Recurrent Coupled Topic Modeling over Sequential Documents

JINJIN GUO, University of Macau
LONGBING CAO, University of Technology Sydney
ZHIGUO GONG, University of Macau

The abundant sequential documents such as online archival, social media, and news feeds are streamingly updated, where each chunk of documents is incorporated with smoothly evolving yet dependent topics. Such digital texts have attracted extensive research on dynamic topic modeling to infer hidden evolving topics and their temporal dependencies. However, most of the existing approaches focus on single-topic-thread evolution and ignore the fact that a current topic may be coupled with multiple relevant prior topics. In addition, these approaches also incur the intractable inference problem when inferring latent parameters, resulting in a high computational cost and performance degradation. In this work, we assume that a current topic evolves from all prior topics with corresponding coupling weights, forming the *multi-topic-thread evolution*. Our method models the dependencies between evolving topics and thoroughly encodes their complex multi-couplings across time steps. To conquer the intractable inference challenge, a new solution with a set of novel data augmentation techniques is proposed, which successfully discomposes the multi-couplings between evolving topics. A fully conjugate model is thus obtained to guarantee the effectiveness and efficiency of the inference technique. A novel Gibbs sampler with a backward–forward filter algorithm efficiently learns latent time-evolving parameters in a closed-form. In addition, the latent Indian Buffet Process compound distribution is exploited to automatically infer the overall topic number and customize the sparse topic proportions for each sequential document without bias. The proposed method is evaluated on both synthetic and real-world datasets against the competitive baselines, demonstrating its superiority over the baselines in terms of the low per-word perplexity, high coherent topics, and better document time prediction.

CCS Concepts: • **Mathematics of computing → Probabilistic inference problems**; **Gibbs sampling**; **Nonparametric statistics**; • **Information systems → Document topic models;**

Additional Key Words and Phrases: Topic modeling, topic evolution, topic coupling, multiple dependency, data augmentation, gibbs sampling, dropout, bayesian network

This work was supported by the National Key D&R Program of China (2019YFB1600704), FDCT (FDCT/0045/2019/A1, FDCT/0007/2018/A1), GSTIC (EF005/FST-GZG/2019/GSTIC), University of Macau (MYRG2018-00129-FST), and GDST (2019B111106001).

Authors' addresses: J. Guo and Z. Gong, State Key Lab of IoTSC and Department of Computer and Information Science, University of Macau, Macau S.A.R, 999078, China; emails: {yb57414, fstzgg}@um.edu.mo; L. Cao, Data Science Lab, University of Technology Sydney, Ultimo, New South Wales 2007, Sydney, Australia; email: LongBing.Cao@uts.edu.au.

1556-4681/2021/06-ART8 $15.00
https://doi.org/10.1145/3451530

ACM Reference format:
Jinjin Guo, Longbing Cao, and Zhiguo Gong. 2021. Recurrent Coupled Topic Modeling over Sequential Documents. *ACM Trans. Knowl. Discov. Data.* 16, 1, Article 8 (June 2021), 32 pages.
https://doi.org/10.1145/3451530

1 INTRODUCTION

The all-the-time update of abundant digital documents, such as Google News, Twitter, and Flickr, have generated large amounts of sequential temporal-tagged documents, exhibiting complex temporal dependencies across the time steps. Such temporal-tagged digital documents have attracted extensive studies on the *time-evolving* nature of topics. A successful way for such a task is to divide the collection into a sequence of document chunks and each chunk corresponds to a time-slice incorporated with topics in the temporal period [2, 4, 9, 21, 31, 47]. Then, the problem of topic evolution could be addressed by studying relationships between topics crossing two adjacent time slices.

Though fruitful results have been obtained in this area, most of the existing approaches are contracted with the single-topic-thread assumption, i.e., a topic in the current time-slice can only develop into a single topic in the subsequent slice [2, 9, 31, 40, 52]. Obviously, this assumption cannot well align with the reality. Taking the news about COVID-19 as an example, the topic of coronavirus outbreak not only develops itself with intensive reports along the time but also triggers other topics such as the shortage of medical masks, shutdown of entertainment venues, and flight suspension. On the other hand, a new topic (e.g., work resumption) could be coupled with multiple prior topics (e.g., the effective control of coronavirus pandemic and market pressure). Such multi-topic coupling relationships over time are quite common and complex in the real world [11, 13, 27, 53], which pose significant challenges to the existing dynamic topic modeling techniques [2, 9, 31, 40, 52]. This article investigates this multi-topic coupling nature by assuming the *multi-topic-thread evolution*, and proposes the ***recurrent Coupled Topic Modeling*** **(rCTM)** to learn the multiple probabilistic dependencies between topics.

1.1 Limitations of the Existing Work

Two limitations of the existing work on topic evolution relevant to this article are discussed in this section.

1.1.1 Single-Topic-Thread Evolution. A well-known mechanism for analyzing the temporal evolution of topics is the state space model for the dynamic topic modeling [9, 52], where the temporal dependency between evolving topics is captured by Gaussian distributions. Another widely used mechanism exploits a Dirichlet distribution [31, 40] to encode the temporal dependency. Despite their difference in encoding the temporal development of topics, one common limitation lies in their single-topic-thread assumption as mentioned above. This violates the nature of many real cases.

As noted in Figure 1, the left side presents a topic evolutionary process following the single-topic-thread assumption, where each topic develops itself in a single thread, and the description words evolve in different slices. For example, the evolution of content about algorithm depends only on its own past state and ignores the influence of other prior topics such as **Natural Language Process (NLP)** and **Computer Vision (CV)**. This oversimplified evolution model does not reflect the reality in the real world. In contrast, the right side in the figure corresponds to an example of multi-thread-dependent evolutionary process, where the content on algorithm not only develops itself but also significantly influences NLP and CV. Further, the content on CV in the last slice

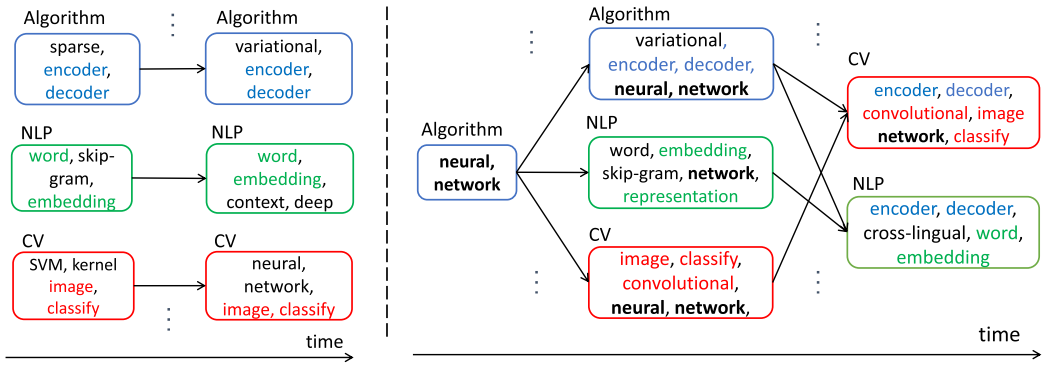


Fig. 1. An example of topic evolving process in terms of the conventional dynamic modeling (left) and the proposed rCTM (right), where the left side denotes topics evolving under the single-topic-thread assumption, while the right side corresponds to the multi-topic-thread evolution.

evolves not only from its past content but also being influenced by algorithm. Such multi-thread influence on the posterior topics is reinforced by the highlighted common words, for example, the content on CV in the last slice shares common words from the prior topics of algorithm and CV.

This example reinforces the fact that the development of topics is not constrained in one thread, rather, multiple topics are interactively coupled with each other [13, 27]. Without thoroughly encoding the complex temporal dependencies between evolving topics, the detected topic sequence from those conventional dynamic models might be defective. Therefore, our work aims to address this problem and learn multiple probabilistic dependencies between topics.

1.1.2 Fixed Topic Number in Documents. With some old topics phasing out and new ones coming in, the number of topics in each time-slice can be significantly different. Though existing studies [2, 51] are enabled to automatically learn the overall topic number for a collection of documents, they ignore the fact that each specific document from the collection may only involve a very small subset of those topics. Associating all topics with each individual document may cause the topic sparsity problem. Taking the collection of the published conference papers in a year as an example, the overall involved topics are diverse and numerous, where each individual paper is only related to very few of those topics, and the topics vary from paper to paper. It is clear that the traditional practice of assigning all topics to each document is very inappropriate, resulting in noisy topics assigned to a document and thus degrading the performance. Typically, the problem gets worse in the task of sequential short texts [37]. The prior work [40, 60] mitigates the sparsity problem to some extent by restricting one-topic assignment for a document; however, such a setting ignores the case of long documents which may contain more topics. Therefore, in terms of topic settings for both document chunks and individual documents, a unified and powerful mechanism is required to simultaneously infer both the overall topic number for each slice and the sparse topic number for individual documents.

1.2 Our Contributions

Motivated by the above discussion, this article introduces a new Bayesian sequential model rCTM over sequential documents through the following proposals.

First, we assume that a topic ϕ_{k_t} ($k_t \in \{1, \ldots, K_t\}$) at slice t evolves from all prior topics Φ_{t-1} of slice $t-1$ with the corresponding coupling weight $\beta_{k_{t-1}k_t}$ w.r.t. a Dirichlet distribution, and the distinguishable weights associated with the prior topics are learned from hierarchical Gamma

distributions. The proposal induces a new and more flexible framework that topic ϕ_{k_t} jointly depends on multiple prior topics, and a prior topic $\phi_{k_{t-1}}$ could also contribute to multiple topics at step t, which breaks the single-topic-thread limitation of the existing dynamic topic modelings [2, 9, 31, 40, 52]. Hence, the complex multi-topic-thread dependencies between time-evolving topics are thoroughly encoded by this proposal.

Second, the above new proposal of the multi-coupling relationships between evolving topics induces an unexplored and intractable inference problem, which significantly challenges the existing inference techniques. To fully solve this problem, we propose a novel solution with a set of novel data augmentation and marginalization techniques, which is the main novel contributions of this article. Our solution also discloses that the coupling weight between consecutive topics is indeed indicated by their shared latent word occurrences, and accordingly a novel negative binomial distribution is incorporated into the inference framework to obtain the latent word occurrences. Finally, with novel data augmentation, the joint multi-dependency between topics is discomposed into separated relationships and each coupling weight turns to be measurably independent, leading to a fully conjugate and interpretable Bayesian model.

Third, with the update of sequential document chunk, no one knows its optimal topic setting at each time-slice. In addition, each document only talks about a sparse number of topics, which remains unknown and varies from document to document. To fully tackle these problems, we leverage a nonparametric prior, a latent **Indian Buffet Process (IBP)** compound distribution [17, 56], to solve the sparsity problem over the document–topic matrix. In addition to the unbound topic number at each slice, the mechanism of IBP allows each document to contain its customized latent topics without bias.

With the aid of novel data augmentation and marginalization techniques, a new Gibbs sampler with a backward–forward filter algorithm is proposed to approximate latent time-evolving parameters. In this algorithm, at each iteration latent word counts are propagated backward from slice T to the initial slice, and the latent parameters are drawn forward from the initial slice to slice T with updated word counts. To validate the significance of multi-topic coupled dependencies from the prior topics, we design a variant model injected by a dropout technique from neural networks to prune the couplings with the prior topics. We explore both synthetic and real-world datasets with varying document lengths to evaluate the performance of rCTM against the competitive baselines. The extensive experimental results confirm the superiority of rCTM in terms of the low per-word perplexity, high topic coherence and better document time prediction.

To our best knowledge, this is the first article to address the *coupled topic modeling* problem, to which we make the following novel contributions:

— A new and general framework of encoding multi-topic-thread evolution is proposed for sequential document analysis, where a topic in the current slice may be flexibly influenced by multiple prior topics, and also develop into multiple threads with corresponding weights in the subsequent slice.
— A novel solution with data augmentations is presented to solve the unexploredly intractable problem and thoroughly decode the complex multi-dependencies between topics. rCTM thus enjoys a full conjugacy, where not only the evolution of topics across the slices but also their coupling relationships are efficiently captured in a closed form.
— Without the manual setting of the topic number, a nonparametric mechanism, a latent IBP compound distribution, is leveraged to automatically learn the whole topic number for a document chunk as well as the sparse topic numbers for individual documents. Such a mechanism solves the topic sparsity problem and flexibly accommodates both long and short documents.

2 THE PROPOSED MODEL

We discretize a collection of temporally sequential documents into T time slices $\{\mathbf{d}_t|1 \le t \le T\}$, where $\mathbf{d}_t$ is the document chunk of the t-th slice with $|\mathbf{d}_t|$ documents, and each document in the chunk is represented by a bag-of-words with t-th timestamp. Given the sequential documents, the word dictionary with V unique words is predefined. Before introducing our multi-topic-thread model, we define some notations and functions.

In what we present below, vectors and matrices are denoted by bold-faced lowercase and capital letters respectively and scalar variables are written in italic. $Dir_V()$, $Gam()$, $Mult()$, $Pois()$, and $Bern()$ stand for the V-dimensional Dirichlet, Gamma, multinomial, Poisson, and Bernoulli distribution, respectively. For a tensor $X \in \mathbb{Z}^{K_1 \times K_2 \times K_3}$, the (k_1, k_2, k_3) entry is denoted by $x_{k_1 k_2 k_3}$. Also $x_{k_1 k_2 \cdot} = \sum_{k_3}^{K_3} x_{k_1 k_2 k_3}$ and $x_{k_1 \cdot \cdot} = \sum_{k_2}^{K_2} \sum_{k_3}^{K_3} x_{k_1 k_2 k_3}$.

2.1 The Multi-Topic-Thread Generative Process

The proposed *rCTM* with multiple threads consists of two important integrated components: (1) the topic proportion learning, which automatically determines the total number of topics over the slice and sparsifies the affinity between topics and documents, and (2) the multi-topic-thread evolution, which incorporates the joint multiple dependencies between consecutive topics.

Topic proportion learning. Given a document chunk $\mathbf{d}_t$ of slice t, the hidden topics not only evolve from prior slice $t-1$, but may also come as new. Hence, the topic number K_t may change from slice to slice. In the existing work, the **Hierarchical Dirichlet Process (HDP)** [51] is widely used to determine the topic number. However, the HDP induces a rich-gets-richer problem, such that the infrequent topics are always overwhelmed by the popular ones [56]. For example, an article from the conference paper collections on Bayesian network could be dominated by the popular topic of neural network in its topic assignment. Furthermore, the HDP ignores the topic sparsity problem for individual documents, which may bring noise topic intruding in the topic assignment.

