-fitting, where the extra parameters in the multiple TT and the multiscale TT models result in better fitting of the training data, but hurt its generalization ability compared to the basic TT model. Notwithstanding this fact, the multiscale model is still very useful since it allows us to visualize data better, through the multiscale analysis as we have shown earlier.

Finally, we note that the DTM [3], on the contrary, reported slightly lower perplexity than LDA. Unlike the DTM which uses full Bayesian estimation, we use MAP estimation to keep the the algorithm simple. The latter setting results in a few free parameters in the form of the priors $(\lambda_\mu, \delta_\mu, \delta_\beta)$. In our experiments, we fixed these priors to the values reported in section 4.1. Since these values are not necessarily optimal, the perplexity values of MTTM we reported in the experiments are actually upper bounds. It is possible to further lower the perplexity of MTTM by tuning the priors through cross-validation. But we did not venture into this direction because our perplexity experiments are only meant to be illustrative and not necessarily conclusive.

5. DISCUSSION

In this work, we presented a new approach to modeling temporal evolution of topics in a large document collection. The new approach, based on non-homogeneous Poisson processes, combined with multi-scale Haar wavelet analysis is a more natural way to do sequence modeling of counts-data than previous approaches. The new model offers us the best features of both the *ToT* [13] and *DTM* [3] models. While *ToT* models the probability of occurrence of a topic with time, *DTM* models the evolution of topic content. The topic tomography model permits us to accomplish both at the same time. In addition, the multiscale analysis used in the model provides us with an additional 'zoom' feature that permits the user to examine the topic evolution at multiple scales of resolution.

One of the limitations of the MTTM lies in its generative process: since the Dirichlet distribution that generates topic proportions is defined over the set of documents in a given epoch, the model permits only generating an entire chunk of documents whose size is equal to the training set chunk size. This is the main reason why we used equal sized training and test sets in our perplexity experiments in section 4.2. We are no longer able to make inference on a single document at a time. One way to overcome this limitation is use a Gamma distribution to generate topic weights for each document independently as done in the *GaP* model. However, multiscale analysis using the Gamma distributed weights becomes tricky due to the coupling between the Poisson parameters and the Gamma weights. In our case, we were able to uncouple the Poisson parameters from the topic proportions using the relation $\sum_d \theta_{tkd} = 1$ (see Eq. (19) in appendix A). Nevertheless, we intend to construct a variational algorithm for multi-scale analysis using a *GaP* like model, as part of our future work.

6. REFERENCES

[1] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International conference on Machine learning*, pages 113–120, 2006.

[4] J. Canny. Gap: a factor model for discrete data. In *International ACM SIGIR conference on Research and development in information retrieval*, pages 122–129, 2004.

[5] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, pages 537–544, 2005.

[6] S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 35:285–295, 1975.

[7] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[8] E. D. Kolaczyk. Bayesian multiscale models for poisson processes. In *Journal of the American Statistical Association*, pages 920–933, 1999.

[9] R. Nowak. Multiscale hidden markov models for bayesian image analysis. *Bayesian Inference in Wavelet Based Models (B. Vidakovic and P. Muller, eds.), Lecture Notes in Statistics 141, Springer-Verlag.*, 1999.

[10] R. Nowak and E. Kolaczyk. A statistical multiscale framework for poisson inverse problems. *Special issue of IEEE Transactions on Information theory on information-theoretic imaging*, 2000.

[11] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *International ACM SIGIR conference on Research and development in information retrieval*, 1994.

[12] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Technical Report 653, Department Of Statistics, UC Berkeley*, 2003.

[13] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *International conference on Knowledge discovery and data mining*, pages 424–433, 2006.

APPENDIX

A. VARIATIONAL INFERENCE

We define the following variational bounds on the log-likelihood of the observed data using Jensen's inequality for the log function as shown below.

$$\begin{aligned}
&\log P(\mathbf{n}, |\alpha, \boldsymbol{\mu}, \boldsymbol{\delta}) = \sum_{t=0}^{2^S-1} \log(\int_{\boldsymbol{\theta}_t} \{\left(\prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\theta}_{tk}|\alpha)\right) \\
&\times \prod_{d=1}^{M}\prod_{w=1}^{V} \mathrm{Poiss}(n_{tdw}|\sum_k \theta_{tkd}\mu_{tkw}^{(S)})\} d\boldsymbol{\theta}_t) + \log P(\boldsymbol{\mu}|\boldsymbol{\delta}) \\
&\geq \sum_{t=0}^{2^S-1} \{\int_{\boldsymbol{\theta}_t} \{\prod_{k=1}^{K} q(\boldsymbol{\theta}_{tk}|\boldsymbol{\gamma}_{tk})(\sum_{k=1}^{K} \log \mathrm{Dir}(\boldsymbol{\theta}_{tk}|\alpha) \\
&+ \sum_{d=1}^{M}\sum_{w=1}^{V}[-\sum_{k=1}^{K} \theta_{tkd}\mu_{tkw}^{(S)} + n_{tdw}\log(\sum_{k=1}^{K} \theta_{tkd}\mu_{tkw}^{(S)}) \\
&- \log(n_{tdw}!)]\} d\boldsymbol{\theta}_t + \sum_{k=1}^{K} H(q(\boldsymbol{\theta}_{tk}|\boldsymbol{\gamma}_{tk}))\} + \log P(\boldsymbol{\mu}|\boldsymbol{\delta}) \\
&\geq \sum_{t=0}^{2^S-1} \{\sum_{k=1}^{K} E_q[\log \mathrm{Dir}(\boldsymbol{\theta}_{tk}|\alpha)] - \sum_{w=1}^{V}\sum_{k=1}^{K} \mu_{tkw}^{(S)} \\
&+ \sum_{d=1}^{M}\sum_{w=1}^{V} n_{tdw}(\sum_{k=1}^{K} \phi_{tdwk}(\log \mu_{tkw}^{(S)} + E_q[\log \theta_{tkd}])) \\
&- \log(n_{tdw}!) + \sum_{k=1}^{K} H(q(\boldsymbol{\theta}_{tk}|\boldsymbol{\gamma}_{tk})) \\
&+ \sum_{d=1}^{M}\sum_{w=1}^{V} n_{tdw} H(\boldsymbol{\phi}_{tdw})\} + \log P(\boldsymbol{\mu}|\boldsymbol{\delta}) \qquad (19)
\end{aligned}$$

