Topic Modeling for Sequences of Temporal Activities

Zhi-Yong Shen, Ping Luo, Yuhong Xiong, Jun Sun, Yi-Dong Shen

HP Laboratories
HPL-2010-160

Keyword(s):

topic modeling, LDA, sequence, temporal activities

Abstract:

Temporally-ordered activity sequences are popular in many real-world domains. This paper presents an LDA-style topic model for sequences of temporal activities that captures three features of such sequences: 1) the counts of unique activities, 2) the Markov transition dependence and 3) the absolute or relative timestamp on each activity. In modeling the first two features we propose the concept of global transition probability and distinguish it with local transition probability used in previous work. In modeling the third feature, we employ a continuous time distribution to depict the time range of latent topics. The combination of the global transition probability and the temporal information helps to refine the mixture distribution over topics for temporal sequence analysis. We present results on the data of distributed denial-of-service attack and system call traces, qualitatively and quantitatively showing improved topics, better next activity prediction and sequence clustering.

External Posting Date: October 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: October 21, 2010 [Fulltext]
Published in the Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, December 6-9, 2009

Topic Modeling for Sequences of Temporal Activities

Zhi-Yong Shen[1,2], Ping Luo[3], Yuhong Xiong[3], Jun Sun[1,2], Yi-Dong Shen[1]
[1]Institute of Software, CAS
[2] Graduate University of Chinese Academy of Sciences
[3] Hewlett Packard Labs China

Abstract

Temporally-ordered activity sequences are popular in many real-world domains. This paper presents an LDA-style topic model for sequences of temporal activities that captures three features of such sequences: 1) the counts of unique activities, 2) the Markov transition dependence and 3) the absolute or relative timestamp on each activity. In modeling the first two features we propose the concept of global transition probability *and distinguish it with* local transition probability *used in previous work. In modeling the third feature, we employ a continuous time distribution to depict the time range of latent topics. The combination of the global transition probability and the temporal information helps to refine the mixture distribution over topics for temporal sequence analysis. We present results on the data of distributed denial-of-service attack and system call traces, qualitatively and quantitatively showing improved topics, better next activity prediction and sequence clustering.*

I. Introduction

Temporally-ordered activity sequences are popular in many real-world domains. For example, web access traces of users are recorded as logs in the servers; multi-step network intrusions may trigger off alert sequences by the Intrusion Detection Systems. These activity sequences are represented as a sequence of symbols. Additionally, the numerical data streams, such as the data from human behavior sensors [7] and the temperature data from the chillers in the data center [10], can also be converted into the symbolic representations. Profiling such data may help to identify interesting latent patterns or malicious behaviors among them.

Such temporal data can be formulated as follows. A *temporal activity* is a pair $\langle w, t\rangle$ where $w \in \mathcal{V}$ is an activity symbol while t is the timestamp of w, and $\mathcal{V}$ is the finite set of all activity symbols. A *sequence of temporal activities* is then an ordered set of activities occurred within some time range, denoted by $s = (\langle w_1, t_1\rangle, \cdots, \langle w_N, t_N\rangle)$, where $w_n \in \mathcal{V}$, and $t_n \leq t_{n+1}$ for $n = 1, \cdots, N-1$. Due to the wide applications of temporal activity sequences we focus on the topic modeling method for such kind of data in this study. Specifically, given a set of sequences of temporal activities, we aim to find the latent topics within these sequences by probabilistic topic models [3], [11]. Such models try to assign a latent topic to each activity and then achieve more compact representation of the sequences.

The latent topics within a sequence of temporal activity sequence may be expressed in the following three features: 1) the counts of unique symbols (global information), 2) the Markov transition dependence (local information) and 3) the time-stamps attached to the symbols (temporal information). To capture the characteristics in these features several probabilistic topic models have been applied or proposed for profiling activity sequences. First, Huynh et al. [7] successfully employ Latent Dirichlet Allocation (LDA) [3] to discovery patterns in the sequences of daily human activities, in which LDA only models the counts of unique words (Feature 1) in a sequence and ignores the activity order. Next, to model the Markov dependence (Feature 2) directly Simplicial Mixture of Markov Chain (SMMC) [4] and the Topic Model proposed in [13], expressed by the conditional probability of a symbol given its very previous state, are proposed and applied into the applications in web page browsing, word processor and telephone usage data to show their ability of sequential activity profiling especially on predicting incoming activities. Finally, Wang and McCallum proposed Topic Over Time (TOT) [15] to capture the temporal information (Feature 3) by associating a continuous time distribution with each topic.

Since none of the above models can capture all these three features in temporal activity sequences, in this paper we aim at a uniform topic model which might capture more

meaningful latent semantics in the data by considering all these three features. To this end, we first propose a new concept of *global transition probability* based on transitions of symbols (bigrams), and claim that it covers the information from the first two features. Then, we inject the temporal information on each activity into the generating process of the sequences. This way the proposed model of T-BiLDA is influenced by all the three features of temporal activity sequences. Finally, we present experimental results with two real-world data sets. On the data of *Distributed Denial-of-Service* (DDoS) attack, we show that T-BiLDA exactly re-constructs the attack scenario, in which each attack phase exactly corresponds to one of the latent topics, and the roles of each computers, such as victim, compromised computers and so on, are also exactly identified. On the data of *system call traces*, we show its micro descriptive ability via next token prediction and its macro descriptive ability via sequence clustering. The qualitative results on the first data set and the quantitative results on the second one both show that T-BiLDA significantly outperforms the other five baseline methods considered in modeling sequences of temporal activities.

II. Topics over Transitions

In this section, we first give the formalized definitions of *global activity probability* together with *local and global transition probability* and propose the model of BiLDA for modeling the transitions in the sequences. Then we discuss the properties of these probabilities.