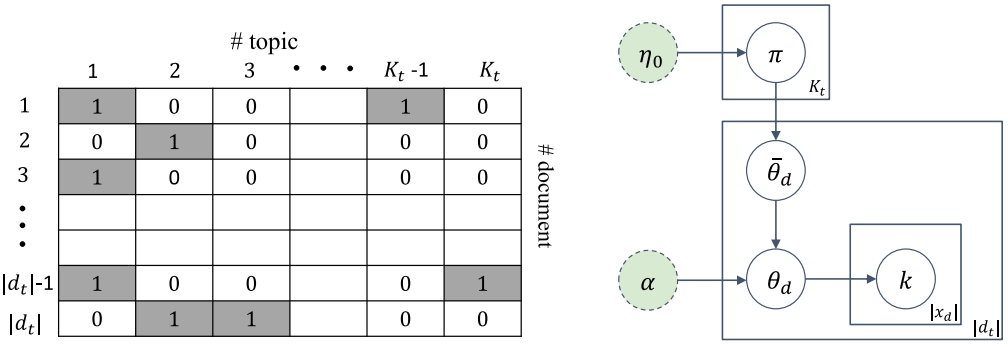
To resolve the above mentioned problems, the latent IBP Compound Distribution is exploited in the proposed model to get rid of the rich-gets-richer harm and boost the rare topics in the topic assignment of documents. In addition, it also enables a sparsity mechanism for each document to select its customized topics via the Bernoulli technique.

In detail, as shown in Figure 2(a), the sparsification of document-topic affinity is specified by a $|\mathbf{d}_t| \times K_t$ matrix $\overline{\theta}$, entries of which are stochastic variables that entry=1 indicates the affinity is true, otherwise false. Hence, not only the overall topic number K_t but also the affinity matrix $\overline{\theta}$ are stochastic variables which need to be inferred simultaneously in the learning process.

The generative process of topic proportion follows the procedure below:

$$\begin{aligned} \pi &\sim IBP(\eta_0), & \overline{\theta}_{dk_t} &\sim Bern(\pi), \\ \theta_d &\sim Dir_K(\overline{\theta}_d \odot \alpha), & k_t &\sim Mult(\theta_d), \end{aligned} \tag{1}$$

where $\odot$ is the element-wise Hadamard product and IBP is the Indian Buffet Process [17, 19], and other notations are presented in Table 1. The process first generates a probability matrix π via the IBP mechanism (the principles will be introduced below); next, taking π as the prior, a sparse document–topic affinity matrix $\overline{\theta}$ is produced via the Bernoulli distribution, indicating document d selects topic k_t if $\overline{\theta}_{dk_t} = 1$, otherwise they have no affinity; then, drawing the topic distribution θ_d for document d via the Dirichlet distribution by taking $\overline{\theta}_d$ as the prior; after that, drawing topic k_t for document d via the multinomial distribution; finally drawing words via the multinomial distribution with the word distribution ϕ_{k_t}, which is introduced in the following component.



(a) The sparse document-topic affinity matrix $\bar{\theta}$ for document chunk $\mathbf{d}_t$.

(b) Graphical representation of topic proportions at slice t.

Fig. 2. The construction of topic proportions at slice t. In Figure(a), each document in $\mathbf{d}_t$ contains its customized topics marked with 1 in the shaded color, and those topics excluded are left as 0 in blank, which are determined via the IBP mechanism. Figure(b) presents the graphical representation of topic proportion construction, where the circles with dash lines indicate the specified hyper-parameters, and the rest denote latent variables.

Table 1. Summary of Notations

Symbol	Description
η_0	the hyper-parameter of the IBP
α	the hyper-parameter of Dirichlet distribution
π	the parameter of the Bernoulli distribution
$\bar{\theta}_d$	the vector of sparse topic affinity with document d
θ_d	topic proportions for document d
ϕ_{k_t}	word distribution of topic k_t at slice t
$\beta_{k_{t-1}k_t}$	evolutionary coupling weight between topic k_{t-1} and k_t
Φ_t	topic set at slice t
$\mathbf{B}_{t-1,t}$	coupling matrix for topics between slice $t-1$ and t
η	hyper-parameter of Dirichlet distribution at slice 1
K_t	the inferred total topic number at slice t
x_{dw}	the observed word w's occurrence in the document d
x_{dwk_t}	the word w's occurrence in the document d assigned to the topic k_t
$\mathbf{x}_{k_t}$	the vector representation of word occurrence assigned to topic k_t
y_{wk_t}	the auxiliary latent word w's occurrence assigned topic k_t via *CRT*
$y_{wk_tk_{t-1}}$	the propagated word w's occurrence from topic k_t to k_{t-1}
$\mathbf{z}_{k_t}$	the vector representation of propagated word occurrence from $t+1$ slice
$y_{\cdot k_tk_{t-1}}$	the sum of propagate word counts from topic k_t to k_{t-1}
ξ_{k_t}	auxiliary variable from the beta distribution
r_{k_t}	shape parameter of the Gamma distribution
c_t, c_0	rate parameter of the Gamma distributions
a_0, b_0, e_0, d_0, r_0	hyper-parameters of the Gamma distributions

Now, we introduce in detail the first step of how to obtain the probability π via the IBP. Assume there are N customers in the restaurant, and each customer encounters a buffet consisting of

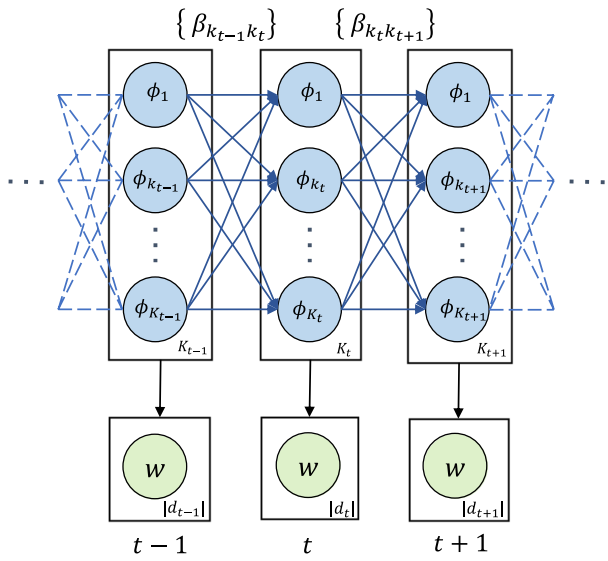


Fig. 3. Graphical representation of recurrent coupled topic evolution crossing three consecutive slices from $t-1$ to $t+1$, where $\phi_{k_t}(k_t \in \{1, 2, \ldots, K_t\})$ represents the hidden topic k at time t denoted by blue circles, the coupling relationships $\{\beta_{k_{t-1}k_t}\}$ between consecutive topics marked by the blue arrows denote their temporal dependencies, and $\mathbf{w}$ in green color represents the observed words from document chunk $\mathbf{d}_t$.

infinitely many dishes arranged in a line. The first customer starts at the left of the buffet and takes a serving from each dish, stopping after $Pois(\eta_0)$ number of dishes as his plate is full. The i-th customer moves along the buffet and samples dishes with proportion to their popularity $\frac{m_k}{i}$, where m_k is the number of previous customers who have taken the k-th dish. At the end of all previously sampled dishes, the i-th customer tries $Pois(\frac{\eta_0}{i})$ number of new dishes.

By analogy to the IBP, the sparse document–topic affinity matrix $\overline{\theta}$ with $|\mathbf{d}_t|$ documents corresponds to the N customers' specific choices over infinite dishes by taking the limit $K_t \to \infty$. The probability matrix π generating $\overline{\theta}$ corresponds to the probabilities of all customers' selection of dishes. Based on π, each document is thus allowed to sequentially select its customized topics via the Bernoulli distribution. Topic proportions θ_d are generated by the Hadamard product between $\overline{\theta}_d$ and hyper-parameter α via a Dirichlet distribution. That means only those selected topics ($\overline{\theta}_{dk_t} = 1$) are endowed with weight α to constitute the topic proportion for document d (i.e., sparsified the document-topic affinity matrix). The graphical representation of topic proportion construction is presented in Figure 2(b).

Multi-Topic-Thread evolution. The other important component of the generative process is how to encode multiple topic dependencies crossing slices. Figure 3 presents a simple scenario of coupled topic evolution crossing three consecutive slices.

At the initial slice $t = 1$, without any prior dependency, the topic ϕ_{k_1} $(k_1 \in \{1, \ldots, K_1\})$ is sampled from the Dirichlet distribution parameterized by η. At slice t, the topic ϕ_{k_t} $(k_t \in \{1, \ldots, K_t\})$ is assumed to evolve from the prior topics with the corresponding coupling weights $\beta_{k_{t-1}k_t}$ via the Dirichlet distribution, where $\beta_{k_{t-1}k_t}$ is drawn from a Gamma distribution to measure the evolutionary closeness to the topic in the prior slice. At the final slice $t = T$, topic ϕ_{k_T} $(k_T \in \{1, \ldots, K_T\})$ evolves depending on the prior topics at $T-1$. Given the defined recurrent topics, words $\mathbf{w}$ from the document chunk $\mathbf{d}_t$ are accordingly generated via the multinomial distributions at each slice.

The recurrent coupled topic sequences smoothly evolve across the T slices according to the following generative process,

$$
\begin{aligned}
&(x_{\cdot wk_1})_{w=1}^{V} \sim Mult(\phi_{k_1}), \quad \phi_{k_1} \sim Dir_V(\eta),\\
&\quad \cdots\\
&(x_{\cdot wk_t})_{w=1}^{V} \sim Mult(\phi_{k_t}), \quad \phi_{k_t} \sim Dir_V\left(\sum_{k_{t-1}=1}^{K_{t-1}} \beta_{k_{t-1}k_t}\phi_{k_{t-1}}\right), \quad \beta_{k_{t-1}k_t} \sim Gam(r_{k_{t-1}}, 1/c_t),\\
&\quad \cdots\\
&(x_{\cdot wk_T})_{w=1}^{V} \sim Mult(\phi_{k_T}), \quad \phi_{k_T} \sim Dir_V\left(\sum_{k_{T-1}=1}^{K_{T-1}} \beta_{k_{T-1}k_T}\phi_{k_{T-1}}\right), \quad \beta_{k_{T-1}k_T} \sim Gam(r_{k_{T-1}}, 1/c_T),
\end{aligned}
\tag{2}
$$

where Gam(-,-) is the Gamma distribution with shape and scale parameters. We further impose Gamma priors on the following variables: $r_{k_{t-1}} \sim Gam(r_0/K_{t-1}, 1/c_0)$, $c_t \sim Gam(e_0, 1/d_0)$ and $c_0 \sim Gam(a_0, 1/b_0)$, where a_0, b_0, d_0, e_0 and r_0 are specified hyper-parameters.

The idea of our recurrent modeling of multiple coupled topic sequences is summarized as follows:

- From a forward–backward view, the proposed model resembles the stochastic feedforward network [50], where the input is a topic set $\{\phi_{k_1}\}_{k_1=1}^{K_1}$ at slice 1, the output is topics $\{\phi_{k_T}\}_{k_T=1}^{K_T}$, the weight matrices are $\{\beta_{k_{t-1}k_t}\}_{k_{t-1},k_t}^{K_{t-1},K_t}$ and the activation functions are Dirichlet distributions.
- When learning the word distribution of topic k_t ($t > 1$), the mixture of prior topics $\{\phi_{k_{t-1}}\}_{k_{t-1}=1}^{K_{t-1}}$ serves as the prior knowledge to initialize topic k_t via a Dirichlet distribution, and the coupling weights $\{\beta_{k_{t-1}k_t}\}_{k_{t-1}}^{K_{t-1}}$ identify the different contributions of prior topics to topic k_t.
- According to the expectation of a Dirichlet distribution, topic ϕ_{k_t} is expected to be the weighted arithmetic mean of prior topics at slice $t-1$, $E(\phi_{k_t}) = \frac{\sum_{k_{t-1}=1}^{K_t} \beta_{k_{t-1}k_t}\phi_{k_{t-1}}}{\sum_{k_{t-1}=1}^{K_t} \beta_{k_{t-1}k_t}}$, implying that the evolution of topic ϕ_{k_t} jointly depends on multiple prior topics with the corresponding weights rather than on a single past topic, and the prior topic $\phi_{k_{t-1}}$ ($k_{t-1} \in \{1, \ldots, K_{t-1}\}$) also contributes to multiple topics at slice t. The coupling weight $\beta_{k_{t-1}k_t}$ is noted to play an important role in measuring the evolutionary distance between two word distributions of topic k_{t-1} and k_t. In addition, this expectation indicates coupling weights $\{\beta_{k_{t-1}k_t}\}_{k_{t-1}=1}^{K_{t-1}}$ associated with topic k_t are not shared with other parallel topics, which allows topics at slice t to evolve differently with flexible dependency on the common priors.
- The coupling weight $\beta_{k_{t-1}k_t}$ is drawn from a hierarchical Gamma prior (its shape parameter r_{k_t} is also drawn from a Gamma). Such a hierarchical design leads to more distinguishable and sparse coupling weights associated with topic ϕ_{k_t} [64].

2.2 A Dropout Technique

In the context of topic evolution with multiple threads, one question is naturally raised about how to validate the significance of multi-dependencies between evolving topic sequences, since each topic evolves from all prior topics. Further, one may argue that the salient coupling connections of one topic during evolving process are a small set and sparsely distributed in practice, e.g., in light of diverse and enormous topics inferred from the computer science articles last year, the topic about Bayesian network only connects with a small number of relevant topics by the salient coupling

weight, while the weights with most unrelated topics are small. Thus, could the proposed model distinguish the salient coupled topics from the less related ones by the weights? To answer this question, we develop a variant of the proposed model named rCTM-D as a comparison to rCTM.

In this approach, we borrow the dropout mechanism from the neural network and inject it into our Bayesian framework. Dropout [49] is one of the most popular and successful regularizers for deep neural network. It randomly drops out each neuron with a predefined probability at each iteration of stochastic gradient descent, to avoid the overfitting problem and reinforce the performance. In our solution, at each iteration of inference process, the topic node ϕ_{k_t} is attached with a probability ρ to drop out coupling connection with prior topics Φ_{t-1}, which is denoted as:

$$\phi_{k_t} \sim Dir_V(\psi_{k_t}), \qquad \psi_{k_t} = \sum_{k_{t-1}=1}^{K_{t-1}} (\beta_{k_{t-1}k_t}(1 - m_{k_{t-1}}))\phi_{k_{t-1}}, \qquad m_{k_{t-1}} \sim Bern(\rho), \tag{3}$$

where $m_{k_{t-1}}$ is the dropout indicator drawn from a Bernoulli distribution with parameter ρ. If $m_{k_{t-1}} = 0$, the coupling connection from prior topic k_{t-1} is preserved with its original weight; otherwise, the connection is dropped out and this prior topic would not participate in the inference to posterior topics.

