Discovering Topical Interactions in Text-based Cascades using Hidden Markov Hawkes Processes

Jayesh Choudhari, Anirban Dasgupta
IIT Gandhinagar, India
Email: {choudhari.jayesh, anirbandg}@iitgn.ac.in

Indrajit Bhattacharya
TCS Research, India
Email: b.indrajit@tcs.com

Srikanta Bedathur
IIT Delhi, India
Email: srikanta@cse.iitd.ac.in

***Abstract*—Social media conversations unfold based on complex interactions between users, topics and time. While recent models have been proposed to capture network strengths between users, users' topical preferences and temporal patterns between posting and response times, interaction patterns between topics has not been studied. We propose the Hidden Markov Hawkes Process (HMHP) that incorporates topical Markov Chains within Hawkes processes to jointly model topical interactions along with user-user and user-topic patterns. We propose a Gibbs sampling algorithm for HMHP that jointly infers the network strengths, diffusion paths, the topics of the posts as well as the topic-topic interactions. We show using experiments on real and semi-synthetic data that HMHP is able to generalize better and recover the network strengths, topics and diffusion paths more accurately than state-of-the-art baselines. More interestingly, HMHP finds insightful interactions between topics in real tweets which no existing model is able to do.**

***Keywords*—Hawkes Process, Textual cascades, Markov Chain, Gibbs Sampling, Topical Interactions**

I. Introduction

A popular area of recent research has been the study of information diffusion cascades, where information spreads over a social network when a 'parent' event from one infected node influences a 'child' event at neighboring node [5], [11], [18], [6], [10]. The action of propagating information between two neighboring nodes depends on various factors, such as the strength of influence between the nodes, the topical content of the parent event and the extent of interest of the child node towards that topic. However, such explanatory variables and also the identity of influencing or parent event for any event, are typically unobserved and need to be inferred.

A recent body of work in this area has proposed increasingly sophisticated models for information cascades. To account for the temporal burstiness of events, Hawkes processes, which are self-exciting point processes, have been proposed to model their time stamps [15]. Also, influences travel more quickly over stronger social ties. This is modeled by the Network Hawkes process [12] by incorporating connection strengths between users into the intensity function of the Hawkes process. When events additionally have associated textual data, parent and child events in a cascade are typically similar in their topical content. The Hawkes Topic Model (HTM) [11] captures this by combining the Network Hawkes process with a LDA-like generative process for the textual content where the topic mixture for a child event is close to that for a parent event.

Despite its strengths, the HTM fails to capture one important aspect of the richness of information cascades. Typically, there are sequential patterns in the textual content of different events within a cascade. In terms of topical representation of the events, topics of parent and child events display repeating patterns. At the same time, these cannot be assigned to the same topic. Instead, our model assigns these to different topics, and recognizes this to be a frequently appearing topic pair in other parent-child tweets. This serves as additional evidence for inferring a parent-child relationship between this tweet pair.

However, just like topic distributions, topic interaction patterns are also typically latent, and need to be inferred. Interestingly, inferring topic interaction patterns, in turn, benefits from more accurate parent and topic assignment to events. This calls for joint inference of topic interaction patterns and the other latent variables.

We propose a generative model for textual cascades that captures topical interactions in addition to temporal burstiness, network strengths, and user topic preferences. Temporal and network burstiness is captured using a Hawkes process over the social network [12]. Topical interactions are modeled by combing this process with a Markov chain over event topics along a path of the diffusion cascade. Such topical Markov chains have appeared in the literature [8], [9], [1], [2], [4], but in the very different context of modeling word sequences in a document. Our model effectively integrates topical Markov chains with the Network Hawkes process and we call it the Hidden Markov Hawkes Process (HMHP).

We derive a simple and efficient collapsed Gibbs Sampling algorithm for inference using our model. This algorithm is significantly more scalable than the variational algorithm for the HTM, and allows us to analyze large collections of real tweets. Using extensive experiments we validate that HMHP fits information cascades better compared to models that do not consider topical interactions. With the aid of semi-synthetic data, we show that modeling topical interactions indeed leads to more accurate identification of event parents and topics. More importantly, HMHP is able to identify interesting topical interactions in real tweets, which is beyond the capability of any existing information diffusion model.