Definition 1 (Global Activity Probability): The global activity probability of the activity with symbol $w \in \mathcal{V}$ is defined as

$$p(w) = \frac{\sharp(w)}{\sum_{w^*} \sharp(w^*)}.$$

Definition 2 (Local Transition Probability): The local transition probability from w to $w'$is defined as

$$p(w'|w) = \frac{\sharp(w \to w')}{\sum_{w^*} \sharp(w \to w^*)}.$$

Definition 3 (Global Transition Probability): The global transition probability from w to w' is defined as

$$p(w \to w^{'}) = \frac{\sharp(w \to w^{'})}{\sum_{w^*,w^{*'}} \sharp(w^* \to w^{*'})}.$$

In the above three definitions $\sharp(\cdot)$ is the counting number of the corresponding symbol or transition occurring. This counting can be conducted within different context. For example, if the counting is in a sequence s, we obtain these probabilities specific to the sequence, denoted as $p(\cdot|s)$ where the '$\cdot$' could be w, $w'|w$ or $w \to w'$. Analogously, we can also compute $p(\cdot|z)$ for each topic z when all symbols or transitions are already assigned to the topics.

TABLE I. Notations used in this paper

Symbol	Description
$\mathcal{V}$	the vocabulary of symbols, with size l
k	number of topics
N_s	number of symbols in the sequence s
w	a symbol in vocabulary
$w \to w'$	a symbol transition where $w, w' \in \mathcal{V}$
w_{sn}	the n-th observed symbol in sequence s
t_{sn}	the timestamp of both w_{sn} and $w_{sn} \to w_{s,n+1}$
z_{sn}	the topic of both w_{sn} and $w_{sn} \to w_{s,n+1}$
θ_s	the multinomial distribution over topics
β_{zw}	the multinomial distribution over w for topic z
$\beta_{z,w'\|w}$	the transition distribution from w for topic z
$\beta_{z,w\to w'}$	the multinomial distribution over $w \to w'$ for topic z
ψ_z	the temporal distribution of time for topic z

The topic modeling is essentially to build a probabilistic mixture model of over $p(\cdot|s)$ via learning the mixture component $p(\cdot|z)$ as well as estimating mixture proportion $p(z|s)$. Various contents in the brackets of $p(\cdot)$ lead to various probabilistic topic models, which are introduced in details as follows.

LDA [3] is a Bayesian network that generates documents using a mixture of topics. LDA ignores the order of the words in a sequence. In its generative process, for each document s, a multinomial distribution $\theta_{sz} = p(z|s)$ over topics is randomly sampled from a Dirichlet with parameter α, and then to generate each word, a topic z is chosen from this distribution, and a word, w, is generated by randomly sampling from a topic-specific multinomial distribution β_{zw}. This model is aimed to learn the distribution over topic θ_s for each sequence s and β_{zw} for each topic z. The marginal distribution of document s in LDA can be represented as:

$$p(\{w_n\}_{n=1}^N|\alpha,\beta,s) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z=1}^{k} \underbrace{p(z|\theta,s)}_{p(z|s)} \underbrace{p(w_n|\beta,z)}_{p(w|z)} \mathrm{d}\theta \quad (1)$$

where the mixture component $p(w|z)$ for each topic and the mixture proportion $p(z|s)$ for each document are determined based on the observed global activity probability $p(w|s)$.

SMMC [4] is proposed to generate the sequence with m-th order Markov dependence. The generative process of the sequences by this model with first-order Markov dependence can be described as follows:

1) set multinomial distribution $\beta_{z,w'|w}$ for each topic z and each $w \in \mathcal{V}$;
2) for each sequence s, draw a multinomial distribution θ_{sz} from a Dirichlet prior α; then for the n-th word w_{sn} in the sequence s:
   a) draw z_{sn} from θ_{sz}; and
   b) draw w_{sn} from $\beta_{z_{sn},w'|w_{s,n-1}}$.

The model of SMMC is aimed to estimate the distribution over topics $\theta_{sz} = p(z|s)$ for each sequence s

(a) Number of transitions

	$\sharp(a \to a)$	$\sharp(a \to b)$	$\sharp(a \to c)$	$\sharp(b \to a)$	$\sharp(b \to b)$	$\sharp(b \to c)$	$\sharp(c \to a)$	$\sharp(c \to b)$	$\sharp(c \to c)$
s_1	92	1	1	1	1	1	1	1	1
s_2	992	1	1	1	1	1	1	1	1

(b) The local transition probability

	$p(a\|a)$	$p(b\|a)$	$p(c\|a)$	$p(a\|b)$	$p(b\|b)$	$p(c\|b)$	$p(a\|c)$	$p(b\|c)$	$p(c\|c)$
s_1	$\frac{92}{94}$	$\frac{1}{94}$	$\frac{1}{94}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
s_2	$\frac{992}{994}$	$\frac{1}{994}$	$\frac{1}{994}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(c) The global transition probability

	$p(a \to a)$	$p(a \to b)$	$p(a \to c)$	$p(b \to a)$	$p(b \to b)$	$p(b \to c)$	$p(c \to a)$	$p(c \to b)$	$p(c \to c)$
s_1	0.92	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
s_2	0.992	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

TABLE II. The statistical numbers in the two example sequences

and $\beta_{z,w'|w} = p(w'|w,z)$ for each topic z and each word w. Analogously, the marginal distribution of Markov transitions in a sequence s in SMMC is:

$$p(\{w_{n+1}|w_n\}_{n=1}^{N-1}, \alpha, \beta, s) = \int p(\theta|\alpha) \prod_{n=1}^{N-1} \sum_{z=1}^{k} \underbrace{p(z|\theta, s)}_{p(z|s)} \underbrace{p(w_{n+1}|w_n, \beta, z)}_{p(w'|w,z)} \mathrm{d}\theta \tag{2}$$

where the mixture component $p(w'|w,z)$ and the mixture proportion $p(z|s)$ are determined based on the observed local transition probability $p(w'|w,s)$.