where in Eq. (19), we used the relation $\sum_{d=1}^{M} \theta_{tkd} = 1$ while $q(\boldsymbol{\theta}_{tk}|\boldsymbol{\gamma}_{tk})$ and $\boldsymbol{\phi}_{tdw}$ are variational posterior Dirichlet and variational multinomial distributions respectively, $E_q[X]$ represents the expectation of the random variable X with respect to the distribution $q()$. $H()$ represents the entropy of the distribution in its argument.

The terms in Eq. (19) can be expanded as follows.

$$\begin{aligned}
E_q[\log P(\boldsymbol{\theta}_{tk}|\alpha)] &= \Gamma(M\alpha) - M\Gamma(\alpha) + \sum_d (\alpha - 1) \\
&\quad (\psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd})) \qquad (20) \\
E_q[\log \theta_{tkd}] &= \psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd}) \qquad (21) \\
H(q(\boldsymbol{\theta}_{tk}|\boldsymbol{\gamma}_{tk})) &= \sum_d \log\Gamma(\gamma_{tkd}) - \log\Gamma(\sum_d \gamma_{tkd}) \\
&- \sum_d (\gamma_{tkd} - 1)(\psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd})) \\
&\qquad (22) \\
H(\boldsymbol{\phi}_{tdw}) &= -\sum_k \phi_{tdwk} \log \phi_{tdwk} \qquad (23)
\end{aligned}$$

Plugging back these expansions in Eq. (19) and calling the expression obtained by collecting terms that contain ϕ_{tdwk}, $L_{[\phi_{tdwk}]}$, we have:

$$\begin{aligned}
L_{[\phi_{tdwk}]} &= n_{tdw}\phi_{tdwk}(\log \mu_{tkw}^{(S)} + \psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd}) \\
&- \log \phi_{tdwk}) \qquad (24)
\end{aligned}$$

Taking the partial derivative of $L_{[\phi_{tdwk}]}$ with respect to ϕ_{tdwk} gives:

$$\begin{aligned}
\frac{\partial L_{[\phi_{tdwk}]}}{\partial \phi_{tdwk}} &= n_{tdw}(\log \mu_{tkw}^{(S)} + \psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd}) \\
&- \log \phi_{tdwk} - 1) \qquad (25)
\end{aligned}$$

Setting the partial derivative to zero and solving yields the maximizing value of the variational parameter ϕ_{tdwk} as shown in Eq. (9).

Similarly, collecting the terms in Eq. (19) that contain γ_{tkd} into $L_{[\gamma_{tkd}]}$, we get:

$$\begin{aligned}
L_{[\gamma_{tkd}]} &= (\psi(\gamma_{tkd}) - \psi(\sum_d \gamma_{tkd}))(\alpha + \sum_w n_{tdw}\phi_{tdwk} \\
&- \gamma_{tkd}) + \log\Gamma(\gamma_{tkd}) - \log\Gamma(\sum_d \gamma_{tkd}) \qquad (26)
\end{aligned}$$

Taking the partial derivative of $L_{[\gamma_{tkd}]}$ with respect to γ_{tdk} gives:

$$\begin{aligned}
\frac{\partial L_{[\gamma_{tkd}]}}{\partial \gamma_{tkd}} &= (\psi'(\gamma_{tkd}) - \psi'(\sum_d \gamma_{tkd})) \\
&\quad (\alpha + \sum_w n_{tdw}\phi_{tdwk} - \gamma_{tkd}) \qquad (27)
\end{aligned}$$

Equating the partial derivative to zero results in the maximizing expression for γ_{tkd} shown in Eq. (10).

B. PROOF THAT Z_{TWK} IS A POISSON VARIABLE WITH MEAN $\mu_{TKW}^{(S)}$

We first start with noting that the counts of a word w in a document d from epoch t is distributed as

$$n_{tdw} \sim \mathrm{Poiss}(\cdot|\sum_k \theta_{tkd}\mu_{tkw}^{(S)})$$

Now, let us define the variable z_{tdwk} denoting the latent counts of the word w from topic k in the same document. Now clearly, $\sum_k z_{tdwk} = n_{tdw}$. Since the summation of two independent Poisson random variables is also a Poisson variable with mean equal to the sum of the means of the original random variables, we can infer that

$$z_{tdwk} \sim \mathrm{Poiss}(\cdot|\theta_{tkd}\mu_{tkw}^{(S)})$$

Now z_{twk} is the latent counts of the word w from topic k in the whole chunk that corresponds to epoch t. Therefore, by definition it follows that

$$\begin{aligned}
z_{twk} &= \sum_{d=1}^{M} z_{tdwk} \\
&\sim \mathrm{Poiss}(\cdot|\sum_{d=1}^{M} \theta_{tkd}\mu_{tkw}^{(S)}) \\
&= \mathrm{Poiss}(\cdot|\mu_{tkw}^{(S)}\sum_{d=1}^{M} \theta_{tkd}) = \mathrm{Poiss}(\cdot|\mu_{tkw}^{(S)}) \qquad (28)
\end{aligned}$$