II. Model

We consider a set of nodes $V = \{1, \ldots, n\}$, representing content producers, and a set of directed edges E among them, representing the underlying graph using which information propagates. For every edge (u, v), W_{uv} denotes the weight of the edge, capturing the extent of influence between content producers or users u and v. For each event e, representing e.g. a social media post, t_e denotes its posting time, c_e the user who creates the post and d_e the document associated with the post. Event e could be *spontaneous*, or a *diffusion* event, meaning that it is triggered by some recent event by one of the followees of user c_e. The variable z_e denotes the unique parent event that triggered the creation of event e if it is a diffusion event. For spontaneous events, z_e is set to 0. The generated event e is then consumed by each of the followers of the user c_e, thereby triggering them in turn to create multiple events and producing an information cascade. Our generative model can be broken up into two distinct stages - generating the cascade events, and then the event documents.

A. Generating Cascade Events

We first generate (t_e, c_e, z_e) for all events using a multivariate Hawkes process (MHP) following existing models [16], [11], [12], where the users can mutually excite each other to produce events. For each node v, we define $\lambda_v(t)$, a rate at time t, as a superposition of the base intensity μ_v for user v, and the impulse responses for each event e_n that has happened at time t_n at a followee node c_n.

$$\lambda_v(t) = \mu_v(t) + \sum_{n=1}^{|\mathcal{H}_{t-}|} h_{c_n,v}(t - t_n) \tag{1}$$

where $h_{c_n,v}(t - t_n)$ is the impulse response of user c_n on the user v, and $\mathcal{H}_{t^-}$ is the history of events upto time t. The impulse response can be decomposed as the product of the influence W_{uv} of user u on v, and a time-kernel as follows:

$$h_{u,v}(\Delta t) = W_{u,v} f(\Delta t) \tag{2}$$

We use a simple exponential kernel $f(\Delta t) = exp(-\Delta t)$, as this is not the main thrust of our work. Following [16], we generate the events using a level-wise generation process. Level 0, denoted as Π_0, contains all the spontaneous events, generated using the base rates of the users. The events Π_l at level l, are generated according to the following non-homogenous Poisson process

$$\Pi_l \sim Poisson\left(\sum_{(t_n,c_n,z_n)\in\Pi_{l-1}} h_{c_n,\cdot}(t - t_n)\right) \tag{3}$$

B. Generating Event Documents

The document d_e for e is drawn using a topic model over a vocabulary of size $\mathcal{W}$. For spontaneous events, the topic choice η_e depends on the topic preference of the user c_e. We deviate from existing literature in modeling the topic of a diffusion event. Instead of being identical or 'close' to the topic of the triggering event [11], the diffusion event may be on a 'related' topic. We model this transition between related topics using a Markov Chain over topics, involving a topic transition matrix $\mathcal{T}$. This enables us to capture repeating patterns in topical transitions between parent and child events. Since the topical sequence is 'hidden' and observed only indirectly through the content of the events, we call our model the Hidden Markov Hawkes Process (HMHP).

We have K topics, denoted $\{\boldsymbol{\zeta}_k\}$, assumed to be probability distributions over words (vocabulary with size $\mathcal{W}$), generated *iid* from a Dirichlet distribution. We also assume the existence of a topic-topic interaction matrix $\boldsymbol{\mathcal{T}}$, where $\boldsymbol{\mathcal{T}}_k$, again sampled from a Dirichlet distribution, denotes the distribution over 'child topics' for a 'parent topic' k. Since our data of interest is tweets (short documents), we model a document d_e as having a single topic η_e. To generate d_e, we first follow a Markov process for sampling the topic η_e conditioned on the topic of its parent event η_{z_e}, followed by sampling the words *iid* according to the chosen topic. For spontaneous events the topic for its document is sampled randomly from the preferred distribution over topics $\boldsymbol{\phi}_u$ for the corresponding user u.

An important consideration in the design of our model is the use of conjugate priors. As we will see in the Inference section such priors play a crucial role in the design of efficient and simple sampling-based inference algorithms. Models such as HTM [11], which have to sacrifice conjugacy to model data complexity, have to resort to more complex variational inference strategies.

We summarize the entire generative process below.

1) Generate (t_e, c_e, z_e) for all events according to the process described in previous sub-section.
2) For each topic k: sample $\boldsymbol{\zeta}_k \sim Dir_{\mathcal{W}}(\boldsymbol{\alpha})$ and $\boldsymbol{\mathcal{T}}_k \sim Dir_K(\boldsymbol{\beta})$
3) For each node v: sample $\boldsymbol{\phi}_v \sim Dir_K(\boldsymbol{\gamma})$
4) For each event e at node $c_e = v$:
   a) **if** $z_e = 0$ (level 0 event): sample topic $\eta_e \sim Discrete_K(\boldsymbol{\phi}_v)$ **else**: sample topic $\eta_e \sim Discrete_K(\boldsymbol{\mathcal{T}}_{\eta_{z_e}})$
   b) Sample document length $N_e \sim Poisson(\lambda)$
   c) For $w = 1 \ldots N_e$: draw word $x_{e,w} \sim Discrete_{\mathcal{W}}(\boldsymbol{\zeta}_{\eta_e})$