For the analysis of sequences with Markov dependence, exploiting the local transition probability $p(w'|w,s)$ by SMMC might be better than considering the $p(w|s)$ by LDA. However, considering only $p(w'|w,s)$ is not enough for this task. See the following two sequences with vocabulary $\mathcal{V} = \{a, b, c\}$:

$$s_1 = \underbrace{a, \cdots, a}_{93}, c, c, c, c, b, a, c, a;$$
$$s_2 = \underbrace{a, \cdots, a}_{993}, b, b, c, c, b, a, c, a.$$

As shown by Table 2(b) the local transition probabilities of the two sequences are the same except on the first three entries. Therefore, the distributions over topics of these two sequences will be very similar if we only consider the local transition probability. However, the number of transitions occurring in the sequence might be the key factor to express the semantic. For example, in the above sequences the transition $a \to a$ may represent some important features. For example, they may be the two successive alarms of network attack. Under this situation the model of SMMC, considering only the local transition probability, might fail to find that s_2 is more likely to be a sequence of network attack than s_1.

In the following we propose the model of BiLDA (Binary LDA), which considers the global transition probability to generate the sequence. The key idea of BiLDA is to consider any transition of two immediate tokens w, w' in a sequence s, denoted by $w \to w' \in \mathcal{V} \times \mathcal{V}$, as a word in the bag-of-words model. Then the same generative process as LDA are carried out to produce this *bag-of-transitions*.

The model of BiLDA is aimed to estimate the distribution over topics θ_{sz} for each sequence s and $\beta_{z,w \to w'}$ for each topic z. Every entry in $\beta_{z,w \to w'}$ is actually the probability that the corresponding transition $w \to w'$ is generated given the topic z, that is $\beta_{z,w \to w'} = p(w \to w'|z)$.

$$p(\{w_n \to w_{n+1}\}_{n=1}^{N-1}|\alpha, \beta, s) = \int p(\theta|\alpha) \prod_{n=1}^{N-1} \sum_{z=1}^{k} \underbrace{p(z|\theta, s)}_{p(z|s)} \underbrace{p(w_n \to w_{n+1}|\beta, z)}_{p(w \to w'|z)} \mathrm{d}\theta \tag{3}$$

where the mixture component $p(w \to w'|z)$ and the mixture proportion $p(z|s)$ are determined based on the observed global transition probability $p(w \to w'|s)$.

A. Discussion

It is clear that global activity probability $p(w)$ is equal to the normalized *bag-of-words* representation in text mining, local transition probability $p(w'|w)$ corresponds to the transition matrix in Markov chains, and global transition probability $p(w \to w')$ is essentially the normalized *bag-of-transitions* representation. Furthermore, from $p(w \to w')$ we can estimate $p(w'|w)$ and $p(w)$ directly, namely

$$p(w'|w) = \frac{p(w \to w')}{\sum_{w^*} p(w \to w^*)}, \quad p(w) \approx \sum_{w^*} p(w \to w^*). \tag{4}$$

The approximation of the right equation in (4) is due to the fact that the counts of $\sharp(w)$ and $\sharp(w \to w')$ are not exactly equal when the ending symbol of a sequence is w. Obviously, this proximation is reasonable when the sequence is long enough. Illustrated by Equation (4) we find that the global transition probability $p(w \to w'|s)$ considers the factors of both the local transition probability $p(w'|w,s)$ and the global activity probability $p(w|s)$. As shown in Table 2(c), the global transition probabilities of the two sequences are quite different, thus, it is more possible to distinguish the two sequences. Therefore, in this paper we suggest using the global transition probability in temporal activity profiling.

III. Topics over Temporal Activities

Each activity in a temporal sequence is always attached with a timestamp. This temporal information might be another important factor to refine the topics. In this section we propose the model of T-BiLDA, which injects the temporal information into BiLDA.

A. Model Description

In T-BiLDA, topic discovery is influenced not only by the co-occurrences of transitions, but also temporal information. We adopt the method in T-LDA[1] [15] to inject the temporal information into the generative process. This method associates with each topic a continuous distribution over time. We adopt normal distribution $\mathcal{N}(\cdot|\mu_z, \sigma_z^2)$ to express this time distribution in this study. The mean parameter μ_z locates the topic z over the time horizon while the standard deviation σ_z facilitates the time range estimation. In each round of transition generation the current topic generates not only a transition according to the global transition probabilities under this topic, but also the timestamp associated to the transition according to the time distribution on the topic. Specifically, the generative process of the T-BiLDA model can be described as follows:

1) draw multinomial distribution $\beta_{z,w\to w'} = p(w \to w'|z)$ over transitions for each topic z; draw Normal distribution ψ_z for each topic z;
2) for each sequence s, draw a multinomial distribution θ_s from a Dirichlet prior α; then for the transition $w_{sn} \to w_{s,n+1}$ in the sequence s:
   a) draw z_{sn} from θ_{sz};
   b) draw $w_{sn} \to w_{s,n+1}$ from $\beta_{z_{sn}}$;
   c) draw a timestamp t_{sn} from $\psi_{z_{sn}}$.

Although T-BiLDA adopts the similar method with T-LDA model [15] to inject the temporal information into the model, the characteristics of the data used in these two models are different. Each document used in the T-LDA model [15] is associated with only one timestamp, while the symbols in a sequence are associated with different timestamps. Additionally, the time ranges of the temporal sequences greatly overlap while the timestamps of the documents used in [15] overlap less. Thus, the temporal information in the temporal sequences can be fully utilized to refine the topics.

With the similar method it is easy to extend the model of SMMC into T-SMMC, which considers both local transition probabilities and temporal information. Figure 1 gives a summary of the topic models with the features they exploit respectively. These six methods will be compared for temporal activity modeling in the experimental section.

[1]The same with the model of TOT [15]. In Sections III and V we rename this model in order to indicate that this model is actually the LDA model plus the temporal information.

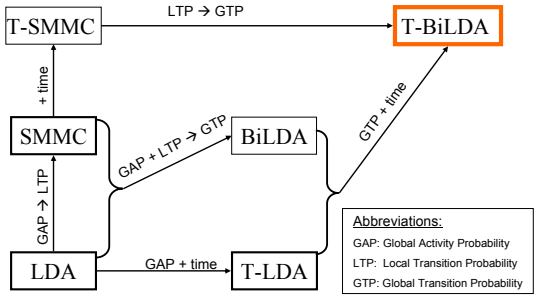

Fig. 1. Summary of Topic Models of this paper.