Let us consider the dropout probability in two extreme cases. (1) If we set the dropout probability $\rho = 0$, then $m_{k_{t-1}} = 0$ ($k_{t-1} \in \{1, \ldots, K_{t-1}\}$), it means all coupling connections are preserved and rCTM-D is recovered to rCTM. (2) If $\rho = 1$, then $m_{k_{t-1}} = 1$ ($k_{t-1} \in \{1, \ldots, K_{t-1}\}$), rCTM-D is thus degraded to T separated topic modelings at each time-slice without any connections. Hence, we would give the dropout probability ρ within the range $(0, 1)$ in the rCTM-D, to see its performance with different ratios of coupling connection dropped out.

3 THE POSTERIOR INFERENCE

Since we induce *multi-topic-thread evolution*, the main challenge for the proposed rCTM is to solve the intractable problem and obtain a closed-form inference to recurrent topics Φ_t as well as their coupling matrix $\mathbf{B}_{t-1,t}$ at each time slice. Such a task has never been explored before. To tackle this problem, a set of auxiliary variables and data augmentation techniques are introduced. In this section, we propose a novel Gibbs sampler with a backward-forward filter algorithm to implement its inference process.

Sampling $\overline{\theta}$***:*** the sparse document-topic affinity matrix $\overline{\theta}$ could be sampled by marginalizing out θ and π_k. First, we note that if the word count $x_{dwk_t} > 0$, then $\overline{\theta}_{dk_t}$ must be 1 because it implies there exists at least one word assigned to topic k_t. Let vector $\overline{\theta}_{d(0)}$ represent the d-th row vector of $\overline{\theta}$ with entries of 0, and vector $\overline{\theta}_{k(0)}$ denote the k-th column vector of $\overline{\theta}$ with entries of 0. If $x_{d\cdot k_t} = 0$, the probability $\overline{\theta}_{dk_t} = 1$ is marginalized as,

$$P(\overline{\theta}_{dk_t} = 1 | \alpha, \eta_0) = \frac{B(\alpha|\overline{\theta}_{d(0)}| + |\overline{\theta}_{d(0)}|, \alpha)(|\overline{\theta}_{k(0)}| + \eta_0)}{B(\alpha|\overline{\theta}_{d(0)}|, \alpha)(|\mathbf{d}_t| - |\overline{\theta}_{k(0)}| + \eta_0)}, \tag{4}$$

where $B(-,-)$ denotes a beta distribution, $|\mathbf{d}_t|$ records the document number at slice t, $|\overline{\theta}_{d(0)}|$, $|\overline{\theta}_{k(0)}|$ record the number of 0 entries in the d-th row vector and k-th column vector of matrix $\overline{\theta}$ respectively, and η_0, α are specified hyper-parameters.

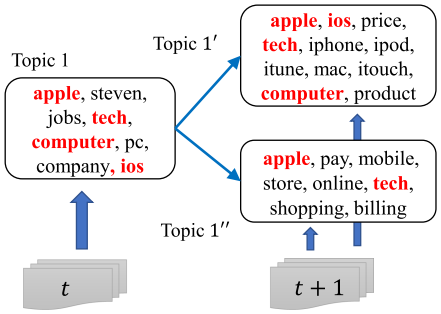
Sampling θ***:*** as we obtain the sparse document-topic affinity matrix $\overline{\theta}$, the topic proportion θ_d for document d is sampled from its conditional posterior distribution as,

$$\theta_d \sim Dir_K(\overline{\theta}_d \odot (\alpha + x_{d\cdot k_t})), \tag{5}$$

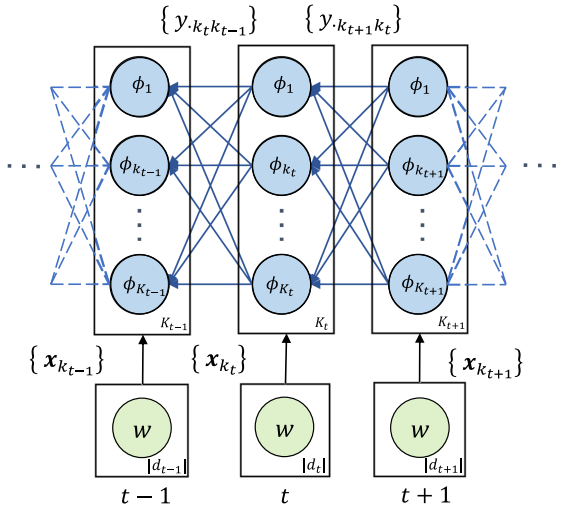
where $x_{d\cdot k_t}$ records the number of words in the document d assigned to the topic k_t.

Table. The difference of word occurrences (red) in the topic $1'$ between prior and no prior.

Word	*apple*	*ios*	*tech*	*computer*	Prior
Topic $1'$	18	10	15	12	**no**
⬇	**6**	**3**	**3**	**5**	-
Topic $1'$	20	19	17	12	**yes**



(a) A motivating example to decode the coupling relationship between consecutive topics.



(b) The inference process with backward propagation.

Fig. 4. In figure (a), each topic is represented by a set of description words, their occurrences with and without prior are respectively listed in black font, and the shared common word occurrences are denoted in red in the table. In figure (b), arrows between consecutive topics denote their shared latent word counts in the backward filter, which are annotated by $\{y_{\cdot k_t k_{t-1}}\}$ in blue, and $\mathbf{x}_{k_t}$ summarizes the inferred vector of word counts assigned to topic k_t in black.

Sampling x_{dwk_t}**:** the observed word w's occurrence in document d is denoted as x_{dw}, and we augment it as $x_{dw} = x_{dw\cdot} = \sum_{k_t} x_{dwk_t}$, indicating the number of word w in the document d assigned to topic k_t, which is sampled as,

$$x_{dwk_t} \sim Mult\left(x_{dw\cdot}, \left(\frac{\theta_{dk_t}\phi_{k_t w}}{\sum_{k_t}^{K_t}\theta_{dk_t}\phi_{k_t w}}\right)_{k_t=1}^{K_t}\right). \tag{6}$$

The vector $\mathbf{x}_{k_t}$ is defined as $\mathbf{x}_{k_t} = [x_{\cdot 1k_t}, x_{\cdot 2k_t}, \ldots, x_{\cdot Vk_t}]$, indicating the vector of all word occurrences from document chunk $\mathbf{d}_t$ assigned to the topic k_t, which is illustrated in Figure 4(b).

Challenges of Inference. We proceed to infer the latent parameters in the component of coupled topic evolution, which is the core part of the solution. There remain two demanding challenges to be solved for a tractable inference, which significantly challenge the existing inference approaches.

— To obtain an independent inference to coupling weight $\beta_{k_{t-1}k_t}$, it is vital to decompose the joint multiple dependencies into individual relationships associated with each prior topic.
— The essence of non-negative weight $\beta_{k_{t-1}k_t}$ connecting topic k_{t-1} to k_t remains unknown.

To solve these challenges, we induce a motivating example to illustrate it. As shown in Figure 4(a), Topic 1 naturally evolves into two different threads from slice t to $t+1$ via the proposed generative process, and their contents are represented by a set of frequent words. Though their word representations differ, it is found that the shared common words, such as "apple," "tech," and "computer," naturally chain the prior Topic 1 and its thread Topic $1'$ together. Indicated by the table in Figure 4(a), the occurrences of these common words in Topic $1'$ with the prior influence of Topic 1 are distinguished from Topic $1'$ without such prior dependency. An insightful fact is found

that occurrences of common words in Topic 1′ with the prior could be decoded into two parts. One part of these occurrences is directly from the documents at time $t+1$, and the other is implicitly contributed from the prior Topic 1, denoted by the numbers in red from the Table. Without such implicit word-level sharing, the dependency relationship between Topic 1 and its thread would not exist. Hence, it is concluded that coupling weight $\beta_{k_{t-1}k_t}$ connecting topic k_{t-1} and its subsequent thread k_t is essentially summarized by their shared latent word occurrences.

Based on the above insightful observations, it is of significance to derive the shared latent word counts between consecutive topics. Since topics at slice $t(1 < t < T)$ are recursively chained and inter-dependent on prior topics at $t-1$, the conventional inference techniques [20, 40], which are implemented as independent at each slice, ignore such *recursive dependency* between slices and they are inapplicable for such an inference task. Therefore, we design a novel backward–forward filter to fully solve it and achieve the tractable inference. In the backward filter, the smart data augmentation techniques unfreeze the limitation of *recursive dependency*, and derive the shared latent word counts between consecutive slices. Then the time-evolving parameters are naturally inferred in a closed-form in the forward filter.

Backward propagating the latent counts. We start from time slice T since no more latent variables depend on it. By integrating out ϕ_{k_T}, we obtain the likelihood of latent word counts $(x_{\cdot wk_T})_{w=1}^{V}$ according to the conjugacy between the Dirichlet and multinomial distributions.

$$\mathcal{L}(x_{\cdot wk_T})_{w=1}^{V} \sim DirMult(\psi_{k_T}), \qquad \psi_{k_T} = \sum_{k_{T-1}=1}^{K_{T-1}} \beta_{k_{T-1}k_T}\phi_{k_{T-1}}, \tag{7}$$

where the multi-dependency ψ_{k_T} associated with all prior topics always appears in the sum form as the parameter of Dirichlet. Since we could not directly obtain the individual dependency $\beta_{k_{t-1}k_t}$ associated with each prior topic, we introduce an auxiliary variable $\xi_{k_T} \sim B(x_{\cdot\cdot k_T}, \psi_{\cdot k_T})$, and further augment $DirMult$. The joint likelihood of $(x_{\cdot wk_T}, \xi_{k_T})$ takes the following form [1],

$$\begin{aligned}\mathcal{L}((x_{\cdot wk_T})_{w=1}^{V}, \xi_{k_T}) &= \mathcal{L}(x_{\cdot wk_T})_{w=1}^{V} \times B(\xi_{k_T}|x_{\cdot\cdot k_T}, \psi_{\cdot k_T}) \\ &\propto \prod_{w=1}^{V} NB(x_{\cdot wk_T}|\psi_{vk_T}, \xi_{k_T}),\end{aligned} \tag{8}$$

where $NB(-,-)$ is the negative binomial distribution, and $B(-,-)$ is the beta distribution. With the auxiliary variable introduced, the variable $x_{\cdot wk_T}$ now follows the negative binomial distribution, which plays a critical role in bridging the Dirichlet and Poisson distributions. Hence, Lemma 3.1 is defined in the following, which presents the transformation relationship from a negative binomial to a Poisson distribution. The property of the Poisson distribution is thus able to be enjoyed when disentangling the joint dependency relationship after the transformation.

Lemma 3.1. *If $m \sim NB(r,p)$ represents that m follows a negative binomial distribution, then the conditional posterior of l given m and r is denoted as $(l|m,r) \sim CRT(m,r)$, a Chinese restaurant table (CRT) count random variable, which can be generated via $l = \sum_{n=1}^{m} z_n$, $z_n \sim Bern(r/(n-1+r))$. It can also be augmented under a compound Poisson representation as $m = \sum_{t=1}^{l} u_t$, $u_t \sim Log(p)$, $l \sim Pois(-rlog(1-p))$ [2, 47].*

According to Lemma 3.1, a new variable $y_{wk_T} \sim CRT(x_{\cdot wk_T}, \psi_{wk_T})$ is obtained based on $x_{\cdot wk_T}$ and ψ_{wk_T}. With further data augmentation applied, an equivalent representation of y_{wk_T} under a compound Poisson form is expressed as,

$$y_{wk_T} \sim Pois(-\psi_{wk_T} ln(1-\xi_{k_T}). \tag{9}$$

Since $\psi_{wk_T} = \sum_{k_{T-1}=1}^{K_{T-1}} \beta_{k_{T-1}k_T}\phi_{vk_{T-1}}$ as defined by Equation (7), we feed it into the above equation, which is extended as

$$y_{wk_T} \sim Pois\left(-\sum_{k_{T-1}=1}^{K_{T-1}} \beta_{k_{T-1}k_T}\phi_{vk_{T-1}} ln(1-\xi_{k_T})\right). \tag{10}$$

We now introduce another auxiliary variable $y_{wk_Tk_{T-1}}$ which is augmented from $y_{wk_T} = y_{wk_T\cdot} = \sum_{k_{T-1}=1}^{K_{T-1}} y_{wk_Tk_{T-1}}$, and the above Equation (10) is thus represented as follows according to the property of Poisson distribution,

$$y_{wk_Tk_{T-1}} \sim Pois(-\beta_{k_{T-1}k_T}\phi_{wk_{T-1}} ln(1-\xi_{k_T})), \tag{11}$$

where the joint coupling dependency is successfully decomposed into separated relationships thanks to the merit of data augmentation technique and the property of Poisson distribution. Since the auxiliary variable $y_{wk_Tk_{T-1}}$ is augmented from the variable y_{wk_T}, now we define Lemma 3.2 in the following to present the relationship between Poisson and multinomial distributions.

Lemma 3.2. *If* $y_\cdot = \sum_{n=1}^{N} y_n$, *where* $y_n \sim Pois(\theta)$ *are independent Poisson-distributed random variables, then* $(y_1, \ldots, y_N) \sim Mult(y_\cdot, (\frac{\theta_1}{\sum_{n=1}^{N}\theta_n}, \ldots, \frac{\theta_N}{\sum_{n=1}^{N}\theta_n}))$ *[65].*

Thus, $y_{wk_Tk_{T-1}}$ is distributed as *Mult* via Lemma 3.2, which is expressed as,

$$y_{wk_Tk_{T-1}} \sim Mult\left(y_{wk_T}, (\frac{\beta_{k_{T-1}k_T}\phi_{wk_{T-1}}}{\sum_{k_{T-1}=1}^{K_{T-1}}\beta_{k_{T-1}k_T}\phi_{wk_{T-1}}})_{k_{T-1}=1}^{K_{T-1}}\right), \tag{12}$$

where $y_{wk_Tk_{T-1}}$ is successfully obtained to denote the shared latent word w' occurrence between topic k_T and K_{T-1} , which indicates the numbers to be inferred denoted in red from the table of Figure 4(a).

