HawkesTopic: A Joint Model for Network Inference and Topic Modeling from Text-Based Cascades

Xinran He[1] XINRANHE@USC.EDU
Theodoros Rekatsinas[2] THODREK@CS.UMD.EDU
James Foulds[3] JFOULDS@UCSC.EDU
Lise Getoor[3] GETOOR@SOE.UCSC.EDU
Yan Liu[1] YANLIU.CS@USC.EDU

[1]University of Southern California, [2]University of Maryland, College Park, [3]University of California, Santa Cruz

Abstract

Understanding the diffusion of information in social networks and social media requires modeling the text diffusion process. In this work, we develop the *HawkesTopic* model (HTM) for analyzing *text-based* cascades, such as "retweeting a post" or "publishing a follow-up blog post." HTM combines *Hawkes processes* and *topic modeling* to simultaneously reason about the information diffusion pathways and the topics characterizing the observed textual information. We show how to jointly infer them with a mean-field variational inference algorithm and validate our approach on both synthetic and real-world data sets, including a news media dataset for modeling information diffusion, and an ArXiv publication dataset for modeling scientific influence. The results show that HTM is significantly more accurate than several baselines for both tasks.

1. Introduction

There has been an increasing interest in understanding the processes and dynamics of *information diffusion* through networks and modeling the *influence* across the nodes of the underlying networks. Such processes play a fundamental role in a variety of domains, such as evaluating the effects of networks in marketing (Domingos & Richardson, 2001; Kempe et al., 2003; Leskovec et al., 2007; Wang et al., 2010), monitoring the spread of news, opinions, and scientific ideas via citation networks (Adar et al., 2004; Gruhl et al., 2004; Leskovec et al., 2005), and detecting the spread of erroneous information (Dong et al., 2009). Most prior work focuses on modeling the diffusion of information by solely exploiting the observed timestamps when different nodes in the network post (i.e., publicize) information. In such scenarios diffusion is modeled either using cascade models, for example, *NetInf* (Gomez-Rodriguez et al., 2012) and *NetRate* (Gomez-Rodriguez et al., 2011), or point process models such as *MMHP* (Yang & Zha, 2013) and *LowRankSparse* (Zhou et al., 2013). These techniques do not leverage textual information and focus mainly on *context-agnostic* tasks such as product purchasing.

However, *text-based cascades* are abundant in a variety of social platforms, ranging from well-established social networking websites such as Facebook, Google+, and Twitter, to increasingly popular social media websites such as Reddit, Pinterest, and Tumblr. Moreover, a growing number of platforms such as GDELT and EventRegistry[1] extract and analyze textual information from diverse news data sources which often borrow content from each other or influence each other (Dong et al., 2010).

Text is, in many cases, the medium by which information is propagated, making it particularly salient for inferring information diffusion. Models that are based on observed timestamps have been shown to become more effective at discovering topic-dependent transmission rates or diffusion processes when combined with the textual information associated with the information propagation (Du et al., 2013; Wang et al., 2014). Nevertheless, this line of work assumes that either the topics associated with the diffusion process are specified in advance or that the influence paths are fully observed. It is easy to see that due to these assumptions, the aforementioned models are not applicable in many scenarios, such as discovering influence relationships between news data sources or users of social media, and detecting information diffusion paths.

In this paper, we focus on the problem of inferring the diffusion of information *together* with the topics characterizing the information. We assume that only the textual information and timestamp of posted information is known. We

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

[1]See gdeltproject.org and eventregistry.org.

do not require prior knowledge of either the structure of the network nor the topics, as in previous approaches. We introduce a novel framework that combines topic models and the Hawkes process (Liniger, 2009) under a unified model referred to as the *HawkesTopic* model (HTM). HTM uses the *Marked Multivariate Hawkes Process* (Liniger, 2009) to model the diffusion of information and simultaneously discover the hidden topics in the textual information.