B. Model Learning by Variational Inference

We now estimate α, β and $\psi = (\mu, \sigma^2)$ as parameters and as well as the literatures [2], [11], [15] we simply fix α to $\vec{1}$. There are some approximate inference techniques available in the literature: variational methods, expectation propagation and Gibbs sampling. We choose a variational Expectation Maximization (EM) procedure, which is the approach taken in [3], [14], [1], [2], [4]. We firstly consider T-LDA and introduce the learning of T-SMMC and T-BiLDA later.

1) Variational E-step: Given a sequence of temporal activity, the likelihood of the this observed sequence is

$$L(\alpha, \beta, \psi) = \int p(\theta|\alpha) \prod_{n=1}^{N} (\sum_{z=1}^{k} p(z|\theta) p(w_n|z, \beta) p(t_n|z, \psi)) \, \mathrm{d}\theta \tag{5}$$

The the parameters are estimated by maximizing $L(\alpha, \beta, \psi)$, which is intractable as with other LDA type models. Therefore, we carry out variational EM. Suppose we have the variational distribution:

$$q(\theta, \vec{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{6}$$

where γ is a Dirichlet parameter vector with dimension k and each ϕ_n parametrizes a multinormial distribution on the latent topics satisfying $\mathrm{E}[Z_n] = \phi_n$.

Variational objective function. We need to maximize the lower bound $\mathcal{L}(\cdot)$ of the log likelihood for a single

sequence:

$$\log\, L(\alpha,\beta,\psi) \geq \mathcal{L}(\gamma,\phi;\alpha,\beta,\psi) =$$

$$\mathrm{E}_q[\log\, p(\theta|\alpha)] + \sum_{n=1}^{N} \mathrm{E}_q[\log\, p(z_n|\theta)] + \sum_{n=1}^{N} \mathrm{E}_q[\log\, p(w_n|z_n,\beta)]$$

$$+ \sum_{n=1}^{N} \mathrm{E}_q[\log\, p(t_n|z_n,\psi)] + \mathrm{H}(q) \tag{7}$$

Note that the first tree terms and the entropy of the variational distribution $\mathrm{H}(q)$ are the same as the original LDA. Expand and maximize the lower bound of log likelihood $\mathcal{L}(\cdot)$ with respect to ϕ_{nz} and γ_z. We have

$$\hat{\phi}_{nz} \propto \beta_{z,w_n} \cdot \mathcal{N}(t_n|\mu_z,\sigma_z^2) \cdot \exp[\Psi(\gamma_z) - \Psi(\sum_{z=1}^{k}\gamma_z)] \tag{8}$$

$$\hat{\gamma}_z = \alpha_z + \sum_{n=1}^{N} \phi_{nz} \tag{9}$$

where $n = 1, 2, ..., N$, Ψ is digamma function and $\mathcal{N}(\cdot|\mu_z,\sigma_z^2)$ is the normal probability density function with parameter $\psi_z = (\mu_z,\sigma_z^2)$. Detailed derivations are given in the Appendix for completeness.

2) M-step: The model parameters are on a set of sequences so we need to maximize the lower bound of the joint log likelihood across sequences by adding up the per-sequence log likelihoods. We add sequence index s to the quantities in the previous section, i.e. t_n becomes t_{sn}, ϕ_{nz} becomes ϕ_{snz} and so on. We maximize the joint low bound $\sum_s \mathcal{L}_s(\cdot)$ with respect to β and $\psi = (\mu,\sigma^2)$. We have

$$\hat{\beta}_{zw} \propto \sum_{s=1}^{S} \sum_{\{n|w_n=w\}} \phi_{snz} \tag{10}$$

with the normalization factor $\sum_w \hat{\beta}_{zw}$, and

$$\hat{\mu}_z = \frac{\sum\limits_{s,n} \phi_{dnz} t_{sn}}{\sum\limits_{s,n} \phi_{snz}}, \quad \hat{\sigma}_z^2 = \frac{\sum\limits_{s,n} \phi_{snz}(t_{sn}-\mu_z)^2}{\sum\limits_{s,n} \phi_{snz}} \tag{11}$$

Additionally, although it doesn't participate in the EM updates, the latent parameter θ can also be estimated after finishing the EM procedure

$$\hat{\theta}_{sz} \propto \sum_{n=1}^{N} \phi_{snz} \tag{12}$$

3) Variational Inference on T-BiLDA and T-SMMC: For the models BiLDA or T-BiLDA, the vocabulary contains the transitions (bigrams), we simply replace the input unigram tokens for the learning procedure of LDA or T-LDA with the bigrams. Then M-step updates of the topics, i.e. β's become

$$\hat{\beta}_{z,w\to w'} \propto \sum_{s=1}^{S} \sum_{\{n|w_n=w,w_{n+1}=w'\}} \phi_{snz} \tag{13}$$

with the normalization factor $\sum_{w,w'} \hat{\beta}_{z,w\to w'}$. To learn SMMC and T-SMMC we need an appropriate normalization factor to make the topics to Markov chains: $\sum_w \hat{\beta}_{z,w\to w'}$. The forms of updating the other parameters for these models remain as the same.

IV. Data Sets

We present experiments on two real-world data sets: the data of DDoS attack and system call traces. These two data sets are use to conduct qualitative and quantitative evaluation respectively.

A. Data of DDoS Attack

Usually, an attacker carries out a DDoS attack by exploiting a group of machines in a network and identifies the vulnerable ones among them as the DDoS *compromised machines*. The intruder then loads cracking tools onto these compromised machines. Finally, with a single command the intruder instructs all the controlled machines to launch flood attacks against a DDoS *victim*. The inundation of packets to the victim causes a denial-of-service.