We now induce the auxiliary variable $z_{wk_{T-1}}$, which is defined as $z_{wk_{T-1}} = \sum_{k_T=1}^{K_T} y_{wk_Tk_{T-1}}$. It is viewed as latent word counts propagated from topic set at slice T. Thus the vector $\mathbf{z}_{k_{T-1}}$ is defined as $\mathbf{z}_{k_{T-1}} = [z_{1k_{T-1}}, z_{2k_{T-1}}, \ldots, z_{Vk_{T-1}}]$. Another highlighted variable is $y_{\cdot k_Tk_{T-1}}$ obtained via $y_{\cdot k_Tk_{T-1}} = \sum_{w=1}^{V} y_{wk_Tk_{T-1}}$ to summarize the sum of shared latent word counts between topic k_T and k_{T-1}, which is computed in advance and cached to be used in the forward filter.

As we continue propagating backward from $t = T-1, \ldots, 2$, the latent word count vectors $\mathbf{z}_{k_{T-2}}, \ldots, \mathbf{z}_1$ are sequentially obtained. It's worth noting that since no more document chunks after slice T, there is no propagated word count at slice T such that $\mathbf{z}_{k_T} = \mathbf{0}$.

In conclusion, the propagating process between consecutive evolving topics from slice $t+1$ to t is summarized as: (1) the latent word count $y_{wk_{t+1}}$ is firstly derived from $x_{wk_{t+1}}$ via the *CRT* distribution, (2) then we distribute $y_{wk_{t+1}}$ according to the *Mult* distribution to obtain the latent count $y_{wk_tk_{t+1}}$, and (3) finally $y_{wk_tk_{t+1}}$ is aggregated to form the latent counts z_{wk_t} at slice t. This process is illustrated in Figure 4(b).

Forward sampling the latent variables. Conditioned on the propagated auxiliary values $\{\mathbf{z}_{k_t}\}_{t=1}^{T-1}$, $\{y_{\cdot k_2k_1}, \ldots, y_{\cdot k_Tk_{T-1}}\}$ and $\{\xi_{k_2}, \ldots, \xi_{k_T}\}$ obtained via the backward propagating filter. We start sampling the latent variables by performing a forward sampling pass from $t = 1, \ldots, T$.

Sampling Φ***:*** based on the conjugacy between the Dirichlet and multinomial distributions, the topics ϕ_{k_1} ($k_1 \in \{1, \ldots, K_1\}$) at slice $t = 1$ is marginalized from its conditional posterior,

$$\phi_{k_1} \sim Dir_V(\eta + \mathbf{x}_{k_1} + \mathbf{z}_{k_1}), \tag{13}$$

where $\mathbf{x}_{k_1} = [x_{\cdot 1k_1}, x_{\cdot 2k_1}, \ldots, x_{\cdot Vk_1}]$ is the word occurrence vector inferred from document chunk $\mathbf{d}_1$ at slice 1 via Equation (6), and $\mathbf{z}_{k_1} = [z_{1k_1}, z_{2k_1}, \ldots, z_{Vk_1}]$ denotes the propagated word count vector from slice 2 to slice 1.

At time $1 < t \leq T$, ϕ_{k_t} ($k_t \in \{1, \ldots, K_t\}$) is sampled as,

$$\phi_{k_t} \sim Dir_V\left(\sum_{k_{t-1}=1}^{K_{t-1}} \beta_{k_{t-1}k_t}\phi_{k_{t-1}} + \mathbf{x}_{k_t} + \mathbf{z}_{k_t}\right), \tag{14}$$

where $\mathbf{x}_{k_t}$ is the word count vector inferred from $\mathbf{d}_t$ at slice t, and $\mathbf{z}_{k_t}$ denotes the propagated word count vector from slice $t+1$ to t. It is noted $\mathbf{z}_{k_T} = \mathbf{0}$ at time slice T.

***Sampling* B:** indicated by Equation (11), $\beta_{k_{T-1}k_T}$ still entwines with $\phi_{wk_{T-1}}$ in the parameter of Poisson distribution. However, via $\sum_{w=1}^{V}\phi_{wk_t} = 1$, it is marginalized as,

$$y_{\cdot k_T k_{T-1}} \sim Pois(-\beta_{k_{T-1}k_T} ln(1-\xi_{k_T})). \tag{15}$$

Recall the prior distribution of $\beta_{k_{t-1}k_t}$, which is defined as $\beta_{k_{t-1}k_t} \sim Gam(r_{k_{t-1}}, 1/c_t)$ in Equation (2), thus it is marginalized via the conjugacy between the Poisson and Gamma distributions,

$$\beta_{k_{t-1}k_t} \sim Gam(y_{\cdot k_t k_{t-1}} + r_{k_{t-1}}, 1/(c_t - ln(1-\xi_{k_t}))), \tag{16}$$

where $y_{\cdot k_t k_{t-1}}$ is the sum of propagated word counts and cached in the backward filter via $y_{\cdot k_t k_{t-1}} = \sum_{w=1}^{V} y_{wk_t k_{t-1}}$, and the auxiliary variable ξ_{k_t} is also induced in the backward filter.

Through a series of novel data augmentation techniques, the inference of recurrent topic ϕ_{k_t} and its coupling strength $\{\beta_{k_{t-1}k_t}\}_{k_{t-1}}^{K_{t-1}}$ are finally tractable at each slice. According to the expectation of Gamma distribution, it is derived that $\beta_{k_{t-1}k_t} \approx (y_{\cdot k_t k_{t-1}} + r_{k_{t-1}})/(c_t - ln(1-\xi_{k_t}))$, implying that the sum of propagated common word counts between consecutive topics k_t and k_{t-1} is an important indicator to the coupling weight $\beta_{k_{t-1}k_t}$.

***Sampling* $r_{k_{t-1}}$:** recall $r_{k_{t-1}}$ is drawn from its prior distribution via $r_{k_{t-1}} \sim Gam(r_0/K_{t-1}, 1/c_0)$ (cf. the Section 2.1), and its likelihood distribution is defined as $\beta_{k_{t-1}k_t} \sim Gam(r_{k_{t-1}}, 1/c_t)$ in Equation (2), the inference of $r_{k_{t-1}}$ incurs the non-conjugate problem between the Gamma and Gamma distributions, hence, we induce Lemma 3.3 in the following to help solve it.

Lemma 3.3. *if* $x_i \sim Pois(m_i r_2)$, $r_2 \sim Gam(r_1, 1/c_0)$, $r_1 \sim Gam(a_0, 1/b_0)$, *then* $(r_1|-) \sim Gam(a_0 + l, 1/(b_0 - log(1-p)))$, *where* $(l|x, r_1) \sim CRT(\sum_i x_i, r_1)$ *and* $p = \frac{\sum_i m_i}{c_0 + \sum_i m_i}$ *[1, 2].*

Based on Equation (15) and the above prior and its likelihood distribution, $r_{k_{t-1}}$ is sampled via Lemma 3.3,

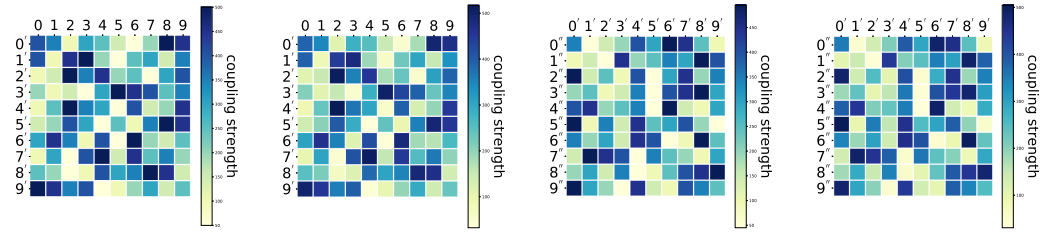
$$\begin{aligned} r_{k_{t-1}} &\sim Gam(r_0/K_{t-1} + l_{k_{t-1}}, 1/(c_0 - log(1-p))), \\ l_{k_{t-1}} &\sim CRT\left(\sum_{k_t=1}^{K_t} y_{\cdot k_t k_{t-1}}, r_{k_{t-1}}\right), \qquad p = \sum_{k_t=1}^{K_t} m_{k_t} \Big/ \left(\sum_{k_t=1}^{K_t} m_{k_t} + c_0\right), \end{aligned} \tag{17}$$

where $m_{k_t} = -ln(1-\xi_{k_t})$.

***Sampling* c_t, c_0:** given the conjugacy between the Poisson and Gamma distributions, c_t and c_0 are sampled respectively as,

$$\begin{aligned} c_t &\sim Gam\left(e_0 + \sum_{k_{t-1}}^{K_{t-1}} r_{k_{t-1}}, 1 \Big/ \left(\sum_{k_{t-1}}^{K_{T-1}} \beta_{k_{t-1}k_t} + d_0\right)\right), \\ c_0 &\sim Gam(e_0 + r_0/K_{t-1}, 1/(r_{k_{t-1}} + d_0)), \end{aligned} \tag{18}$$

where K_{t-1} is inferred topic number from a topic proportion process via the latent IBP compound distribution, $r_{k_{t-1}}$ and $\beta_{k_{t-1}k_t}$ are sampled from the prior steps, and the rest variables e_0, d_0, r_0 are specified hyper-parameters.



(a) The true coupling weights between topics at $t = 1$ and $t = 2$

(b) The estimated coupling weights between topics at $t = 1$ and $t = 2$

(c) The true coupling weights between topics at $t = 2$ and $t = 3$

(d) The estimated coupling weights between topics at $t = 2$ and $t = 3$

Fig. 5. The comparison of the true and estimated coupling weights between consecutive topics.

The whole Gibbs sampling with a backward–forward filter is presented in Algorithm 1. At each iteration, $\mathbf{x}_{k_t}$ is firstly sampled via the *Mult* distribution from the document chunk $\mathbf{d}_t$ at each slice. During the backward filter, the auxiliary variable ξ_{k_t} and y_{wk_t} are induced, and the sequence of propagated word counts $\mathbf{z}_{k_t}$ is obtained sequentially from slice T to slice 2 via repeated data augmentation and marginalization techniques. Conditioned on latent counts from the above filter procedure, the recurrent topics Φ_t and their coupling matrix $\mathbf{B}_{t-1,t}$ are updated in a closed-form at each slice. These steps are repeated *MaxIteration* times until the joint posterior distribution converges. The latent parameters are thus estimated based on the stable samples.

By far, we have introduced our novel recurrent multi-topic modeling and the corresponding novel and effective inference method. In the backward filter, the adoption of novel *NB* augmentation into the dynamic Dirichlet chain is non-trivial, which bridges the gap between the Dirichlet and Poisson distributions. Such an infusion plays an important role in unfreezing the limitation of *recursive dependency* and deriving the shared latent word counts between consecutive topics, leading to an efficient and tractable inference for recurrent topics and their multi-dependencies. Note that none of the existing work on the temporal topic modeling proposes such assumptions that naturally fit the complex sequential data, facilitates the interpretability of latent states, and yields a closed-form and straight-forward update in the inference.

4 EXPERIMENTS

4.1 Experiments with Synthetic Data

To verify whether rCTM is capable to capture the multi-thread coupling weights between recurrent topics, we manually create a synthetic dataset with predefined coupling relationships over three slices.

Referring to the emprical study [2], 3 $1,000 \times 1,000$ document-word matrices with $1,000$ documents over the vocabulary size of $1,000$ are created sequentially at each slice according to the following steps. At slice $t = 1$, we initialize ten topics $\{0, 1, \ldots, 9\}$ via the Dirichlet distributions, and randomly use them to generate the first $1,000$ documents. Given the specified coupling weight matrix in Figure 5(a), topics $\{0', 1', \ldots, 9'\}$ at slice $t = 2$ are produced according to the proposed evolutionary process, and randomly to generate the second $1,000$ documents. Similarly, topics $\{0'', 1'', \ldots, 9''\}$ at slice $t = 3$ and the corresponding document chunk are also created.

Based on three synthetic document–word matrices, we utilize the proposed rCTM to recover the recurrent topics and their coupling weights to see whether the proposed model is able to decode the intricate dependency relationship between topics. Due to the space limit, only the comparison between the true coupling weights and the estimated weights by the rCTM is indicated by Figure 5. Noted from Figure 5(a) and (b), the estimated 10×10 coupling weights between topics

ALGORITHM 1: Gibbs Sampling with the backward-forward filter

```
Initializing topic assignments randomly for all documents in {d_t}_{t=1}^T;
Initializing variables B_{t-1,t}, (r_{k_t})_{k_t=1}^{K_t}, c_t and c_0 at each slice;
for iter ∈ 1, 2, ..., MaxIteration do
    for d ∈ {d_t}_{t=1}^T do
        Sampling the document-specific θ̄_d, θ_d by Equation (4) and Equation (5);
        for w ∈ {w_1, ..., w_{|x_d|}} do
            Sampling the latent word count x_{dwk_t} by Equation (6);
        end
    end
    Backward propagating: initialize t = T;
    while t > 0 do
        Sampling the auxiliary variable ξ_{k_t} from beta distribution;
        Sampling the latent count y_{wk_T} via CRT;
        Sampling the latent count y_{wk_T k_{T-1}} by Equation (12);
        Caching the auxiliary variable y_{·k_t k_{t-1}} to use in the forward pass;
        t = t − 1;
    end
    Forward sampling: initialize t = 1;
    while t ≤ T do
        Sampling φ_{k_t} by Equation (13) or Equation (14);
        Sampling β_{k_{t-1} k_t} by Equation (16);
        Sampling r_{k_t} by Equation (17);
        Sampling c_t, c_0 by Equation (18);
        t = t + 1;
    end
end
```

$\{0, 1, \ldots, 9\}$ at slice $t = 1$ and topics $\{0', 1', \ldots, 9'\}$ at slice $t = 2$ precisely match with the true matrix. Similarly, the estimated coupling weights between topics $\{0', 1', \ldots, 9'\}$ at slice $t = 2$ and topics $\{0'', 1'', \ldots, 9''\}$ at slice $t = 3$ also highly resemble the true weights denoted by Figure 5(c) and (d). Furthermore, not only strong coupling but also weak dependency relationships between consecutive topics are successfully discriminated by rCTM. As the coupling weights between consecutive topics are precisely captured, the discovery of precise topics are naturally followed. This comparison highlights rCTM is capable to decode the intrinsic multi-thread dependency between evolving topics.

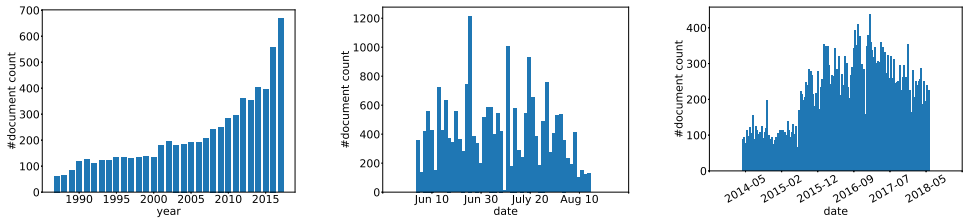
4.2 Experimental Setup with Real-world Data

We use five real-world datasets from different domains to evaluate all algorithms. The statistics of datasets are summarized in the Table 2.