Specifically, our model captures the posting of information from different nodes of the hidden network as *events* of a Hawkes process. The *mutually exciting nature* (Yang & Zha, 2013) of a Hawkes process, i.e., the fact that an event can trigger future events, is a natural fit for modeling the propagation of information in domains such as the ones mentioned above. To address the limitation that the thematic content of the available textual information is unknown, HTM builds upon the Correlated Topic Model (CTM) (Blei & Lafferty, 2006a) and unifies it with a Hawkes process to discover the underlying influence across information postings, and thus, the hidden influence network. We derive a joint variational inference algorithm based on the mean-field approximation to discover both the diffusion path of information and its thematic content.

2. HawkesTopic Model

We consider text content cascades among a set of nodes $V = \{1, 2, \ldots, |V|\}$. Each node v can represent a news domain propagating news stories, a researcher publishing scientific papers or a user in a social network (e.g., Facebook). The nodes can influence each other via a *hidden* diffusion network G. We observe a sequence of posting activities $D = \{a_i | i = 1, 2, \ldots, |D|\}$, e.g., a series of news article or research paper publications, or a series of user postings on Facebook. We denote each posting activity as a tuple $a_i = (t_i, v_i, X_i)$. This means that node v_i posts document X_i at time t_i. Documents $\{X_1, \ldots, X_{|D|}\}$ are represented as a bag-of-words with vocabulary size W.

Given this input, our goal is to infer the hidden diffusion network G and the topics characterizing the observed textual information. We adopt a model that jointly reasons about the posting time and the content of documents to: (1) accurately infer the hidden diffusion network structure; and (2) track the thematic content of documents as they propagate through the diffusion network. Next, we discuss the two components of our HawkesTopic model in detail. The notation of our model is summarized in Table 1.

Table 1. Notation used in this paper.

Notation	Definition
V	Set of nodes
E	Set of events
t_e	Posting time of event e
v_e	Node who carries out event e
X_e	Document of event e
$P_e = (P_{e,0}, \{P_{e,e'}\}_{e' \in E})$	Parent indicator of event e
η_e	Topics parameters of document X_e
$\boldsymbol{\beta} = \beta_{1:K}$	Collection of all topics
$\lambda_v(t)$	Intensity process of node v
μ_v	Base intensity of node v
$\kappa_e(t, v)$	Impulse response of event e
$\boldsymbol{A} = \{A_{u,w}\}$	Node influence matrix

2.1. Modeling the posting time

The first component of our framework models the node posting times via the *Multivariate Hawkes Process* (MHP) (Liniger, 2009). Document modeling is described in Section 2.2. In the MHP model, each node v is associated with a point process $N_v(t)$, where $N_v(t)$ is the number of posting activities of node v in the time interval $[0, t]$ (assuming the process starts at time 0). Following the traditional notation of point processes, we define each posting activity $a_i = (t_i, v_i, X_i)$ as an event $e_i = (t_i, v_i)$ of the process associated with node v_i. We use $E = \{e_1, \ldots, e_{|D|}\}$ to denote all events. Figure 1 provides an example of the MHP model. The sources correspond to two users v_1 and v_2 that publish alternatively on the same subject. Here, $e_{2,1}$ is a response to $e_{1,1}$ published by v_1, and $e_{1,2}$ is v_2's response to v_1's document $e_{2,1}$.

In the MHP model, the occurrence of an event can lead to a chain of future events. For example, a seminal paper may start a new field of study generating a large amount of follow-up work. The *mutually exciting* property makes the MHP model a perfect fit for cascades of posting activities and can be captured effectively via the intensity process $\lambda_v(t|H)$ defined as:

$$\lambda_v(t|H) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N_v(t + \Delta t) - N_v(t)|H_t)]}{\Delta t}$$

where H_t is the history of all events before time t. Intuitively, $\lambda_v(t|H)\Delta t$ corresponds to the expected number of events for node v occurring in a small time interval $[t, t + \Delta t]$. For MHP, the intensity process $\lambda_v(t|H)$ takes the form:

$$\lambda_v(t) = \mu_v + \sum_{e:t_e<t} \kappa_e(t, v),$$

where μ_v is the base intensity of the process, while each previous event e adds a nonnegative impulse response $\kappa_e(t, v)$ to the intensity, increasing the likelihood of future events. We decompose the impulse response $\kappa_e(t, v)$ into two factors that capture both the influence between nodes and the temporal aspect of diffusion:

$$\kappa_e(t, v) = A_{v_e,v} f_\Delta(t - t_e)\,.$$

Here, $\boldsymbol{A} = \{A_{u,w}\}$ is a non-negative matrix modeling the strength of influence between nodes. $A_{u,w}$ is the expected number of events that a single event at node u can trigger in the process of node w. Nonzero $A_{u,w}$ entries correspond to edges (u, w) in the diffusion network. The larger $A_{u,w}$, the stronger the influence node u has on node w. Also, $f_\Delta(\cdot)$ is the probability density function for the *delay distribution*. It captures how long it takes for an event at one node to