The resultant joint likelihood can we written as follows:

$$P(E, \boldsymbol{\Phi}, \boldsymbol{\mathcal{T}}, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{\mu}) =$$
$$\prod_{v\in V} P(\boldsymbol{\phi}_v \mid \boldsymbol{\gamma}) \times \prod_{k=1}^{K} P(\boldsymbol{\zeta}_k \mid \boldsymbol{\alpha}) \times \prod_{k=1}^{K} P(\boldsymbol{\mathcal{T}}_k \mid \boldsymbol{\beta})$$
$$\times \prod_{e\in E} \left\{ \left[\prod_{e':t_{e'}<t_e} P(\eta_e|\boldsymbol{\mathcal{T}}_{\eta_{z_e}})^{\delta_{z_e,e'}} \right] P(\eta_e|\boldsymbol{\phi}_v)^{\delta_{z_e,0}} \right\}$$
$$\times \prod_{e\in E} \left[\prod_{w=1}^{N_e} P(x_{e,w}|\eta_e, \boldsymbol{\zeta}_{\eta_e}) \right]$$
$$\times \prod_{v\in V} \left[exp\left(-\int_0^T \mu_v(\tau)d\tau \right) \prod_{e\in E} \mu_v(t_e)^{\delta_{c_e,v}\delta_{z_e,0}} \right]$$
$$\times \prod_{e\in E}\prod_{v\in V} \left[exp\left(-\int_{t_e}^T h_{c_e,v}(\Delta\tau)d\tau \right) \prod_{e'\in E} h_{c_e,c_{e'}}(\Delta(t_{e'}))^{\delta_{c_{e'},v}\delta_{z_{e'},e}} \right] \tag{4}$$

Here, $\delta_{c_e,v}$ is an indicator for the event e being triggered at node v, $\delta_{z_{e'},e}$ is an indicator for event e being parent of event e', $\Delta\tau$ is $(\tau - t_e)$, $\Delta(t_{e'})$ is $(t_{e'} - t_e)$ and T is the time horizon.

III. Inference

The latent event variables are the topic η_e and the diffusion parent z_e.The process parameters to be estimated include the user-user influence values W_{uv}, the topic transition matrix $\mathcal{T}$, and the user-topic preferences.

By making use of the conjugate priors, we perform inference using collapsed Gibbs sampling. We integrate out the topic distributions over words $\boldsymbol{\zeta}$, the topic interaction distributions $\boldsymbol{\mathcal{T}}$ and the user-topic preference distributions $\boldsymbol{\Phi}$. The continuous valued latent variables that remain are the connection strengths $W_{u,v}$. The resulting collapsed Gibbs sampling algorithm iteratively samples individual topic and parent assignments and the connection strengths from their conditional distributions given current assignments to all other variables until convergence. We next describe the conditional distributions for sampling the different variables in the algorithm. The overall algorithm is described in Algorithm 1 below.

Algorithm 1 Gibbs Sampler

Initialize η_e for all events
Initialize z_e for all events
Initialize $W_{u,v}$ for all u, v in the followers map
for $iter = 0$; $iter\ != maxIter$; $iter++$ **do**
  for $e \in allEvents$ **do** ▷ Sampling topic
    $\eta_e \sim P(\eta_e = k \mid \boldsymbol{\eta}_{\neg e}, \boldsymbol{z}, \{X\}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu})$
  for $e \in allEvents$ **do** ▷ Sampling Parent
    $z_e \sim P(z_e = e' | E_t, \boldsymbol{z}_{\neg e}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{\mu})$
  for $(u, v) \in Edges$ **do** ▷ Estimate user-user influence
    $W_{u,v} = Mean(Gamma(N_{uv} + \alpha', N_u + \beta'))$

a) Topic assignment: The conditional probability for topic η_e of a diffusion event e ($z_e \neq 0$) being k when the current parent has topic k' is the following:

$$P(\eta_e = k \mid \{x_{e\cdot}\}, \eta_{z_e} = k', \boldsymbol{\eta}_{\setminus z_e}, \{z_e\}) \propto \frac{\beta_k + N^{(\neg(z_e,e))}_{k',k}}{(\sum_l \beta_l) + \boldsymbol{N}^{(\neg(z_e,e))}_{k'}} \times \frac{\prod_{l'=1}^{K} \prod_{i=0}^{N^{(C_e)}_{k,l'}-1} (\beta_{l'} + N^{(\neg C_e)}_{k,l'} + i)}{\prod_{i=0}^{\boldsymbol{N}^{(C_e)}_{k}-1} ((\sum_{l'} \beta_{l'}) + \boldsymbol{N}^{\neg C_e}_{k} + i)} \times \frac{\prod_{w \in d_e} \prod_{i=0}^{N^w_e - 1} (\alpha_w + \mathfrak{T}^{\neg e}_{k,w} + i)}{\prod_{i=0}^{N_e - 1} ((\sum_{w \in \mathcal{W}} \alpha_w) + \mathfrak{T}^{\neg e}_{k} + i)} \tag{5}$$

Here $N^{(C)}_{k',k}$ denotes the number of parent-child event pairs with topics k and k', $(\neg(z_e, e))$ denotes all edges excluding (z_e, e), C_e is the set of edges from event e to its child events ($\neg C_e$ being its complement), and $\mathfrak{T}^{\neg e}_{k,w}$ is the number of occurrences of word w under topic k in events other than e. Further, $\boldsymbol{N}^C_k = \sum_{k'} N^C_{k,k'}$, and N^w_e is count of w in d_e. The first term is the conditional probability of transitioning from parent topic k' to this event's topic k, the second term is that of transitioning from this topic k to each child event's topic l', and the third term is that of observing the words in the event document given topic k. It is important to observe how this conditional distribution pools together three different sources of evidence for an event's topic. Even when the document words do not provide sufficient evidence for the topic, the parent and children topics taken together can significantly reduce the uncertainty.

For spontaneous events ($z_e = 0$), the conditional probability looks similar to Eqn. 5. Only the first term changes to $\frac{\gamma_k + \mathfrak{U}^{(\neg e)}_{v,k}}{(\sum_k \gamma_k) + \boldsymbol{\mathfrak{U}}^{(\neg e)}_{v}}$, where $\mathfrak{U}^{(\neg e)}_{v,k}$ is the number of events by user v with topic k, discounting event e. This captures the probability of node v picking topic k from its preferred topics.

b) Parent assignment: The conditional probability of event e' being the parent z_e of event e looks as follows:

$$P(z_e = e' | E_t, \boldsymbol{z}_{\neg e}, \boldsymbol{W}) \propto \frac{(\beta_k + N_{k',k} - 1)}{((\sum_{k=1}^{K} \beta_k) + \boldsymbol{N}_{k'} - 1)} \times h_{u_{e'}, u_e}(t_e - t_{e'})$$

Here the first term is the transition probability from topic k' of the proposed parent event e' to this event's topic k. The second term is the probability of t_e being the time of this event e given the occurrence time $t_{e'}$ of the proposed parent event e'. As with topic identification, we see that evidence for the parent now comes from two different sources. When there is uncertainty about the parent based on the event time, common patterns of topic transitions between existing parent-child events helps in identifying the right parent.

The conditional probability of event e being a spontaneous event with no parent is given as:

$$P(z_e = 0 | E_t, \boldsymbol{z}_{\neg e}, \boldsymbol{\mu}) \propto \frac{(\gamma_k + \mathfrak{U}_{u_e,k} - 1)}{((\sum_{k=1}^{K} \gamma_k) + \boldsymbol{\mathfrak{U}}_{u_e} - 1)} \times \mu_{u_e}(t_e)$$

We limit the maximum number of parent candidates for an event to a parameter M (set to 100 in our experiments) and by limiting the maximum delay to D (set to 1 day).

c) Updating network strengths: For the network strength $W_{u,v}$, using a Gamma prior $Gamma(\alpha, \beta)$, the posterior distribution can be approximated as follows:

$$P(W_{u,v} = x \mid E^{(u,v)}_t, \boldsymbol{z}) \propto \quad x^{\alpha_1} \exp(-x\beta_1)$$

where $\alpha_1 = (N_{u,v} + \alpha - 1)$ and $\beta_1 = (N_u + \frac{1}{\beta})^{-1}$, and $N_{u,v}$ is the number of parent-child events pairs between nodes u and v, and N_u is the number of events at node u. This is again a Gamma distribution $Gamma(\alpha_1, \beta_1)$. Instead of sampling W_{uv}, we set it to be the mean of the corresponding Gamma distribution. Note that in each iteration, we update W_{uv} only for edges that have at least one parent-child influence.