— NIPS corpus [26]. This benchmark dataset consists of the abstracts of papers appearing in the NIPS conference from the year 1987 to 2017. After the standard pre-processing and the removal of the most frequent and the least words, the size of corpus is reduced to 6,753 documents and 4,434 unique vocabularies, and the average document length is about 50.

Table 2. The Statistics of Five Real-wold Datasets

Dataset	#document	#vocabulary	time span	#average document length
NIPS	6,753	4,434	1987–2017	50
Flickr	2,1435	1,887	Jun 1,2010–Aug 22, 2010	10
News	29,247	12,204	April 17, 2014–May 25, 2018	20
ACL	974	9,494	2016–2018	87
SOTU	227	6,570	1790–2016	1,152



(a) NIPS dataset (b) Flickr dataset (c) News dataset

Fig. 6. The temporal densities of the real-world datasets.

— Flickr dataset. We collect the Flickr images within the city of Paris from June 1, 2010 to August 22, 2010. Since each image is annotated by a set of tags denoted by users, we assume those tags associated with an image make up a document. After the pre-processing, we obtain 21,435 documents with 1,887 vocabularies. Each document averagely contains 10 words.
— News [43] dataset. In this dataset, only the news labeled politics is used. There are 29,247 documents with 12,204 unique vocabularies from April 17, 2014 to May 25, 2018 after the standard pre-processing. Each document averagely contains 20 words.
— ACL dataset. It consists of the accepted papers from ACL Anthology in the three consecutive years. After the preprocessing, there are 974 documents with the dictionary size of 9,494. Each document averagely contains 87 words.
— SOTU dataset. It contains the text of annual speech transcripts delivered by the President of United States from 1790 to 2016. After the preprocessing, we obtain 227 documents with 6,570 vocabularies. Each document is averagely composed of 1,152 words.

In addition, the temporal densities of documents from NIPS, Flickr, and News datasets are presented in Figure 6. On the ACL Anthology dataset, there are 265 papers in 2016, 324 papers in 2017, and 550 papers in 2018, respectively, and SOTU dataset is composed of the annual transcripts. The various densities of document arrivals in these datasets as well as the time spans thus form a good testing environment for the proposed rCTM and the other baselines.

We compare the proposed model with the following state-of-the-art algorithms.

— **DTM**, short for the dynamial topic modeling [9], is the seminal dynamic model for topic evolution discovery, where the dynamics of both topic proportions and word distributions are captured via the state space models.

— **DCT**, short for the Dynamic Clustering Topic model [40], is one of the existing models which dynamically learns the topic evolution along the time slices, where both topic popularity and word evolution is captured by Dirichlet chains.
— **rCRP**, a recurrent Chinese Restaurant Process [4], is regarded as one of the benchmark algorithms in modeling dynamics topics where evolving topics are chained by using a recurrent Chinese Restaurant Process.
— **ST-LDA**, short for Streaming LDA [6], originally learns the dynamic topic evolution between consecutive individual documents via a Dirichlet distribution. We extend it to capture the topic dependencies between the consecutive document chunks. In this approach, the topic evolution is chained by the Dirichlet distribution with a balanced scale parameter.
— **DP-density** [23], explores the density of document arrivals to detect the dynamic topics based on the social media data streams, where a Dirichlet process is used to infer the topic number and density estimation technique is exploited to learn the dynamics of topics.
— **MStream**, a model-based text stream clustering algorithm [60], deals with the concept drift problem for the short text streams. To accommodate the topic drift in the long text, we revise its assumption of one-topic proportion to multi-topic proportions for each document.
— **DM-DTM** is short for the Dual Markov Dynamic Topic Model [2] that exploits two Markov chains to capture both topic popularity and topic evolution. In this approach, the topic popularity is captured by the Gamma Markov chain, and topic evolution is modeled by the Dirichlet chain.
— **RNN-RSM** is an abbreviation for Recurrent Neural Network-Replicated Softmax Model [25], where the topic discovery and sequential documents are jointly modeled in the undirected **raplicated softmax (RSM)**[29] and the **recurrent neural network (RNN)** conveys the temporal information for the bias parameters of RSM.

Besides these competitive baselines, the following are our proposed models and their variations.

— **rCTM** is the proposed recurrent Coupling Topic Model, where a new proposal of the multi-topic-thread is induced to describe the topic evolution, and the IBP compound distribution is exploited to infer the topic number as well as sparse topic proportions for each document.
— **rCTM-D** refers to the variant model with the dropout technique. In this approach, the dropout is facilitated over the coupling connections of topics to randomly drop out connections with the given probability, which is used to validate the significance of multi-dependency between evolving topics.
— **rCTM-F** is another variant of rCTM, where the topic number is specified by a fixed number without resorting to the latent IBP compound distribution, and a customized topic proportion of each document is replaced by the fixed common topics at each slice.

In the experiment, we divide a dataset into a sequence of equidistant time slices chronologically, and each document chunk corresponds to a slice. NIPS, Flickr, News, and ACL datasets are divided per three years, per fortnight, per month, and per year, respectively, and SOTU dataset is divided into five slices and each slice spans 45 years. At slice $t = 1$, topics are directly learned from the document chunk $\mathbf{d}_1$ without prior dependency. When time $t > 1$, the evolution of topics depends on their prior states and their coupling relationships. The experimental settings for all models are as follows. (1) In each dataset, the time division is the same for those document chunk-based models including DTM, DCT, rCRP, ST-LDA, DM-DTM, RNN-RSM, the proposed rCTM and its two variants, and document stream-based models DP-density and MStream do not require this setting. (2) Regarding the topic number setting, the nonparametric models including rCTM, rCTM-D, DM-DTM, DP-density, and rCRP are able to automatically learn topic number without such a setting,

while DTM, DCT, ST-LDA, MStream, RNN-RSM, and rCTM-F are specified with the same topic number as those non-parametric models in each dataset. (3) In terms of the document length, the one-topic assumption from DCT, MStream, and DP-density is retained on the NIPS, Flickr, and News datasets, and the assumption is extended to a multi-topic assignment to adapt to the long text of ACL and SOTU dataset. (4) In rCTM, the hyper-parameter is given as $\eta_0 = 0.1$, $\alpha = 0.1$ in the component of topic proportion construction, while the rest are tuned as $\eta = 0.1$, $a_0 = b_0 = 1$, $e_0 = 1$, $d_0 = 10$, $r_0 = 1$ by grid search based on the metric of perplexity during the topic evolution process. We run 1,000 Gibbs samplings to implement the inference process. The parameter settings in other baselines firstly refer to their original papers if available, otherwise we set them at their optimal performance.

4.3 Quantitative Results

Traditionally, *perplexity* [10] is defined to measure the goodness-of-fit of topic modeling by randomly splitting the dataset into training set and testing set, and it is popularly used in the recent topic modeling work [2, 21, 30, 64]. In addition, several new metrics of topic coherence evaluation have been proposed for a comparative review. Among all the competing metrics, the topic coherence [36, 45] matches human judgement most closely, so we adopt it in this work. We also report perplexity, primarily as a way of evaluating the generativeness of different approaches.

4.3.1 Perplexity Over Held-out Set. Following the setting in [25], we randomly hold out p fraction of the dataset ($p \in \{0.6, 0.7, 0.8, 0.9\}$) at each time slice, and train a model with the rest and predict on the sum of held-out sets. A lower perplexity indicates a better generation of the model. For comparing multiple modelings with different assumptions as well as different inference mechanisms, the per-word perplexity on the sum of held-out sets is formally defined as

$$perplexity = exp\left(\frac{-\sum_{t=1}^{T}\sum_{d=1}^{|\mathbf{d}_t|}\sum_{w=1}^{V} log\, P(\sum_{k_t=1}^{K_t}\theta_{dk_t}\phi_{k_t w})}{\sum_{t=1}^{T}\sum_{d=1}^{|\mathbf{d}_t|}\sum_{w=1}^{V} x_{tdw}}\right),$$

where $|\mathbf{d}_t|$ records the number of documents at slice t, x_{tdw} indicates the observed occurrence number of word w in the document d at slice t, while θ_d and ϕ_{k_t} are estimated in the inference procedure.

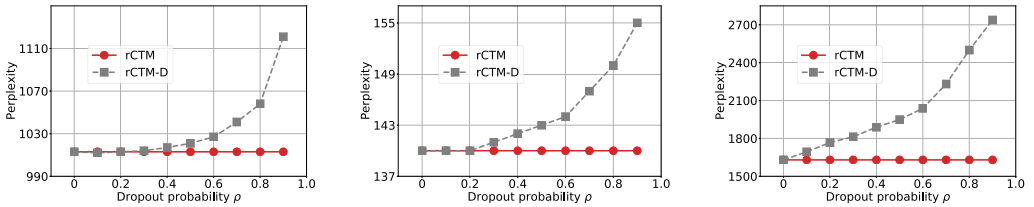
Table 3 reports the *perplexity* performance over varying ratios of held-out sets on the short-text datasets including NIPS, Flickr, and News dataset. By examining the performance result on each dataset, we have the following remarks.

Document chunk-based models. (1) Though the density of document arrivals as well as time span on three datasets are very different, the proposed rCTM and its variant rCTM-F consistently outperform the other baselines with a significant decrease in the perplexity at varying ratios of held-out sets, which confirms the superiority of the proposal of multiple dependencies between evolving topic sequences. Moreover, though rCTM-F is specified with the same topic number learned from rCTM, the perplexity difference between them validates the advantage of the latent IBP compound process in the task of inferring topic number and sparse topic proportion construction. (2) Except for the rCTM and its variant, ST-LDA achieves the best performance on the varying ratios on the NIPS and Flickr datasets, while DM-DTM wins at high ratios on the News dataset, which may be explained that the Gamma Markov Chain in DM-DTM is more fit for the topic weight evolution than others on the News dataset. Among the rest document chunk-based models, the nonparametric model rCRP consistently performs better than DCT and DTM at varying ratios on the three datasets.

Table 3. Perplexity Results of the Increasing Training Data with Varying Ratios $p \in \{0.6, 0.7, 0.8, 0.9\}$ on the NIPS, Flickr, and News Datasets

Models	NIPS				Flickr				News			
	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
DTM	1,658	1,632	1,619	1,597	522	496	487	469	2,880	2,854	2,821	2,785
DCT	1,718	1,669	1,592	1,534	430	412	398	390	2,682	2,609	2,571	2,555
rCRP	1,516	1,457	1,395	1,360	413	398	388	381	2,692	2,588	2,559	2,499
MStream	1,328	1,324	1,303	1,297	419	395	380	375	2,749	2,677	2,637	2,578
DP-density	1,488	1,457	1,450	1,446	347	344	342	340	2,622	2,570	2,522	2,490
ST-LDA	1,207	1,203	1,198	1,195	249	245	240	237	2,864	2,724	2,625	2,549
DM-DTM	1,400	1,364	1,333	1,309	427	382	354	330	2,401	2,365	2,323	2,300
rCTM	**1,057**	**1,045**	**1,015**	**1,013**	**179**	**163**	**152**	**140**	**1,824**	**1,773**	**1,749**	**1,630**
rCTM-F	1,090*	1,075*	1,040*	1,036*	192*	178*	166*	160*	2,099*	2,002*	1,931*	1,859*

The best performance is highlighted in boldface, the second best is emphasized with * and the third best is denoted in underlined.



(a) NIPS dataset (b) Flickr dataset (c) News dataset

Fig. 7. Perplexity comparison between rCTM and its variant rCTM-D with different dropout probabilities $\rho \in [0, 1)$ on the NIPS, Flickr, and News datasets. Both models are measured over the training data with the split ratio $p = 0.9$.

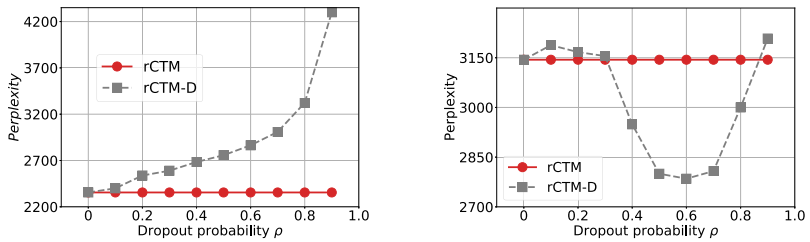
Document stream-based models. It is noted that the performance of MStream and DP-density is different on the three datasets, though both models target at the document streams. DP-density achieves a better result than MStream on the Flickr and News dataset while its performance decreases on the NIPS dataset. That is because DP-density incorporates the arriving density of document streams to determine the dynamics of topics while MStream does not, thus DP-density is more suitable for the social media data with dense arrivals of documents. Such a comparison also implies the proposed generic rCTM is robust to the datasets with different temporal densities.

Dropout-based model. The comparison between rCTM and the variant rCTM-D with varying dropout probabilities on the three datasets is shown in Figure 7. Both proposed models are measured with training data ratio $p = 0.9$. In rCTM-D, the dropout indicator m is drawn from the Bernoulli distribution with ρ. If $m = 0$, the coupling connection is preserved, otherwise it is pruned. On the NIPS dataset, it is observed that perplexity results dramatically increase when the dropout probability $\rho \geq 0.3$. The large dropout probability would drop out the most coupling connections, and topics thus evolve with little dependency from their prior states, leading to the corrupted

Table 4. Perplexity Performance of the Increasing Training Data with Varying Ratios $p \in \{0.6, 0.7, 0.8, 0.9\}$ on the ACL and SOTU Datasets

Models	ACL				SOTU			
	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
DTM	2,599	2,581	2,567	2,555	4,060	4,181	4,312	4,479
DCT	3,340	3,295	3,263	3,175	4,211	3,910	3,774	3,477
rCRP	3,314	3,271	3,233	3,140	4,089	3,826	3,701	3,416
MStream	3,109	3,065	3,025	2,800	3,461	3,400	3,350	3,317
DP-density	3,105	3,049	2,980	2,803	3,480	3,445	3,415	3,389
ST-LDA	2,612	2,549	2,512	2,488	**3,256**	**3,214**	3,181*	3,160*
DM-DTM	2,766	2,704	2,648	2,591	3,502	3,467	3,430	3,399
rCTM	**2,426**	**2,388**	**2,371**	**2,355**	3,336*	3,239*	**3,165**	**3,144**
rCTM-F	2,479*	2,445*	2,416*	2,401*	3,472	3,304	3,206	3,166

The best performance is highlighted in boldface, the second best is emphasized with * and the third best is denoted in underlined.