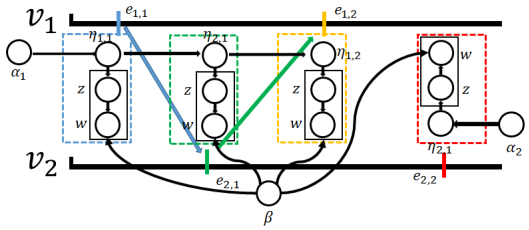


Figure 1. Graphical representation of HawkesTopic model.

influence other nodes (i.e., trigger events at other nodes) and how long this influence will last.

A Hawkes process can be treated as a clustered Poisson process (Simma, 2010) where each event triggers a homogeneous Poisson process with intensity $\kappa_e(t, v)$. The process $N_v(t)$ of node v is a superposition of a homogeneous Poisson process $\mathcal{P}_v$ with intensity μ_v , as base process, and the Poisson processes triggered by previous events.

Viewing the Hawkes process from this perspective has two advantages. First, it provides a method to generate the events of different processes in a breadth first order (Simma, 2010). The main idea is to first generate all events corresponding to the base process of each node, referring to them as *events at level zero* and then generate events at level ℓ from processes triggered by events at level $\ell - 1$. This process is repeated until all processes triggered by events at level L are empty. This construction dictates an explicit *parent relationship* denoted by an indicator vector $P_e = (P_{e,0}, \{P_{e,e'}\}_{e' \in E})$ for each event e. If event e is generated from the process triggered by previous event e', we say that e' is the parent of e and denote it by $P_{e,e'} = 1$. In short, we say that event e' triggers event e. Otherwise, if event e is generated from the base process, we say that it has no parent and denote it as $P_{e,0} = 1$. Equivalently, we say that node v_e carries out event e spontaneously. The colored arrows in Figure 1 depict the parent relationship. In this example, events $e_{1,1}$ and $e_{2,2}$ have no parent, while event $e_{1,1}$ is the parent of event $e_{2,1}$, which itself is the parent of event $e_{1,2}$. The parent relationship is essential to model the evolution of the content information, since our model should capture the intuition that the document associated with an event is supposed to be similar to the document of its parent.

2.2. Modeling the documents

One approach for reasoning about the information of documents is to generalize the MHP to the *Marked Multivariate Hawkes Process* (MMHP) (Liniger, 2009) by treating the words of documents associated events as the *marks* associated with those events. In the MMHP model, events are extended to triples $e = (t_e, v_e, x_e)$ where the mark value x_e is an additional label characterizing the content.

However, this naive approach suffers from two major drawbacks: (1) using words as marks leads to noisy representations due to polysemy and synonyms; more importantly, (2) in the traditional MMHP model, the mark value depends only on the time of the event and is not affected by what triggers the event (Liniger, 2009). This assumption is acknowledged in (Yang & Zha, 2013). The documents as marks associated with the events are just drawn independently from the same language model without considering what is the source of the influence. In other word, if a user posts something influenced by the post of her friend, then the content of the user's post is independent of the content of her friend's post. We can see that this assumption is unrealistic in many real-word scenarios including social or online news media, as posts of users that influence each other (e.g., they are friends) should exhibit dependencies.

Topics as Marks: To overcome the first disadvantage, we propose using the *topics* of the event documents as marks in our HawkesTopic model. Topics, as an abstraction of the actual words, provide a less noisy and more succinct representation of the documents' content.