Most edges have very few (typically just one) influence propagation event. This makes statistical estimation of their strengths infeasible. To get around this problem, we share parameters across edges. We group together edges that have the same value for the tuple (out-degree(source), in-degree(destination)). The intuition is that the influence of the edge (u, v) is determined uniquely by the the popularity (outdegree) of u and the number of different influencers (indegree) of v. We then pool data from all edges in a group and estimate a single connection strength for a group.

After convergence, the parameters $\zeta_{k,r}$ and $\mathcal{T}_{k,k'}$ that were integrated out are estimated using the samples.

IV. Experiments

In this section, we empirically validate the strengths of HMHP against competitive baselines over a large collection of real tweets as well as semi-synthetic data. We first discuss the baseline algorithms for comparison, the datasets on which we evaluate these algorithms, the tasks and the evaluation measures, and finally the experimental results.

A. Evaluated Models

Recall that our model captures network structure, textual content and timestamp of the posts / tweets, and identifies the topics and parents of the posts, in addition to reconstructing the network connection strengths. Considering this, we evaluate and compare performance for the following models:

- HMHP: This is the full version of our model with Gibbs sampling-based inference. It performs all reconstruction tasks mentioned above jointly, while assuming a single topic for a post and topical interaction patterns.
- HTM: This is the state-of-the-art model closest to ours that addresses the same tasks. The key modeling differences

are two-fold: HTM assumes a topical admixture for each document instead of a mixture, and secondly it models parent and child events to be 'close' in terms of their topical admixture. As a result, it cannot capture any sequential pattern in the topical content of cascades.

- HWK+Diag: This is a simplification of our model where the topic-topic interaction is restricted to be diagonal. In other words, each topic interacts only with itself. This model helps assess the value of topical interaction and serves as a crude approximation of HTM since the original did not scale for our larger datasets.
- HWK×LDA: This is motivated by the Network Hawkes model [12], which jointly models the network strengths and the event timestamps, and therefore performs parent assignment and user-user influence estimation jointly. However, it does not model the textual content of the events. Therefore, we augment it with an independent LDA mixture model [14] (LDAMM) to model the textual content. Thus, comparison against this model helps in analyzing the importance of doing topic, parent and network strength estimation jointly. The network reconstruction component of HWK×LDA is similar to the Network Hawkes model [12].

B. Datasets

We perform experiments on two datasets, that we name `Twitter` and `SemiSynth`. The `Twitter` dataset was created by crawling $7M$ seed nodes from Twitter for the months of March-April 2014. We restrict ourselves to $500K$ tweets corresponding to top 5K hashtags from the most prolific $1M$ users generated in a contiguous part of March 2014. For each tweet, we have the time stamp, creator-id and tweet-text. Note that gold-standard for the parameters or the event labels that we look to estimate is not available in this `Twitter` dataset. We do not know the true network connection strengths, event topics or cascade structure of the tweets. While retweet information is available, it is important to point out that retweets form a very small fraction of the parent-child relations that we hope to infer.

Since we need gold-standard labels to evaluate performance of the models, we additionally create a `SemiSynth` dataset using the generative process of our model while preserving statistics of the real data to the extent possible. From a sample of the `Twitter` data we retain the underlying set of nodes and the follower graph. Then we estimate all the parameters of our model (the base rate per user, user-user influence matrix, the topic distributions, resulting topical interactions and the user-topic preferences) from `Twitter` data. The document lengths are randomly drawn from $Poisson(7)$, since 7 was the average tweet length in the `Twitter` dataset. We finally generate 5 different samples of $1M$ events using our generative model. All empirical evaluations are averaged over the 5 generated samples (together termed as `SemiSynth`). Due to lack of space, we omit the details of parameter estimation for generating `SemiSynth` data.

C. Tasks and experimental results

We address three tasks. (1) *Reconstruction accuracy*, (2) *Generalization performance* and (3) *Discovery and analysis of topic interactions*. Now we describe the experimental setup for each of the tasks and then the results.

Table I: Cascade Recons. Accuracy and Recall @ top-K

	HMHP	HWK+Diag	HWK×LDA
Accuracy	0.581	0.362	0.370
Recall@1	0.595	0.373	0.380
Recall@3	0.778	0.584	0.589
Recall@5	0.838	0.674	0.678
Recall@7	0.87	0.73	0.733

Table II: Network Reconstruction Error (as fraction)

	HMHP	HWK+Diag	HWK×LDA
Mean APE	0.448	0.565	0.552
Median APE	0.255	0.283	0.287
Mean APE ($N_u \geq 100$)	0.398	0.520	0.496
Median APE ($N_u \geq 100$)	0.235	0.265	0.264