The data set used in the experiments[2] contains the alert logs on the Intrusion Detection System (IDS) during a DDoS attack. This attack can be divided into 4 phases: 1) an outside attacker probed the inside network and identified three compromised machines; 2) the attacker broke into these compromised machines; 3) the attack installed software onto these compromised machines; 4) a DDoS attack was launched to an victim in the internal network. We also know the exact time ranges for these 4 phases. Thus, we are given a long sequence of 504 alert symbols with timestamps, which can be divided into 4 subsequences according to the time ranges for the 4 attack phases. Then, the 4 sub-sequences, corresponding to the 4 attack phases, are presented in the top panel of Figure 3, where we arrange these 4 attack phases onto a time horizon and mark the time ranges they took place as well.

In the experiments, we try to identify the 4 attack phases as well as their time ranges via topic modeling. However, it is impossible to launch a topic model on a single sequence, thus, we need to reconstruct multiple sequences, each of which only contains the alert symbols on a certain IP. Since all the alert symbols involve with 22 inside IPs and 291 outside IPs which can be viewed as two sets of indices for sequence segmentation, we obtain 22 sequences of inside IP and 291 sequences of outside IP as two distinct data sets. One, denoted by D_{in}, contains the sequences on the inside IPs, while the other, denoted by D_{out}, includes the

[2]http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/2000/LLS_DDOS_1.0.html

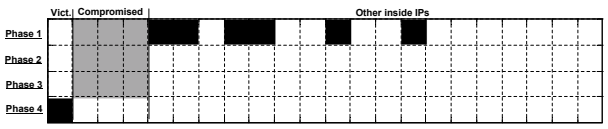

Fig. 2. Normalized Involvement Matrix

sequences on the outside IPs. All these sequences share the vocabulary of 25 symbols, thus, may contain $25 \times 25 = 625$ transitions.

To further visualize the the relationship between the attack phases and the IPs we construct the following *involvement matrix* M, where $M(i, j) = 1$ when the j-th IP are involved with the i-th attack phase, otherwise $M(i, j) = 0$. Each row in M corresponds to an attack phase while each column corresponds to an IP. We then normalize the entry values by columns. Thus, the *normalized involvement matrix* can be visualized by encoding the entry values into gray scales and filling them onto the corresponding blocks. This way the attack scenario behind the sequences in D_{in} is shown in Figure 2. This figure shows that 1) the first IP was the victim since it was only involved with the 4-th attach phase; 2) the next three IPs were compromised machines since they were all involved with the first three phases; 3) some other IPs were also probed in the first phase, however, they were not broken into. In the experiments we will reconstruct this visualization by topic modeling. The data set in this subsection is used for qualitative evaluation of the topic modeling methods.

B. Data of System Call Traces

The University of New Mexico provides various sets of system call traces[3]. The traces contain the temporally-ordered system calls together with the IDs of the processes that generated them. In this data set we construct the sequences of system calls with respect to each ID. In this case, we have no absolute timestamps on the system calls, thus, the relative position n/N is used as the timestamp of the n-th symbol of a sequence (N is the length of this sequence). For this purpose we manually select the sequences within the length interval of $[50, 200]$. Finally, we obtain 747 sequences from three categories of traces: 650 from *synthetic UNM lpr*, 49 from the *CERT synthethic sendmail* and 48 from the *daemon* part. The category information can be used as the true class label to evaluate the performance of clustering on the resultant topic distributions of sequences. All these sequences share the vocabulary of 58 symbols. The data set in this subsection is used for quantitative evaluation of the topic modeling methods.

[3]http://www.cs.unm.edu/ immsec/systemcalls.htm

V. Experimental Results

In this section we present the topics discovered by the T-BiLDA model and compare them with the topics with the other 5 baseline methods in Figure 1. In the qualitative evaluation on the first data set we reconstruct the attack scenario by topic modeling. In the quantitative evaluation on the second data set we demonstrate the ability of the T-BiLDA model in next token prediction and clustering.

A. Topics Discovered for DDoS

In this subsection we evaluate the ability of the discovered topics in attack scenario reconstruction. Ideally, some of the discovered topics should corresponds to the DDoS attack phases. Additionally, we show that the global transition probabilities learned from (T-)BiLDA precisely carry the information of both global symbol probabilities and local transition probabilities. In these experiments we fix the number of the topics $T = 6$, which is bigger than the number of the attack phases. After topic modeling we map each of the 4 attack phases to one of the 6 resultant topics by the following method. Clearly we can estimate the means of timestamps μ_i and the proportions of each alert symbol (or transitions) β_i of Phase i. Then, the topic corresponds to Phase i is determined by:

$$\arg\min_z \left(|\mu_z - \mu_i| + \lambda \cdot \mathrm{KL}(\beta_z, \beta_i)\right)$$

where $\mathrm{KL}(\cdot)$ is the Kullback-Leibler (KL) divergence, and λ is the tradeoff parameters between these two factors.

1) Attack Scenario Reconstruction for DDoS: Attack scenario reconstruction is one of the intrusion detection tasks. We need to specify *where* and *when* the events happen as well as the contents of the events, i.e. *what* the events are.

We first answer the question of *where* via the latent parameter θ, which encodes the relationship between the sequences and the the latent topics. $\theta_{sz} > 0$ means the s-th IP is involved in the z-th latent topic. Therefore, we can evaluate θ via comparing it with the normalized involvement matrix shown in Figure 2. Figure 4 gives the topic distributions from all the 6 topic models considered in this paper for the sequences in D_{in}. It shows that 1) T-BiLDA fits the real normalized involvement matrix best; it succeeds in identifying the victim, the compromised machines and the other machines; 2) The models with temporal information outperform the models without the temporal information; 3) T-SMMC fails to identify the