(a) ACL dataset (b) SOTU dataset

Fig. 8. Perplexity comparison between rCTM and the variant rCTM-D with different dropout probabilities $\rho \in [0, 1)$ on the ACL and SOTU datasets. Both models are measured over the training data with the split ratio $p = 0.9$.

evolving topic sequences. The comparison results imply the significance of multiple couplings between topic chains. On the NIPS dataset, when the dropout probability $\rho \leq 0.3$, it implies most of the multi-coupling connections are maintained, and the perplexity results are stable and nearly approach the optimal performance. On the Flickr dataset, rCTM-D achieves the best performance when the dropout probability $\rho \leq 0.2$, and it is more evident that rCTM-D on the News dataset obtains its best performance only when the dropout probability $\rho = 0$, which means all coupling relationships are preserved. The results from rCTM-D on the three datasets further confirms the proposal of multi-coupling relationships between evolving topics.

Indicated in the Table 4 and Figure 8, the *perplexity* analysis of all competitors on the two long-text datasets including ACL and SOTU datasets, is in the following.

Document chunk-based model. On the ACL dataset, (1) both rCTM and rCTM-F achieve the best performance with an evident decrease in the perplexity at different ratios, followed by the ST-LDA and DTM. Such a comparison once again validates the proposal of multi-topic-thread evolution. With the same topic number setting, the distinct difference between rCTM and rCTM-F

in the perplexity is credited to the latent IBP compound process in the construction of sparsely customized topic proportions for documents. (2) Among the rest document chunk-based models, DM-DTM performs better than rCRP, followed by the performance of DCT.

On the SOTU dataset, (1) the competitor of ST-LDA and the proposed rCTM achieve comparable results at varying ratios, while the former performs better at ratio $p = 0.6$ and $p = 0.7$ and the latter stands out at $p = 0.8$ and $p = 0.9$. However, both strong methods are defeated by the variant of rCTM-D, which earns a much lower perplexity result when the dropout probability $\rho \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$, indicated by Figure 8(b). Specifically, rCTM-D reaches its optimal performance at the dropout probability $\rho = 0.6$. Such results imply that the performance of rCTM improves when a large portion of topic couplings between evolving topics are dropped out. After carefully checking the word distributions of topics as well as their coupling weights, we find this phenomenon is caused by the characteristics of the long-term dataset. In this case, the SOTU dataset is divided into 5 slices, and each slice is allocated with 45 documents spanning 45 years. A portion of topics crossing two slices are actually weakly coupled during the 90-year time, even though some topics seem similar by sharing common frequent words (e.g., power, president, and right). Hence, some of their dependency connections could be dropped out. This phenomenon remains at different time divisions. And not coincidentally, it also occurs in other baselines, whose performance degrades on this dataset. However, rCTM-D survives by dropping out some topic coupling connections between evolving topics on the dataset. (2) Among the rest document chunk-based models, their performance is ranked as DM-DTM > rCRP > DCT > DTM.

Document stream-based model. On the ACL dataset, the difference between DP-density and MStream is slight at varying ratios in terms of a 3-year timespan, while MStream performs better than DP-density on the SOTU dataset. It is because the annual transcripts from the SOTU dataset may not be a good clue to the density estimation in DP-density and its performance is thus compromised.

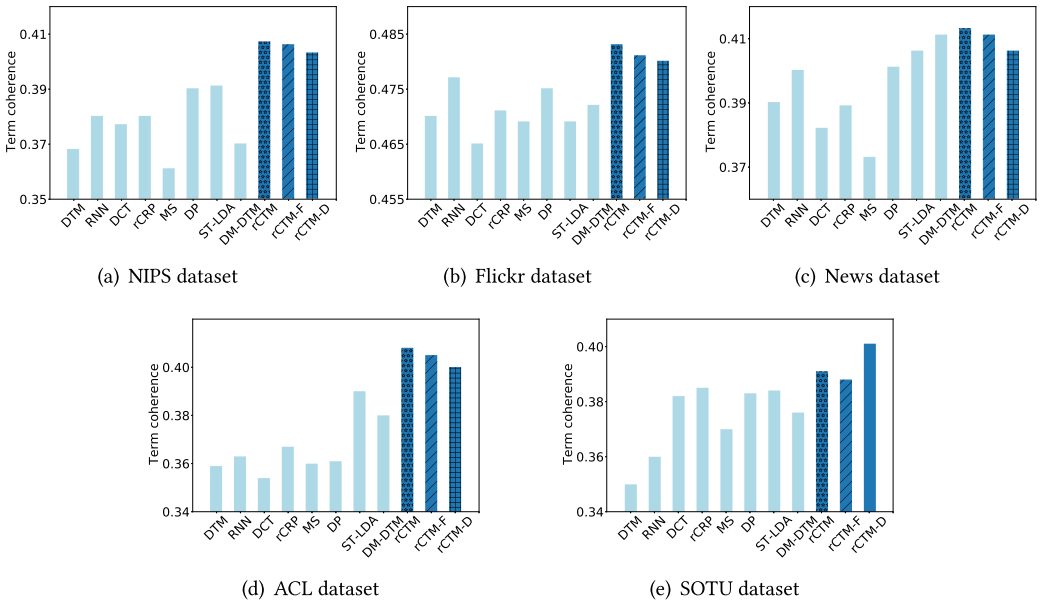
Dropout-based model. Indicated by Figure 8 (a), only when the dropout probability $\rho = 0$, rCTM-D obtains its best performance on the ACL dataset when all coupling relationships are preserved, which confirms the significance of multi-coupling relationships between topic chains. Distinct from the aforementioned datasets, SOTU dataset contains the long-range annual transcripts from 1790 to 2016, which results in the weak connection between topics in consecutive slices. Therefore, rCTM-D reaches the lowest perplexity by dropping out some topics coupling connections between topics.

4.3.2 Topic Coherence. We proceed to evaluate the interpretability of detected topics based on the important measure of topic coherence normalized **Pointwise Mutual Information (*PMI*)** [36], which is formally defined as based on the top-k terms within a topic,

$$C = \frac{2}{k(k-1)\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} nPMI(w_i, w_j)}, \qquad nPMI(w_i, w_j) = \frac{log\frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-log(P(w_i, w_j)+\epsilon)},$$

where $P(w_i, w_j)$ denotes the probability of co-occurrence of w_i and w_j in one document and $P(w_i)$ is the probability of w_i appearing in the document. A higher *PMI* value indicates the terms within the topics are more consistent and interpretable. To obtain an unbiased result, we resort to the large-scale external Wikipedia data [45] to measure the top-10 coherence values for all competitor models. The average topic coherence results based on all topics from each slice are presented in Figure 9, and the analysis of all competitor models is in the following.

On the NIPS and Flickr dataset, the topic coherence results are presented in Figure 9(a) and (b). The dropout probability of rCTM-D is set as $\rho = 0.2$ on both datasets, at which rCTM-D



(a) NIPS dataset (b) Flickr dataset (c) News dataset

(d) ACL dataset (e) SOTU dataset

Fig. 9. Average term coherence results from all competitor models on the five datasets, where the proposed rCTM, rCTM-F, and rCTM-D are emphisized by the rightmost three bars. For the purpose to align with their coherence bars, RNN refers to RNN-RSM, MS corresponds to MStream, and DP is short for DP-density here.

obtains the lowest perplexity. It is observed rCTM and its variants achieve the highest coherence scores among all competitors, and rCTM is superior to its variants with a higher coherence score, indicating more interpretable and coherent topical terms therein. The superiority of rCTM-based models over the other baselines with the single-topic-thread evolution assumption vouches for the significance of multi-thread couplings between evolving topics. In addition, DP-density is noted to retain its advantage over other baselines with a higher coherence score, while ST-LDA degrades and RNN-RSM, MStream, and DM-DTM improve their performance by a large margin on the Flickr dataset.

On the News dataset, the dropout probability in rCTM-D is set as $\rho = 0.1$. The coherence results from all competitors are presented in Figure 9(c). It is noted that rCTM still outperforms others with the highest coherence score, and baselines including RNN-RSM, DP-density, ST-LDA, DM-DTM, and rCTM-F offer the competitive coherence scores. Among them, DM-DTM is tied with rCTM-F for second place and performs better than ST-LDA and rCTM-D. Besides, RNN-RSM consistently outperforms DTM by a large margin, and rCRP also performs better than DCT with a higher coherence score. In contrast, the performance of MStream degrades, implying its disadvantage on a dataset with dense arrivals.

On the ACL and SOTU datasets, the dropout probability of rCTM-D in these two long-text datasets is given as $\rho = 0.1$ and $\rho = 0.6$, respectively. The coherence results are presented in Figure 9(d) and (e). On the ACL dataset, the proposed rCTM is superior to its variants, and ST-LDA is followed among the rest competitors. In constrast, on the SOTU dataset, rCTM-D with $\rho = 0.6$ is the winner with the highest coherence value and rCTM is the runner-up compared with other baselines. Besides the proposed model, the performance of rCRP, DCT, DP-density, and ST-LDA is close in the coherence measure, while the performance of DTM and RNN-RSM decreases in these two long-text documents.

Table 5. Document Time Stamp Prediction Accuracy Results Over the Five Datasets

	DTM	RNN-RSM	DCT	rCRP	ST-LDA	DM-DTM	rCTM	rCTM-F	rCTM-D
NIPS	0.50	0.52	0.45	0.50	0.45	0.53	**0.55**	0.53	0.54
Flickr	0.46	0.48	0.45	0.46	0.53	0.50	**0.54**	0.52	0.54
News	0.53	0.53	0.49	0.50	0.54	0.52	**0.55**	0.54	0.54
ACL	0.60	0.62	0.63	0.64	0.67	0.62	**0.68**	0.67	0.66
SOTU	0.40	0.43	0.45	0.45	0.47	0.44	0.48	0.48	**0.51**

The best performance is highlighted in boldface.

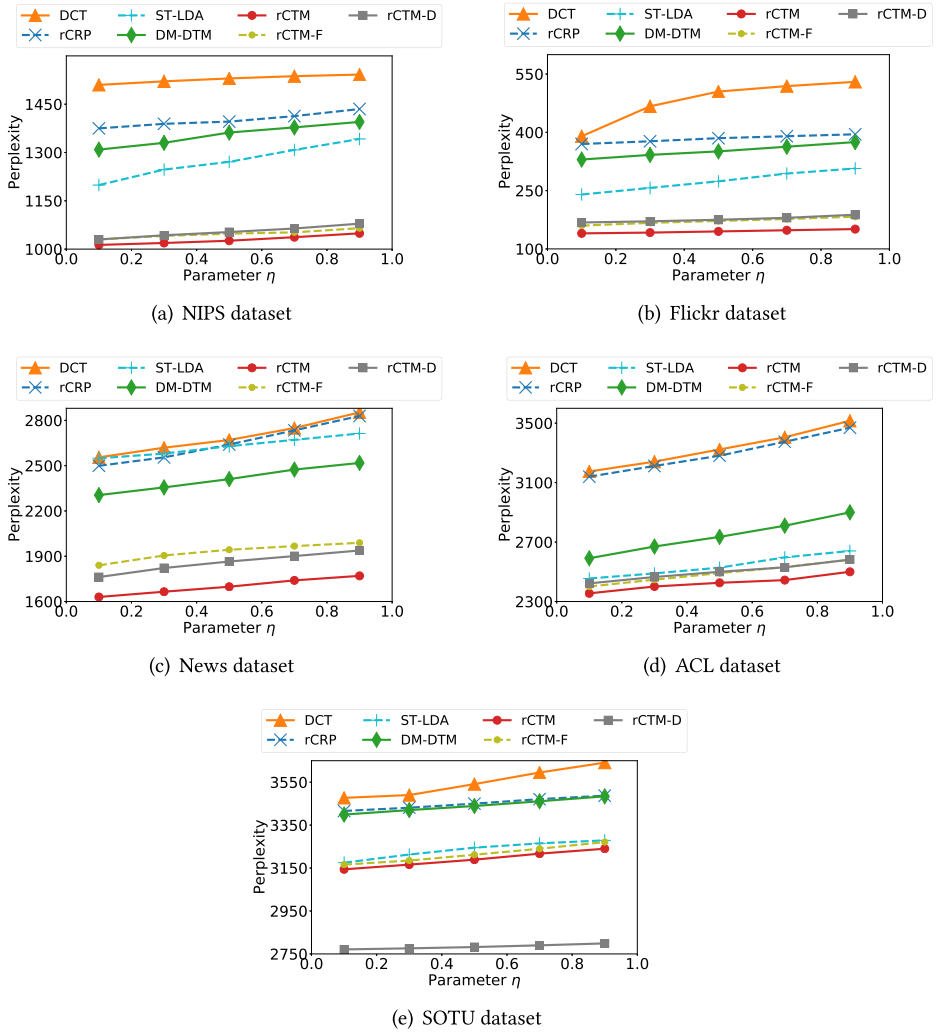
In a nutshell, the proposed rCTM and its variants exhibit superiority to other baselines in terms of the coherence metric on different datasets. Such performance is roughly consistent with the perplexity results, which once again confirms the significance of modeling multiple couplings between evolving topics as well as the sparse customization of topic proportions in rCTM. Besides, DP-density, ST-LDA, DCT, and rCRP are robust on different datasets without a drastic change in the coherence values. However, RNN-RSM and DTM are advantageous on the short-text datasets, while DM-DTM more fits the datasets with densely irregular document arrivals. The performance of MStream is not satisfactory on the NIPS, News, ACL, and SOTU datasets.

4.3.3 Document Time Stamp Prediction. To further evaluate these recurrent modelings, referring to the empirical study [25], we split the sequential documents at each time slice when the ratio $p = 0.9$, and predict the time stamp of a document on the held-out dataset by finding the most likely location based on the topics with maximum likelihood over the timeline. The document stream-based methods are excluded due to the different settings and the results of document time stamp prediction accuracy from the rest competitors are presented in the Table 5.

It is noted that the proposed rCTM as well as its variants rCTM-F and rCTM-D outperform the other baselines with the higher prediction accuracy over five datasets, implying the higher semantic match between the held-out documents and recognized topics along the timeframe. Among the rest of competitors, ST-LDA outperforms the other baselines on the Flickr, News, ACL, and SOTU dataset. DM-DTM and RNN-RSM gain comparable results over the five datasets while enjoying the advantage in the short-text datasets, which is true of DTM. In addition, rCRP obtains a better prediction accuracy than DCT over the five different datasets.