1) Reconstruction Accuracy: We address three reconstruction tasks: *parent identification* (`PI`), *network reconstruction* (`NR`), and *topic identification* (`TI`). Since the gold-standard values are not known for the `Twitter` data, we address these tasks only on the `SemiSynth` data. For `PI` (or cascade reconstruction), given a ranked list of predicted parents for each event, we measure accuracy and recall of the parent prediction. The accuracy metric is calculated as the ratio of number events for which the parent is identified correctly to the total number of events. For `NR`, we measure distance between estimated and true W_{uv} values. We report the error in terms of the median Average Percentage Error (APE), defined as $\sum_{u,v} \frac{|W_{uv}-\hat{W}_{uv}|}{W_{uv}}$, where, $\hat{W}_{uv}$ is the estimated W_{uv} value. For comparison with HTM, we use Total Absolute Error (TAE) (the sum of the absolute differences between the true and estimated W_{uv} values) which was used in the original paper. For `TI`, we take a (roughly balanced) sample of event pairs, and then measure (using precision/recall/F1) whether the model accurately predicts the event pairs to be from the same or different topics.

The model that is closest to ours is HTM [11]. However, the HTM assigns each event to a topic distribution instead of a single topic. Also, the available HTM code cannot scale to the size of our `SemiSynth` dataset. Therefore, we compared our model with HTM separately. We first present comparisons with the other baselines on the `SemiSynth` dataset. All the presented numbers are averaged over five different semi-synthetic datasets generated independently.

Table I presents the accuracy and the recall@k values for the task of predicting parents, and thereby the cascades. For all the algorithms, the recall improves significantly by the time the top-3 candidates are considered. For recall@3, HMHP performs about 32% better than both the other baselines, whereas recall@1, as well as the accuracy of HMHP are at least 57% better than corresponding number of the baselines.

For the `NR` task, Table II describes the various APE values for models HMHP, HWK+Diag, and HWK×LDA. Given that in our simulation, event generation was truncated by limiting the number of events, the last level of events contributes to negative bias affecting the APE, since the algorithms are still trying to assign children to these events. We report both the mean and the median errors for this task. The mean error for HMHP is about 18% lower than other baselines, and the median error is about 10% lower. We also present the results

Table III: Event Topic Identification (P/R/F1)

Topic	HMHP	HWK+Diag	HWK×LDA
Precision	0.893	0.123	0.781
Recall	0.746	0.367	0.752
F1	0.811	0.18	0.765

Table IV: Held-out Log-Likelihood for Tweet Content/Activation Time/Aggregate

#Topics	Log-Likelihood	HMHP	HWK+Diag	HWK×LDA
25	Content	-30499278	-33356945	-30532938
	Event Time	-4236958	-4042903	-4299630
	Total	-34736237	-37399849	-34832568
50	Content	-30141081	-33427354	-30089733
	Event Time	-4288438	-4510072	-4343571
	Total	-34429519	-37937426	-34433305
75	Content	-29860909	-33433922	-29861050
	Event Time	-4285293	-4510535	-4373736
	Total	-34146202	-37944457	-34234787

for the set of nodes where the number of generated children (N_u) is at least 100 (this is an easier task). Here also the mean APE for HMHP is at least 20% lower.

Table III presents the result of the TI task, measured by calculating the precision-recall over event pairs as defined earlier. The results show that HMHP performs much better than HWK+Diag and at least $5-6\%$ better than HWK×LDA. The above experiment confirms that joint inference of topics, parents, and network strengths by modeling topical interactions is more accurate than that ignoring topical interactions and also decoupled inference. While the trend is not surprising as the data was generated using our model, the numbers confirm that when topical interactions patterns are present in the data, our inference algorithm is able to detect and make use of these. Also, the improvement over the baselines for all tasks is both significant in magnitude as well as statistically, showing the importance of topical interactions and joint inference for reconstructions tasks.

We also conducted in depth comparison with HTM by considering synthetic datasets generated based on models suggested in the HTM paper. HMHP gives 40% lesser TAE for the NR task and at least 25% better accuracy for PI task for both synthetic and Arxiv datasets mentioned in HTM [11]. Due to lack of space, we could not include these results, which are available in [3].

2) Data fitting quality: We now evaluate goodness of fit for different models in the held out setting. All the models were trained on 500K events and held-out log-likelihood was calculated on the 500K events immediately following the training events. Since this does not require gold-standard labels, this can be evaluated on Twitter as well as SemiSynth data. Here we present results only for the Twitter data, since that is the harder task for our model. Table-IV shows the values of the log-likelihood for all the three models. We observe that the likelihood number for HMHP is roughly 5% better than the ones for both the HWK+Diag and HWK×LDA. We also observe the test likelihood of HMHP improves with the number of topics.