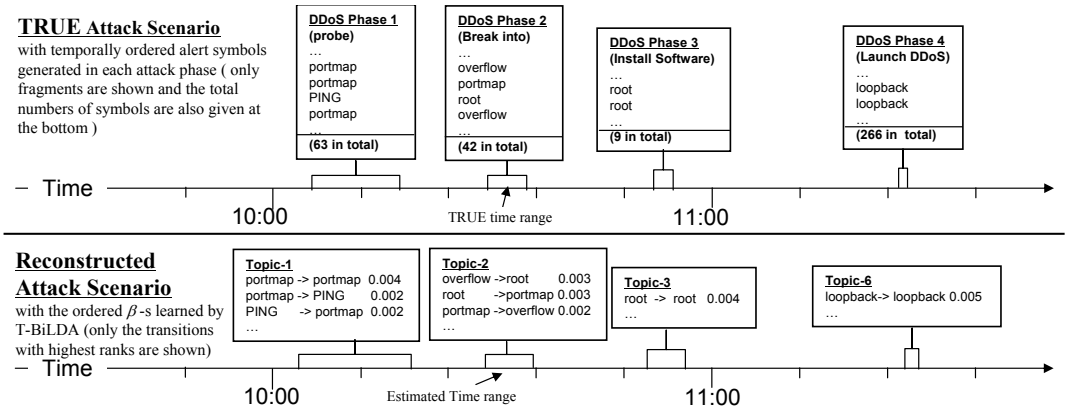

Fig. 3. True and Reconstructed Attack Scenario

roles of compromised machines; it considers some machines, which were not broken into by the attacker, as the compromised machines.

For the *when* question the normal distribution over time on each topic provides the method to estimate the time range of each topic. Specifically, with the temporal normal distribution $\mathcal{N}(\cdot|\mu_z, \sigma_z^2)$ of the z-th topic, we estimate the time range $[t_z^{\text{start}}, t_z^{\text{end}}]$ for this topic via

$$\hat{t}_z^{\text{start}} = \mu_z - 1.96\ \sigma_z, \quad \hat{t}_z^{\text{end}} = \mu_z + 1.96\ \sigma_z, \qquad (14)$$

Here we use the fact that the 95% confidence interval for a normal distribution is roughly 1.96 standard deviations from the mean. For the *what* question we use the transitions or symbols with significant high weights in β as the representative contents of each topic. The bottom panel of Figure 3 is the answer of *when* and *what* from the model of T-BiLDA based on the data set of D_{in}. It shows that T-BiLDA precisely reconstruct the attack phases both on the time range and the content of each phase.

2) From Global Transition Prob. to Local Transition Prob. and Global Activity Prob.: Given the global transition probabilities of $p(w \to w')$, we can get estimated global activity probabilities $\hat{p}(w)$ and local transition probabilities $\hat{p}(w'|w)$ using Equation (4). Next, we will show that the global transition probabilities learned from (T-)BiLDA precisely carry the information of both global activity probabilities and local transition probabilities.

Since we know all the alert symbols in each of the 4 attack phases, the true global activity probabilities and local transition probabilities can be computed for each attack phase. We compare these true values with the estimated values by the resultant topics from different methods. In these comparisons only the 8 symbols appearing in the 4 attack phases are selected to compute these values. Here, we only show the estimations from the temporal models

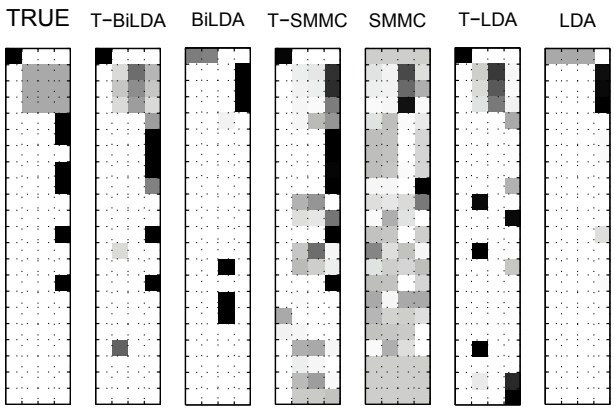

Fig. 4. Comparison between the normalized involvement matrix and the distributions over topics (rotated for saving space)

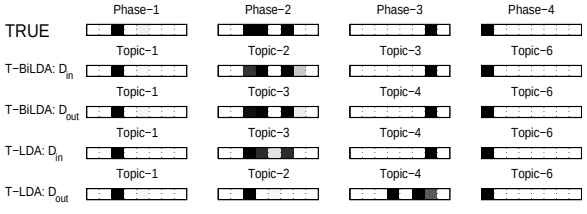

Fig. 5. Estimating global activity probabilities

since the non-temporal models fails to detect the topics of the sequences as shown in Figure 4.

In Figure 5 we compare the estimated global activity probabilities with the true values for each attack phase. It shows that both T-LDA and T-BiLDA precisely depict the global activity probabilities. Note that it is impossible

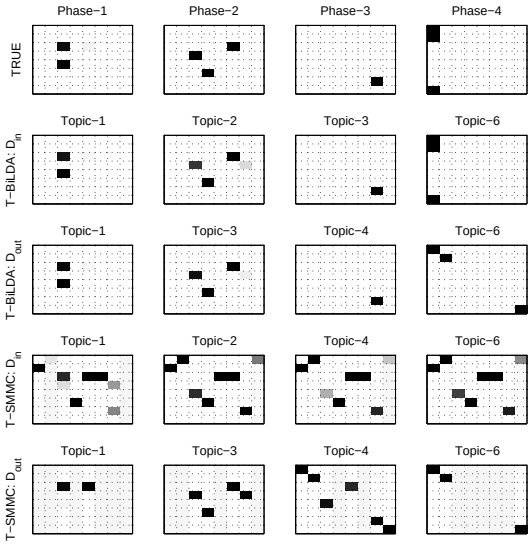

Fig. 6. Estimating local transition probabilities

to estimate global symbol probabilities from the model of (T-)SMMC.

In Figure 6 we compare the estimated local transition matrices with the true values for each attack phase. It is clear that the estimation by T-BiLDA outperforms that by T-SMMC. This conclusion coincides with the fact that T-SMMC fails to detect the topics of some sequences as shown in Figure 4. Note also that it is impossible to estimate local transition probabilities from the model of (T-)LDA.

These results further prove our analysis in Section II that the global transition probabilities considered by (T-)BiLDA covers the information of both global activity probabilities and local transition probabilities.

B. Experiments on System Call Traces

The experiments on system call traces quantitatively evaluate the performance of the topic models. Specifically, we show their micro descriptive ability by next token prediction and their macro descriptive ability by sequence clustering.