4.3.4 Effects of Varying η. Since topics at slice $t = 1$ are directly learned via $\phi_{k_1} \sim Dir_V(\eta)$ ($k_1 \in \{1, 2, \ldots, K_1\}$) without prior dependency. Then they serve as the input to the recurrent coupled topic sequences, and the posterior topics as well as their coupling relationships at slice $t > 1$ are sequentially learned. Hence, the results of topics at slice $t = 1$ are important for the whole topic evolutionary process, which is also true for other dynamic topic models. To see the effects of varying η on the overall performance, topics at slice $t = 1$ are initialized with varying η in these competitors following the prior work [2, 40], and the overall performance on the five datasets is presented in Figure 10.

The results of the document stream-based approaches as well as DTM and RNN-RSM are excluded due to the different settings. The perplexity performance is measured on the held-out set when $p = 0.9$ on the five datasets, and the dropout probability in rCTM-D is set $\rho = 0.2$ on the NIPS, $\rho = 0.2$ on the Flickr, $\rho = 0.1$ on the News, $\rho = 0.1$ on the ACL, and $\rho = 0.6$ on the SOTU datasets. We observe that, in addition to the lowest perplexity results, the proposed rCTM and its two variants acquire a slower increase than other baselines with η growing, which demonstrates



(a) NIPS dataset

(b) Flickr dataset

(c) News dataset

(d) ACL dataset

(e) SOTU dataset

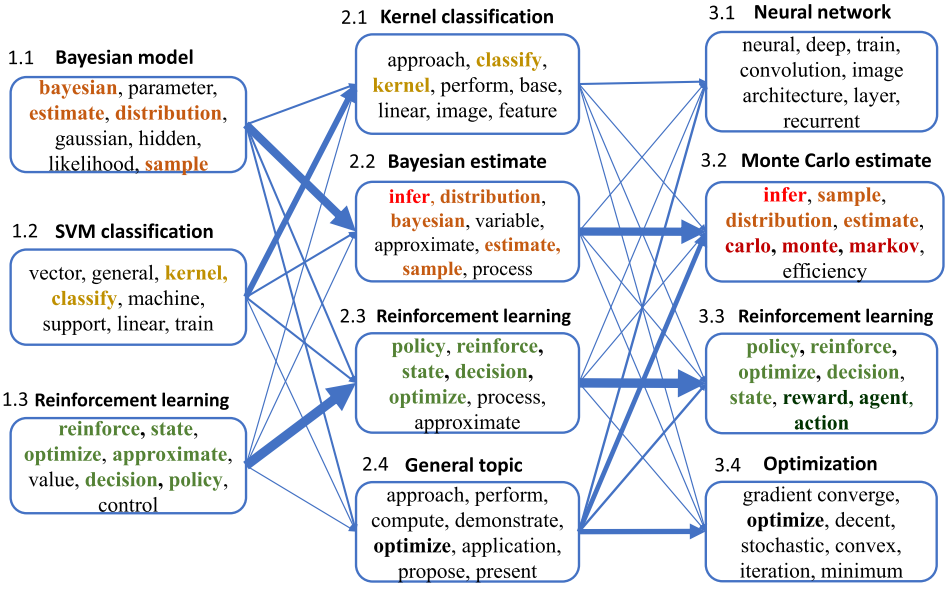
Fig. 10. Performance comparison with varying the parameter $\eta \in (0, 1)$ on the five datasets.

the merit of rCTM and its variants that they are robust and less sensitive to the growing η with the multi-topic-thread evolution assumption. On the other hand, the increasing perplexity from all competitors indicates that varying η affects their performance in the task of evolving topic sequences, and a small value η to initialize topics at slice $t = 1$ is preferred by these document-chunk based topic modelings.

4.4 Qualitative Results

4.4.1 Topic Evolving Sequence with Coupled Dependencies. To have an intuitive understanding of evolving topics as well as their multi-dependency relationships, we present two representative examples to exhibit the evolutionary process.

Figure 11 (a) presents the recurrent topics on the NIPS dataset, which is divided into four equidistant time-slices, and topics in the last three slices are exhibited considering the space limit. Figure 11(b) provides the corresponding weights between consecutive topics, which summarizes the



(a) Recurrent topic sequences.

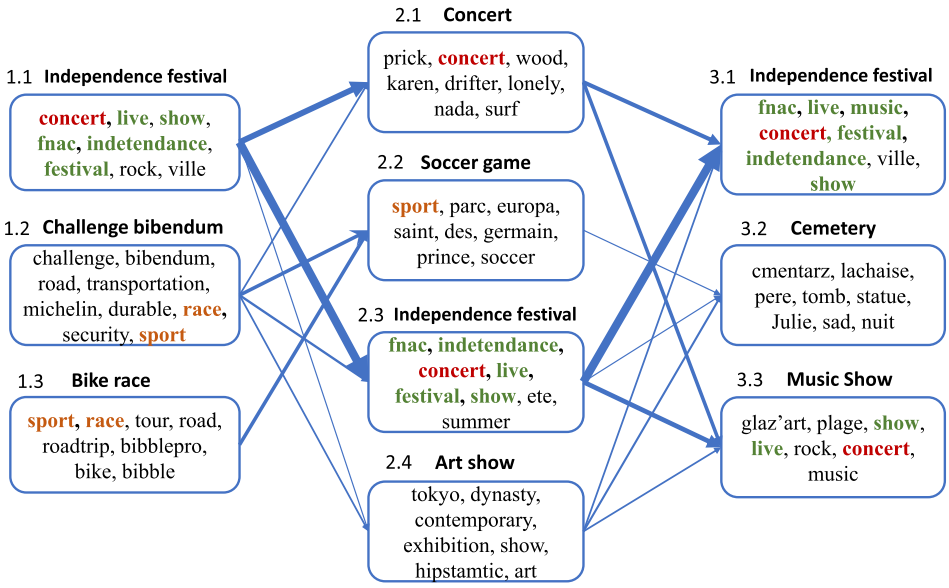
	Topic 2.1	Topic 2.2	Topic 2.3	Topic 2.4
Topic 1.1	6.2	339	8.4	10.7
Topic 1.2	115	9.1	8.9	1.4
Topic 1.3	1.3	4	400	3.1

	Topic 3.1	Topic 3.2	Topic 3.3	Topic 3.4
Topic 2.1	9.3	1.3	4.4	1.5
Topic 2.2	0.4	485	1.1	2.5
Topic 2.3	1.5	1.6	1079	3.8
Topic 2.4	38	127	54	154

(b) Corresponding coupling weights between topics in two consecutive slices.

Fig. 11. The example of topic evolving sequence discovered on the NIPS dataset. In figure (a), topics are annotated by their time-slice and index, each of them is represented by the top-8 words within the rectangular, common words between consecutive topics are highlighted by different colors and the thickness of arrows between consecutive topics indicates their coupling strengths. Figure (b) summarizes the exact coupling weights between consecutive topics.

sharing of latent word counts between them. Our observations are in the following. (1) Topics in each column are semantically meaningful by the most probable words, and similar topics are closely coupled with the highlighted common words across the slices. For example, the topic sequence about Bayesian method evolves from topic 1.1 -> topic 2.2 -> topic 3.2 across three slices with strong coupling weights indicated in Figure 11(b), and the highlighted common words in their contents, such as "bayesian," "distribution," and "estimate," are shared crossing the slices. In addition, the topic sequence about Reinforcement Learning develops from topic 1.3 -> topic 2.3 -> topic 3.3 with strong coupling dependencies across the slices. (2) Besides long-term topic sequences, topic 1.2 about SVM classification in the first column evolves to the subsequent topic 2.1, which weakly connects with the posterior topics in the last slice. In comparison, the general topic of topic 2.4 contributes to all posterior topics with different coupling weights. (3) Coupling weights indicate that new topic 3.1 about Neural Network weakly connects with the priors and no common words are shared, which fits the fact that the topic of the neural network gets its popularity in recent years.



(a) Recurrent topic sequences.

	Topic 2.1	Topic 2.2	Topic 2.3	Topic 2.4
Topic 1.1	2.2	0	5.3	0.1
Topic 1.2	0.2	1.2	0	0.2
Topic 1.3	0	1.1	0	0

	Topic 3.1	Topic 3.2	Topic 3.3
Topic 2.1	0.8	0	0.6
Topic 2.2	0	0.1	0
Topic 2.3	7.9	0.2	1.2
Topic 2.4	0.2	0.3	0.2

(b) Corresponding coupling weights between topics in two consecutive slices.

Fig. 12. The example of topic evolving sequence discovered on the Flickr dataset. In figure (a), topics are annotated by their time-slice and index, each of them is represented by the top-8 words within the rectangular, common words between consecutive topics are highlighted by different colors and the thickness of arrows between topics indicates their coupling strengths. The arrows with coupling weight 0 are not plotted. Figure (b) summarizes the exact coupling weights between consecutive topics.

Distinct from scientific topic sequences on the NIPS dataset, the Flickr dataset records real social activities in the world and its topic sequences together with their coupling weights are presented in Figure 12. We report the recurrent topics in the last month and each slice lasts for 10 days. We observe that (1) the coupling weights between consecutive social activities on the Flickr dataset are small compared with scientific examples on the NIPS dataset, and some coupling weights are 0. That is because each Flickr document contains fewer words and the discovered topics from Flickr are real and different activities. The coupling weights between them are thus small. (2) Relevant topics are coupled while their coupling weights and shared common words are distinguishable, for example, even though the topic sequence about Concert evolves as (topic 1.1) -> (topic 2.1, topic 2.3) -> (topic 3.1, topic 3.3) in multiple threads across three slices. Indicated by the coupling weights in Figure 12(b), topic 2.1 couples the posterior topics with the small weights, and topic 3.3 also weakly couples its prior topics. Though both of them talk about music, they are distinguished from the other strong coupled topic sequences on Independence Festival, which shares more

common frequent words, e.g., "fnac" and "indetendance," highlighted in green color. In addition, the topics about sport (topic 1.2, topic 1.3) -> (topic 2.2) are naturally chained, and the small weights between them indicate each topic records a different sports event, which is reinforced that no more common words are shared between them except for "sport" and "race." (3) Unrelated topics are naturally identified by the small coupling weights. For example, topic 2.2 is about Soccer Game, whose connections with the posterior topics are denoted by the small weights, and it is also true for topic 2.4, which is unrelated to the posterior topics.

Two intuitive examples from the NIPS and Flickr datasets further prove the effectiveness of multi-dependencies associated with prior topics. And the flexible weights learned via the hierarchical Gamma distribution successfully identify the evolutionary closeness between consecutive topics.

5 RELATED WORK

Most of the dynamic topic models are built under the single-topic-thread assumption that the current state of one topic solely depends on its own historical states without referring to other topics. We summarize the related dynamic models in three aspects. The first two modelings are inherited from temporal topic modeling, and the third one is founded on **Poisson factor analysis (PFA)**. Last but not the least, we briefly compare language models with RNNs.

State space modeling. One of the benchmark models learning, the evolution of topics is the state space model, in which the V-dimensional topic ϕ_t at step t evolves via $\phi_{k_t} \sim \mathcal{N}(\phi_{k_{t-1}}, \Sigma)$ [9] or the linear form $\phi_{k_t} \sim \mathcal{N}(\beta\phi_{k_{t-1}}, \Sigma)$ [47]. The seminal work **dynamic topic model (DTM)** [9] captures the evolution of topics by the state space models over the sequence of discrete time-slices, where Kalman filter [34] infers temporal update of the state space parameters. Comparing with the classic DTM, our model significantly differs in three aspects. Firstly, instead of fixed topic number setting in DTM, our model is capable to automatically learn the topic number at each slice as well as the sparse topic proportions of each document and thus accommodate new topics along the time. Second, the topic evolves in a single thread in DTM and it fails to identify the influences from other related topics. However, our model breaks such a limitation and supposes topic evolves in multiple threads with corresponding dependencies on the previous. Furthermore, our model induces a tractable and efficient inference method with data augmentation techniques, and such an inference problem cannot be solved by DTM. The later continuous time DTM [52] replaces the discrete state space model and detects the evolution of topics over continuous documents using Brownian motion, where variational Kalman filtering is exploited to infer the parameter in the continuous time setting. To relieve the manual setting of the topic number, the work [3–5] facilitates the nonparametric prior of a Dirichlet process to automatically derive the topic number for the sequential documents. Among them, topics (storylines) in [3, 4] are chained via a **recurrent Chinese Restaurant Process (rCRP)**, which allows topics to evolve with genesis and death. While the base measure of topics in [5] is tied via the rCRP, documents are generated from an epoch-specific HDP. In these scenarios, the number of topics or themes are flexibly learned rather than predefined and their topic transitions are chained by Gaussian state space models. Successful as state space modelings are, one of the main deficiencies is that these models suffer from a heavy computational cost due to the non-conjugate problem, and their scalability would be prohibitive in the high dimensional data. To this end, a line of research work is thus developed to mitigate the problem. The work [41] employs the Pólya-Gamma augmentation trick to provide a conditionally conjugate scheme for Gaussian priors. To mitigate the scalability limit from the state space modeling, the work [33] presents a generalized class of tractable priors and scalable approximate inference to explore both long-term and short-term

evolving topics, while the work [7] proposes a parallelizable inference using Gibbs sampling with Stochastic Gradient Langevin Dynamics to scale up the dynamic topic modeling in both single and distributed environments. Besides the scalability, the work [12, 32, 44] focuses on the evolving topics with various time-scales of the resolution, which allows topics to evolve in the different scales.

Even though significant progress has been made in state space-based models in the task of topic evolution, such models restrict the evolving topics under the single-topic-thread assumption and fail to capture the potential multiple dependencies between evolving topics. Without thoroughly encoding the complex temporal relationships between time-evolving topics, the learned evolution of topics might be defective.

Dynamic modeling with the Dirichlet chain. Extensive studies exploit the Dirichlet distribution to chain the dynamics of topics over the sequence of discrete slices, in which the tractable inference of sampling topics becomes the advantage over the state space models due to benefit of the Dirichlet distribution. The topic tracking model [31] and DCT [40] harness a Dirichlet distribution to chain the consecutive evolving topics over the text stream, where the evolution of topic popularity and word distributions depend on their prior states via two Dirichlet Markov chains. In comparison, the DM-DTM [2] employs two different Markov chains to detect the topic evolution in the count data, where the topic popularity is modeled by the Gamma Markov chain, and the evolution of topics is also captured by the Dirichlet distribution. The work [39] turns the problem of user interest drifting to be the evolution of topics, where the dynamics of topics over time is also captured via a Dirichlet chain. Despite the wide application of the Dirichlet chain, little attention is paid to the inference of evolutionary weight between consecutive topics, which is still intractable and incurs heavy computation. Furthermore, approaches with a Dirichlet chain ignore multi-dependency relationships between time-evolving topics. In contrast, though the proposed rCTM also exploits a Dirichlet distribution to chain evolving topics, it breaks the limitation of the single-topic-thread evolution and proposes a new framework where the current topic evolves from all prior topics with the corresponding coupling weights. To avoid the confusion with **correlated topic modeling (CTM)** [8, 28], we clarify the major difference in two aspects. First, the correlation between topics in CTM indicates the existence of correlation in their proportions via the logistic normal distribution. In comparison, the couplings between evolving topics are defined as the coupling closeness between their word distributions via the hierarchical Gamma distributions. Second, our proposed model aims at encoding the complex temporal correlations between evolving topics in the dynamic context while CTM is limited to a static text dataset.