Just as for reconstruction accuracy on SemiSynth data, we see that modeling and detecting topical interaction patterns jointly with other tasks leads to significantly better generalization performance for real data. From this we may claim (of course without direct validation) that HMHP would perform more accurate reconstruction as well on Twitter data.

Table V: Sample hashtags from asymmetric topics pairs.

Parent topic hashtags	**Child topic hashtags**
steelers, browns, seahawks, fantasyfootball, nfl	mlb, orioles, rays, usmnt, redsox
idf, rnb, bds, gaza, israel	russia,iran,syria,crimea,ukraine
thewalkingdead, theamericans, onceuponatime, houseofcards, tvtag	arrow, agentsofshield, truedetective, longislandmedium, tvtag
egypt,gaza,israel,syria,iran	kiev, putin, russia, crimea, ukraine

3) Discovery and Analysis of topic interactions: The final task is to discover statistically significant topical interactions from textual information cascades and investigate what actionable insights can be drawn from such topical patterns. We stress that this task can only be performed using HMHP, and as such there is no baseline algorithm for this task. In previous experiments, we demonstrated that HMHP outperforms the baselines on both SemiSynth and real data. This lends credibility to the topical interactions discovered by HMHP model, even though there is no ground truth for these.

We show the usefulness of modeling topic-topic interaction by first providing anecdotes of sets of inter-topic relationships, then by using a personalized pagerank based method to discover related topics for any given topic. We discuss how this can lead to new ways of topic-specific user targeting and influence maximization.

a) Anecdotal parent-child topics: To identify interesting examples of topic-topic interaction, we ignore topic pairs that do not correspond to same underlying topic that was inappropriately partitioned into many finer-grained topics. Such a phenomenon would cause a block-diagonal structure in the topic-topic interaction matrix. We identify the most asymmetric topic pairs which cannot have come from topic splitting by sorting using $\mathcal{T}_{kk'} - \mathcal{T}_{k'k}$. We use their top-5 hashtags to illustrate the selected topic-pairs in Table V. In the first row, the parent and child hashtags are related to American football and baseball. These are different topics, but indicate that tweets related to a particular sport (football) trigger tweets about some other sport (baseball), perhaps due to common users participating in both. Table VI shows a few actual parent-child tweets from the selected asymmetric topic pairs. Note that these tweet pairs are not retweets. Also, though all these tweet pairs are clearly on related topics, often (like the first pair about *MH370*) they do not share any hashtags or even other significant words, which some naive strategy might use to detect relationships. Therefore, in both the cases, we see evidence of interesting topic interactions that are discovered by HMHP. Such discoveries can serve to improve downstreams tasks by getting a global view of ongoing discussions, get better estimates of viral topics etc.

b) Modeling Topic Drifts via random walks: One of the compelling uses of cascade modeling is in targeting a small set of participants in the network to encourage conversations (tweets) about topics of interest to an advertiser. We claim that our model, by capturing the topic-topic interactions, provides alternate topics or hashtags for such advertisers to exploit. To demonstrate this, we choose few of the topics and then run personalized pagerank [13] from these topic nodes on the Markov chain underlying the topic-topic interaction matrix. Table VII shows the results where each chosen topic in

Table VI: Example Parent-Child Tweets

Parent Tweet	Child Tweet
[#MASalert] Statement By Our Group CEO, Ahmad Jauhari Yahya on MH370 Incident. Released at 9.05am/8 Mar 2014 MY LT	Missing #MalaysiaAirlines flight carrying 227 passengers (including 2 infants) of 13 nationalities and 12 crew members.
Certain people are ruining their reputations tonight-really sad! #Oscars	I should host the #Oscars just to shake things up - this is not good!
Gellman:My definition of whistle blowing:are you shedding light on crucial decision that society should be making for itself. #snowden	Gellman we are living inside a one way mirror,they & big corporations know more and more about us and we know less about them #sxsw

Table VII: Example Related Topics using Personalized Pagerank (after emoving some top generic topics)