1) Next Activity Prediction: We demonstrate the descriptive abilities of the proposed models via making the next activity predictions on the data set of system call traces. The probabilities of observing symbol w_{n+1} given the very previous symbols w_n based on the 6 topic models of LDA, T-LDA, SMMC, T-SMMC, BiLDA, and T-BiLDA are given as follows respectively:

$$\begin{aligned}
P(w_{n+1}=w|\theta,\beta) &\propto \sum_{z=1}^{k}\theta_z\cdot\beta_{zw},\\
P(w_{n+1}=w|t_n;\theta,\beta,\psi) &\propto \sum_{z=1}^{k}\theta_z\cdot\beta_{zw}\cdot\mathcal{N}(t_n|\psi_z),\\
p(w_{n+1}=w|w_n;\theta,\beta) &\propto \sum_{z=1}^{k}[\theta_z\cdot\beta_{z,w|w_n}],\\
p(w_{n+1}=w|w_n,t_n;\theta,\beta,\psi) &\propto \sum_{z=1}^{k}[\theta_z\cdot\beta_{z,w|w_n}\cdot\mathcal{N}(t_n|\psi_z)],\\
p(w_{n+1}=w|w_n;\theta,\beta) &\propto \sum_{z=1}^{k}[\theta_z\cdot\beta_{z,w_n\to w}],\\
p(w_{n+1}=w|w_n,t_n;\theta,\beta,\psi) &\propto \sum_{z=1}^{k}[\theta_z\cdot\beta_{z,w_n\to w}\cdot\mathcal{N}(t_n|\psi_z)].
\end{aligned} \tag{15}$$

In the experiments we randomly sample 100 sequences as the training set to learn the models with different topic numbers of k, and the remaining sequences are held out for testing. In testing, we first infer the distribution over topics of each sequence in the test set using the learned models. Then, for each sequence we randomly sample one token w_n and predict its next token w_{n+1}. It seems that the settings of this prediction are not realistic since the topic distribution of the whole sequence is required in advance[4]. However, these experiments are still reasonable to check the descriptive ability of the topic models. Moreover, since the prediction is made on a held out testing set, we actually evaluate the generalization performance of the models by this experiment.

Two measures are used for the evaluation, *accuracy* and *predictive perplexity*. The accuracy value is simply the ratio between the number of right predictions and the total prediction number, and the predictive perplexity is defined as

$$\exp\{-\frac{1}{M}\sum_{m=1}^{M}p(w_{m,n+1})\} \tag{16}$$

where $p(w_{m,n+1})$ is the probability of the true next symbol given all the other information and M is the number of testing documents. The smaller predictive perplexity score indicates the better generalization performance. We sample the training sets and the symbols on each test sequence for both 10 times to calculate the average values of the two measures. Figure 7 shows the prediction performance under different topic numbers. It is clear that 1) T-BiLDA consistently performs the best under different settings of topic number; 2) the temporal information always improves the prediction performance.

[4]For the real-world problem of next activity prediction we can infer the topic distribution and the sequence length N based on the given incomplete sequence, and then make the prediction using the similar method.

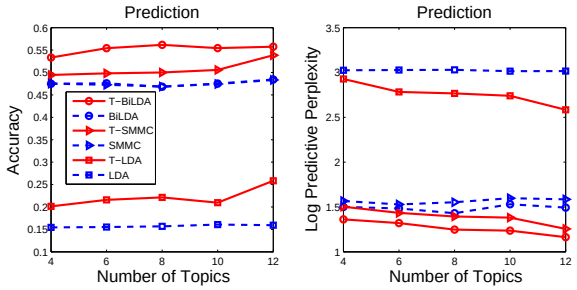

Fig. 7. Results of prediction

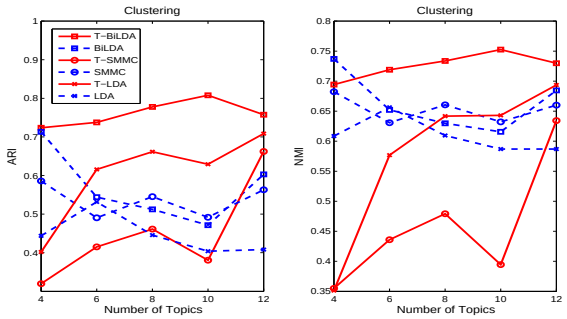

Fig. 8. Results of clustering

C. Clustering for System Call Traces

Next we evaluate the topic models on their descriptive abilities in a macro way. Specifically, we use the distribution over topics of a sequence as the feature vector for clustering. Here, K-means is adopted to clustering the sequences. The three classes of sequences known in advance is used as the ground truth of the clustering. Two evaluation measures, *Adjusted Rand Index* (ARI) [6] and *Normalized Mutual Information* (NMI) [12], are calculated for evaluation. Larger ARI and NMI scores mean better clustering performance. As the same with the prediction task we also randomly sample 100 sequences as the training set to learn the models, and the remaining sequences are held out for testing.

The average scores over 25-round running of K-means are shown in Figure 8. It shows that 1) T-BiLDA still outperforms the other models; 2) T-LDA outperforms T-SMMC since the global activity probabilities are more important than the local transition probabilities in the clustering task.

VI. Related Work

Discovering sequential patterns via topic models has been extensively studied in the literatures. Simplicial Mixture of Markov Chain (SMMC) [4], as introduced in Section II, is a representative one of topic models. Sequential Generative Topographic Mapping (SGTM) models [8] replace the Dirichlet prior in SMMC with a much more complex prior. The prior in SGTM offers more flexibility than the Dirichlet prior so that SGTM is robust against the data size. HMMLDA [5] is a generative composite model that applies hidden Markov model to characterize syntax information and use LDA to model semantic information. Topical N-grams [16] discovers topics as well as phrases, i.e. the local dependency between words. All of the above models have no capability of modeling temporal information.