Besides the above dynamic modelings in the context of discrete slices, a line of studies captures the dynamics of topics from the continuous document streams. The work [18] combines Dirichlet and Hawkes processes to capture the dynamics of topics from sequential documents, in which the Hawkes process learns temporal density of topics with multiple predefined Gaussian kernels. The work in [23] further mitigates the restriction from the predefined kernels and exploits the density estimation technique to incrementally learn the dynamics of topics with the sliding window. In addition, the Temporal LDA [55] aims at predicting the transition of topic weight in the future documents while ignoring the transition of its word distribution. In comparison, the work in [6] puts forward a Bayesian model named streamLDA to learn the transition of topic weight as well as its word transition between consecutive documents. In addition, a large number of research studies focus on continuous streaming short texts from social media to reveal the topic drift. The work in [60, 61] makes an effort to incrementally cluster the short text streams from social media and uncover the dynamic clusters (topics) by assigning one topic to each short text. A joint model

in [57] handles the Chinese streaming short text by integrating the prior of rCRP and biterm topic model [58] to detect the dynamic topics. Given the meta features of social media data, the work in [62, 63] incrementally groups the continuous tweets into different varying topic sets according to the combination of textual contents, spatial and temporal features.

Dynamic Poisson factor analysis. Targeting at the count data, it is a matrix factorization method for the discrete sequential count data under the PFA [1]. Though some applications of PFA are not the focus of this article, for the sake of completeness, we discuss some representative work to introduce how the latent variables evolve over the count data. The work [47] proposes a **Poisson-Gamma dynamic system (PGDS)** for sequentially count data, where the latent states of topic proportions are chained via the Gamma shape parameter. Its later deep variant, [22], extends the Poisson-Gamma dynamic system by constructing a hierarchical latent structure for the topic proportions, which allows both first-order and long-range temporal dependencies. We credit the data augmentation technique in the proposed model to these approaches. The recent work in [46] closely relates with PGDS and presents the Poisson-randomized Gamma dynamic system for the sequential biased data with sparsity or burstiness. In addition, the work in [16] models the evolution of latent factors in terms of user preferences and item features in the context of recommender system via the Gamma scale parameters. Some studies based on the dynamic relational data [38, 59] learn the evolution of node membership by leveraging data augmentation technique under the framework of PFA.

Comparison with the RNN-based language models. In addition to the Bayesian approaches, a line of studies [15, 21, 35, 54] integrates topic models and language models and inherits merits from both sides. Among them, the work in [15] develops the model TopicRNN, where the global semantics is captured by the topic modeling while the local dependency between words within a sentence is detected by a RNN. The work in [35] integrates two components to jointly learn topics and word sequence, where a word sequence is predicted via the RNN. The work in [54] simultaneously learns the global semantics of a document via a neural topic model and uses the learned topics to build a mixture-of-experts language modeling based on RNN. The model of RNN-RSM in [25] also aims at recurrent topic discovery, and it leverages the **Restricted Boltzmann Machines (RBMs)** to define the interaction between topics and words and the RNN is used to convey the temporal information and update the bias parameters of RBMs. In this solution, consecutive topics are not directly connected, which is a marked contrast to the stochastic multi-topic-thread assumption between topics in our model. Additionally, it adopts the contrastive divergence algorithm to estimate the parameters, which also differs from the Gibbs sampling in our Bayesian network. The most recent paper [21] uses a recurrent deep topic model to guide a stacked RNN for language modeling, and thus the words from a document are jointly predicted by the learned topics via the topic modeling and its preceding words via the RNN. It is noted that most of the RNN-based language models are typically applied at the word level and learn the local temporal dependency between the words. Such a task is quite different from ours. First, the topics (word distributions) capture the global semantics of the corpus by word occurrences across the documents. Such long-range dependency and global semantics may not be captured well by the RNN-based language models [15, 21, 35, 54]. In addition, our proposed work aims at encoding the temporal dependency between two sets of latent topics across the time steps, which is distinct from the syntactic dependency between words in the RNN-based models. Though the task of encoding dependency is quite different in dynamic topic modelings and RNN-based language models, they could work together to cooperatively capture both global semantics and local dependency for language generation.

6 CONCLUSION AND FUTURE WORK

We introduce a novel nonparametric Bayesian model, a rCTM over sequentially observed documents. The multi-fold contributions are summarized in the following. (1) This model breaks the limitation of single-topic-thread evolution from most of the existing work and induces a new and flexible proposal of the multi-topic-thread evolution. Accordingly, the current topics evolve from all prior topics with the corresponding topic coupling weights. Such a flexible proposal naturally adapts to the sequential documents with complex relationships. (2) To tackle the unexplored and intractable inference challenge, we present a novel solution with data augmentation and marginalization techniques to decompose the joint multi-dependencies between topics into separated relationships. A novel Gibbs sampler with a backward–forward filter algorithm is exploited to efficiently infer the fully conjugate model in a closed form. (3) Without tuning the topic number in sequential documents, we leverage the latent IBP compound distribution to automatically infer the overall topic number and customize the sparse topic proportions for each document, where both short text and long documents are flexibly adapted. To further validate the significance of topic couplings, we borrow the dropout technique from deep learning and incorporate it into the proposed rCTM as a counterpart. Evaluation on both synthetic and real-world datasets demonstrates that rCTM infers a highly interpretable dynamic structure, and the multi-coupling relationships learned between time-evolving topics are significant to infer the topical structure in future. Further, the experimental results also indicate rCTM is superior over the competitive baselines in terms of low per-word perplexity, high topic coherence and high time prediction accuracy.

Although the analytic posterior of rCTM results in an efficient Gibbs sampling, rCTM is limited by two main disadvantages: (1) Gibbs sampling is a time-consuming batch method when inferring high-dimensional latent parameters compared with the gradient-based optimization methods in the neural networks. (2) It is not easy to plug the valuable side information into the Bayesian network with a predefined structure, e.g., document labels or promising word embeddings [14, 24, 42], otherwise, the structure of Bayesian network has to been reformulated. Therefore, one future attempt is to marry rCTM with neural networks and incorporate the variational autoencoder [48] into the proposed model, where the pretrained word embeddings are possibly incorporated. Another promising direction is to extend the evolving topics into the hierarchically recurrent coupled topics, where not only the coupled topic evolution but also the hierarchical topics from general to specific could be captured.

REFERENCES

[1] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2015. Nonparametric bayesian factor analysis for dynamic count matrices. In *AISTATS*. 1–9.
[2] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2018. A dual Markov chain topic model for dynamic environments. In *ACM SIGKDD*. 1099–1108.
[3] Amr Ahmed, Qirong Ho, Choon Hui Teo, Jacob Eisenstein, Alex Smola, and Eric Xing. 2011. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *AISTATS*. 101–109.
[4] Amr Ahmed and Eric Xing. 2008. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *SDM*. SIAM, 219–230.
[5] Amr Ahmed and Eric P. Xing. 2010. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *UAI*. 20–29.
[6] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. 2016. Streaming-LDA: A copula-based approach to modeling topic dependencies in document streams. In *ACM SIGKDD*. 695–704.
[7] Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. 2016. Scaling up dynamic topic models. In *WWW*. 381–390.
[8] David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *NeurIPS*. 147–154.
[9] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *ICML*. 113–120.
[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR* 3, Jan (2003), 993–1022.
[11] Longbing Cao. 2015. Coupling learning of complex interactions. *Inf. Process. Manage.* 51, 2 (2015), 167–186.

[12] Xilun Chen, K. Selçuk Candan, and Maria Luisa Sapino. 2018. IMS-DTM: Incremental multi-scale dynamic topic models. In *AAAI*.
[13] Xin Cheng, Duoqian Miao, Can Wang, and Longbing Cao. 2013. Coupled term-term relation analysis for document clustering. In *IJCNN*. 1–8.
[14] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *ACL*. 795–804.
[15] Adji B. Dieng, Wang Chong, Jianfeng Gao, and John Paisley. 2016. TopicRNN: A recurrent neural network with long-range semantic dependency. In *ICLR*.
[16] Trong Dinh Thac Do and Longbing Cao. 2018. Gamma-Poisson dynamic matrix factorization embedded with metadata influence. In *NeurIPS*. 5824–5835.
[17] Finale Doshi-Velez, Byron C. Wallace, and Ryan P. Adams. 2015. Graph-sparse LDA: A topic model with structured sparsity. In *AAAI*. 2575–2581.
[18] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J. Smola, and Le Song. 2015. Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. In *ACM SIGKDD*. 219–228.
[19] Thomas L. Griffiths and Zoubin Ghahramani. 2011. The Indian buffet process: An introduction and review. *JMLR* 12, 32 (2011), 1185–1224.
[20] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS* 101, Suppl 1 (2004), 5228–5235.
[21] Dandan Guo, Bo Chen, Ruiying Lu, and Mingyuan Zhou. 2020. Recurrent hierarchical topic-guided RNN for language generation. In *ICML*.
[22] Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. 2018. Deep poisson gamma dynamical systems. In *NeurIPS*. 8442–8452.
[23] Jinjin Guo and Zhiguo Gong. 2017. A density-based nonparametric model for online event discovery from the social media data. In *IJCAI*. 1732–1738.
[24] Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. 2019. Document informed neural autoregressive topic models with distributional prior. In *AAAI*, Vol. 33. 6505–6512.
[25] Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Bernt Andrassy. 2018. Deep temporal-recurrent-replicated-softmax for topical trends over time. *NAACL-HLT*.
[26] Ben Hamner. [n.d.]. NIPS Papers.
[27] Shufeng Hao, Chongyang Shi, Zhendong Niu, and Longbing Cao. 2018. Concept coupling learning for improving concept lattice-based document retrieval. *Eng. Appl. of AI* 69, 1 (2018), 65–75.
[28] Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P. Xing. 2017. Efficient correlated topic modeling with topic embedding. In *ACM SIGKDD*. 225–233.
[29] Geoffrey E. Hinton and Russ R. Salakhutdinov. 2009. Replicated softmax: An undirected topic model. In *NIPS*. 1607–1614.
[30] Ting Hua, Chang-Tien Lu, Jaegul Choo, and Chandan K Reddy. 2020. Probabilistic topic modeling for comparative analysis of document collections. *ACM TKDD* 14, 2 (2020), 1–27.
[31] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. 2009. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI*.
[32] Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. 2010. Online multiscale dynamic topic models. In *ACM SIGKDD*. 663–672.
[33] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. 2018. Scalable generalized dynamic topic models. In *AISTATS*, Vol. 84. 1427–1435.
[34] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *J. Basic Eng.* 82, Series D (1960), 35–45.
[35] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *ACL*. 355–365.
[36] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *ACL*. 530–539.
[37] Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM TOIS* 36, 2 (2017), 1–30.
[38] Yaqiong Li, Xuhui Fan, Ling Chen, Bin Li, and Scott A Sisson. 2020. Recurrent Dirichlet belief networks for interpretable dynamic relational data modelling. In *IJCAI*.
[39] Shangsong Liang. 2019. Collaborative, dynamic and diversified user profiling. In *AAAI*, Vol. 33. 4269–4276.
[40] Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. 2016. Dynamic clustering of streaming short documents. In *ACM SIGKDD*. 995–1004.
[41] Scott Linderman, Matthew J. Johnson, and Ryan P. Adams. 2015. Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation. In *NeurIPS*. 3456–3464.

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
[43] Rishabh Misra. 2018. News Category Dataset. DOI: https://doi.org/10.13140/RG.2.2.20331.18729
[44] Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. 2007. Multiscale topic tomography. In *ACM SIGKDD*. 520–529.
[45] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *ACM WSDM*. 399–408.
[46] Aaron Schein, Scott Linderman, Mingyuan Zhou, David Blei, and Hanna Wallach. 2019. Poisson-randomized gamma dynamical systems. In *NeurIPS*. 781–792.
[47] Aaron Schein, Hanna Wallach, and Mingyuan Zhou. 2016. Poisson-gamma dynamical systems. In *NeurIPS*. 5005–5013.
[48] Akash Srivastava and Charles A. Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.
[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR* 15, 1 (2014), 1929–1958.
[50] Charlie Tang and Russ R. Salakhutdinov. 2013. Learning stochastic feedforward neural networks. In *NeurIPS*. 530–538.
[51] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet processes. *JASA* 101, 476 (2004), 1566–1581.
[52] Chong Wang, David Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *UAI*. 579–586.
[53] Can Wang, Zhong She, and Longbing Cao. 2013. Coupled attribute analysis on numerical data. In *IJCAI*. 1736–1742.
[54] Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2018. Topic compositional neural language model. In *AISTATS*. 356–365.
[55] Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: Efficient online modeling of latent topic transitions in social media. In *ACM SIGKDD*. 123–131.
[56] Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. 2010. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*. 1151–1158.
[57] Yunfeng Xu, Hua Xu, Longxia Zhu, Hanyong Hao, Junhui Deng, Xiaomin Sun, and Xiaoli Bai. 2018. Topic discovery for streaming short texts with CTM. In *IJCNN*. IEEE, 1–7.
[58] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *WWW*. 1445–1456.
[59] Sikun Yang and Heinz Koeppl. 2018. Dependent relational gamma process models for longitudinal networks. In *ICML*. 5551–5560.
[60] Jianhua Yin, Daren Chao, Zhongkun Liu, Wei Zhang, Xiaohui Yu, and Jianyong Wang. 2018. Model-based clustering of short text streams. In *ACM SIGKDD*. 2634–2642.
[61] Jianhua Yin and Jianyong Wang. 2016. A text clustering algorithm using an online clustering scheme for initialization. In *ACM SIGKDD*. 1995–2004.
[62] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. 2017. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *ACM SIGKDD*. 595–604.
[63] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han. 2016. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *ACM SIGIR*. 513–522.
[64] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. 2018. Dirichlet belief networks for topic structure learning. In *NeurIPS*. 7955–7966.
[65] Mingyuan Zhou and Lawrence Carin. 2013. Negative binomial process count and mixture modeling. *PAMI* 37, 2 (2013), 307–320.

Received September 2020; revised January 2021; accepted February 2021