Source Topic Hashtags	Hashtags from top-3 Transitioned Topics
[ukraine, crimea, russia, putin, syria]	[utpol, raisethewage, cdnpoli, obamacare, aca], [irs, tcot, teaparty, gophatesvets, uniteblue], [worldbookday, amwriting, books, litfestlive]
[rhlaw, clinton, r4today, ecommerce, cadem14]	[irs, tcot, teaparty, gophatesvets, uniteblue], [soundcloud, hiphop, mastermind, nowplaying, music], [iahsbkb, nba, iubb, lakers, rockets]
[agentsofshield, arrow, tvtag, supernatural, chicagoland]	[idol, bbcan2, havesandhavenots, pll, thegamebet], [tvtag, houseofcards, agentsofshield, arrow, theamericans], [soundcloud, hiphop, mastermind, nowplaying, music],

first column, the second column contains top-3 topics on their personalized pagerank score, obtained after removing general topics with high authority-score. It is evident that the transitioned topics, given in the right column, are directly related to the start topic. For e.g., the first start topic is about Russia, Syria and Ukraine, which transitions into topics containing hashtags related to politics in US, offering us a hint about the connection between conversations on these topics in Twitter. Similarly, in the last row the start topic is about TV shows, and the transitioned topics also contain hashtags which are about some other prominent TV shows. One interpretation of this is that the average conversation that starts with any topic from the first column, can be expected to drift to one of the topics in the corresponding row.

V. Related Work

There has been a lot of work on network reconstruction based on the observations of event times [7], [17]— these models often ignore the content information of the events. Our work shows that such content information, when present, can profitably be used for a better estimation of the network strengths. [1] studies a related model in which activation times and some side information (e.g. tweet content) about cascades are observed, but not the set of users.

The Dirichlet Hawkes Process (DHP) [5] and the HTM [11] both extend Hawkes Processes to model textual content associated with events. However, neither captures topical interactions. The DHP lacks the notion of a unique parent for any event and also any notion of users or networks. Our model is most similar to the HTM. But while the HTM forces parent and child events to be topically close, it does not capture any parent-child topical patterns beyond this. One consequence is that no two events are topically identical, while we can group events according to assigned topics and, additionally, parent-child relations according to topic pairs. HTM's document model is an admixture of topics, which is more powerful in general, but in the context of short documents such as tweets, this complexity is not as necessary, as we demonstrate in our experiments.

Models for sequence data have blended Markovian dynamics with topic models [8], [9], [4], [2]. This thread, is focused on modeling richer sequential structure within a single document, but has not looked at cascades of events with associated time stamps and networked users. Our contribution is in merging this line of research with the modeling of information cascades.

VI. Conclusion

In summary, we propose a generative model for information diffusion cascades accounting for topical interactions, by coupling a Network Hawkes process with a Markov Chain over topics for diffusion paths. This enables us to fit real Twitter conversations better, and the use of topic interactions and collective inference using our model also leads to more accurate reconstruction of network strengths, diffusion paths and event topics. Using comparisons on a number of datasets, we show that our model outperforms the existing baselines on the standard metrics. On top of this, using our model, we are able to derive insights about topical interactions, which existing models cannot.

References

[1] N. Barbieri, G. Manco, and E. Ritacco. Survival factorization on diffusion networks. In *ECML-PKDD*, 2017.

[2] N. Barbieri, G. Manco, E. Ritacco, M. Carnuccio, and A. Bevacqua. Probabilistic topic models for sequence data. *Mach. Learn.*, 93(1):5–29, 2013.

[3] S. Bedathur, I. Bhattacharya, J. Choudhari, and A. Dasgupta. Discovering topical interactions in text-based cascades using hidden markov hawkes processes. In *arXiv*, 2018.

[4] L. Du, W. Buntine, H. Jin, and C. Chen. Sequential latent dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503, 2012.

[5] N. Du, M. Farajtabar, A. Ahmed, A. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *SIGKDD*, 2015.

[6] S. Gao, J. Ma, and Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In *WSDM*, 2015.

[7] M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(1):26–65, 2014.

[8] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *NIPS*, 2005.

[9] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden topic markov models. In *AISTATS*, 2007.

[10] A. Guille, H. Hacid, C. Favre, and D. Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2), 2013.

[11] X. He, T. Rekatsinas, J. Foulds, L. Getoor, and Y. Liu. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML*, 2015.

[12] S. Linderman and R. Adams. Discovering latent network structure in point process data. In *ICML*, 2014.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

[14] L. Rigouste. *Inference and evaluation of the multinomial mixture model for unsupervised text clustering*. PhD thesis, Télécom ParisTech, 2006.

[15] M. Rizoiu, Y. Lee, S. Mishra, and L. Xie. A tutorial on hawkes processes for events in social media. In *arXiv*, 2017.

[16] A. Simma. *Modeling Events in Time Using Cascades of Poisson Processes*. PhD thesis, 2010.

[17] S. Wang, X. Hu, P. Yu, and Z. Li. Mmrate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *SIGKDD*, 2014.

[18] Q. Zhao, M. Erdogdu, H. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.