There are another sort of models derived from LDA with the consideration of temporal information such as Dynamic Topic Models (DTM) [1], Continuous Time Dynamic Topic Models (cDTM) [14], Multiscale Topic Tomography Model (MTTM) [9] and Topic over Time (TOT) [15]. DTM and cDTM both model the evolution of topics over time while TOT uses the temporal information to refine the topics. MTTM accomplishes both the two objectives, however, it cannot be used to generate the sequential activities considered in this paper.

VII. Conclusions

In this paper we proposed a probabilistic mixture model, called T-BiLDA, for modeling sequences of temporal activities. Compared to the previous work, T-BiLDA jointly exploits both the Markov dependence between the immediate activities and the temporal information on each activity in a sequence. To modeling the Markov dependence we propose the concept of global transition probability, and show that this concept covers the information of both global activity probabilities and local transition probabilities. The experiments on the data of DDoS attack and system call traces show the effectiveness of this model qualitatively and quantitatively. In the near future we will apply this model into more other applications with temporal patterns, such as routine detection on the human daily activities [7] and sustainability characterization of the temperature curves on the ensemble of chiller units in a data center [10].

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

[2] D. M. Blei and J. McAuliffe. Supervised topic models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128, Cambridge, MA, 2008. MIT Press.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning Research*, 3:993–1022, Januaray 2003.

[4] M. Girolami and A. Kabán. Sequential activity profiling: Latent dirichlet allocation of markov chains. *Data Min. Knowl. Discov.*, 10(3):175–196, 2005.
[5] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, pages 537–544, 2005.
[6] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, pages 193–218, 1985.
[7] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In H. Y. Youn and W.-D. Cho, editors, *UbiComp*, volume 344 of *ACM International Conference Proceeding Series*, pages 10–19. ACM, 2008.
[8] A. Kaban. Predictive modeling of hterogeneous sequence collections by topographic ordering of histories. *Machine Learning*, 68:63–95, 2007.
[9] R. Nallapati, S. Ditmore, J. D. Lafferty, and K. Ung. Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 520–529, 2007.
[10] D. Patnaik, M. Marwah, R. Sharma, and N. Ramakrishnan. Sustainable operation and management of data center chillers using temporal data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. to appear*, 2009.
[11] M. Steyvers. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning*, pages 427–448, 2007.
[12] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, 3:583–617, 2002.
[13] H. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.
[14] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 579–586, 2008.
[15] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.
[16] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the Seventh IEEE International Conference on Data Mining*, pages 697–702, 2007.

Appendix

E-step: Expand the right side of Equation (7) we have:

$$\begin{aligned}\mathcal{L}(\gamma,\phi;\alpha,\beta,\psi) &= \log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{z=1}^{k}\log\Gamma(\alpha_z)\\ &+ \sum_{z=1}^{k}(\alpha_z-1)[\Psi(\gamma_z)-\Psi(\sum_{z=1}^{k}\gamma_j)]\\ &+ \sum_{n=1}^{N}\sum_{z=1}^{k}\phi_{nz}(\Psi(\gamma_z)-\Psi(\sum_{j=1}^{k})\gamma_j) + \sum_{n=1}^{N}\sum_{z=1}^{k}\phi_{nz}\log\beta_{z,w_n}\\ &+ \sum_{n=1}^{N}\sum_{z=1}^{k}\phi_{nz}\log[\mathcal{N}(t_n|\mu_z,\sigma_z^2)]\\ &-\log\,\Gamma(\sum_{j=1}^{k}\gamma_j) - \sum_{z=1}^{k}(\gamma_z-1)[\Psi(\gamma_z)-\Psi(\sum_{z=1}^{k}\gamma_z)]\\ &- \sum_{n=1}^{N}\sum_{z=1}^{k}\phi_{nz}\log\,\phi_{nz}\end{aligned} \tag{17}$$

Estimation of ϕ_{nz}: isolating the terms in Equation (17) that reference ϕ and including the constraint for $\sum_{z=1}^{k}\phi_{nz}=1$.

$$\begin{aligned}\mathcal{L}_{[\phi_{nz}]} &= \phi_{nz}[\Psi(\gamma_z)-\Psi(\sum_{j=1}^{k}\gamma_z)] + \phi_{nz}\log\beta_{z,w_n} - \phi_{nz}\log\phi_{nz}\\ &+\phi_{nz}\log\mathcal{N}(t_n|\mu_z,\sigma_z^2) + \lambda_n(\sum_{z=1}^{k}\phi_{nz}-1)\end{aligned} \tag{18}$$

Differentiating with respect to the parameter and setting equal to 0 gives

$$\begin{aligned}\tfrac{\partial\mathcal{L}_{[\phi_{nz}]}}{\partial\phi_{nz}} &= \Psi(\gamma_z) - \Psi(\sum_{z=1}^{k}\gamma_z) + \log\beta_{z,w_n} - \log\phi_{nz}\\ &+\log\mathcal{N}(t_n|\mu_z,\sigma_z^2) + \lambda_n = 0\end{aligned} \tag{19}$$

$$\hat{\phi}_{nz} \propto \beta_{z,w_n}\cdot\mathcal{N}(t_n|\mu_z,\sigma_z^2)\cdot\exp[\Psi(\gamma_z)-\Psi(\sum_{z=1}^{k}\gamma_z)] \tag{20}$$

The estimation of γ is the same as in [3].

M-step: We only need to consider the inference of μ and σ^2. The inference for β and α is the same as [3].

$$\mathcal{L}_{[\mu_z,\sigma_z^2]} = \sum_{s=1}^{M}\sum_{n=1}^{N}\phi_{dnz}[-\frac{1}{2}\log(2\pi\sigma_z^2) - (t_{dn}-\mu_z)^2/2\sigma_z^2] \tag{21}$$

Let $\nabla\mathcal{L}_{[\mu_z,\sigma_z^2]}=0$ we have

$$\hat{\mu}_z = \frac{\sum_{s,n}\phi_{snz}t_{sn}}{\sum_{s,n}\phi_{snz}}, \quad \hat{\sigma}_z^2 = \frac{\sum_{s,n}\phi_{snz}(t_{sn}-\mu_z)^2}{\sum_{s,n}\phi_{snz}} \tag{22}$$