Assuming a fixed set of topics $\boldsymbol{\beta} = \beta_{1:K}$ for all documents, we use a topic vector η_e to denote the topics in document X_e associated with event e. We assume that the actual words of the document are generated similar to the Correlated Topic Model (CTM) (Blei & Lafferty, 2006a). The generative process for a document X_e with N_e words is as follows: Let $\pi(\cdot)$ be the softmax function $\pi_k(\eta_e) = \frac{\exp(\eta_{e,k})}{\sum_j \exp(\eta_{e,j})}$ and $\beta_{1:K}$ be the discrete distributions over words characterizing each of the K topics.

- For $n = 1 \ldots, N_e$:
  1. Draw topic assignment $z_{e,n} \sim \text{Discrete}(\pi(\eta_e))$.
  2. Draw word $x_{e,n} \sim \text{Discrete}(\beta_{z_{e,n}})$.

We choose the logistic normal distribution for variable $z_{e,n}$ as it provides more flexibility than the multinomial distribution in the LDA model. The logistic normal distribution does not constrain the document-topic parameters to the probability simplex, making it a better representation for modeling the dynamics of topic diffusion.

Diffusion of Topics: To overcome the second disadvantage, i.e., modeling the dependencies across marks of events that influence each other, our HTM model *explicitly reasons* about the diffusion of topics across events that influence each other. We distinguish between two types of events: (i) those occurring *spontaneously* and (ii) those *triggered* by previous events. For example, in Figure 1 events $e_{1,1}, e_{2,2}$ belong to the first type, while the remaining events belong to the second type. We assume each node has a prior of interests in different topics. For example, one Facebook user may be interested in sports and politics while the other in music and movies. The content of documents corresponding to spontaneous events are determined by the topic prior of the node. If an event is triggered by another event, its document should be similar to the document of the triggering event. This suggests that the content

of a user's post, influenced by her friend's previous post, should have similar content to her friend's post.

We use the parent relationship generated in event time modeling to realize the above intuition. Let α_v be the parameter describing the topic prior for node v. The topics of spontaneously posted documents for node v are generated from a Gaussian distribution with mean α_v, i.e., $\eta_e \sim N(\alpha_v, \sigma^2 I)$. The topics of triggered events η_e are also generated from a Gaussian distribution, but with mean parameter $\eta_{\text{parent[e]}}$, i.e., $\eta_e \sim N(\eta_{\text{parent[e]}}, \sigma^2 I)$, where parent[e] is the event triggering event e. To promote simplicity, our model uses an isometric Gaussian distribution. In example of Figure 1, the topics distribution $\eta_{1,1}$ is drawn from a Gaussian distribution with mean α_1. Similarly, the topics distribution $\eta_{2,2}$ is drawn from a Gaussian distribution with mean α_2. On the other hand, the topics associated with event $e_{2,1}$ are influenced by event $e_{1,1}$ as that is the one triggering it. More specifically, we draw $\eta_{2,1}$ from a Gaussian distribution with mean $\eta_{1,1}$.

2.3. Summary and Discussion

We summarize the generative process of our model and discuss how it compares against existing models. The notation of our model is summarized in Table 1. The generative process of our model is:

1. Generate all the events and the event times via the Multivariate Hawkes Process, as in Section 2.1.
2. For each topic k: draw $\beta_k \sim Dir(\alpha)$.
3. For each event e of node v:
   (a) If e is a spontaneous event: $\eta_e \sim N(\alpha_v, \sigma^2 I)$. Otherwise $\eta_e \sim N(\eta_{\text{parent[e]}}, \sigma^2 I)$.
   (b) Generate document length $N_e \sim Possion(\lambda)$.
   (c) For each word n:
   $z_{e,n} \sim \text{Discrete}(\pi(\eta_e)), x_{e,n} \sim \text{Discrete}(\beta_{z_{e,n}})$.

Our model generalizes several existing models. If we ignore the event documents, our model is equivalent to the traditional MHP model. To the other extreme, if we only consider the documents associated with spontaneous events of node v, our model reduces to the CTM (see Section 1) with hyperparameter $(\alpha_v, \sigma^2 I)$. HTM further models the contents of triggered events using the diffusion of topics. We also considered other alternatives for modeling the diffusion of documents. For example, it can be modeled via controlling the parameters that determine the generation of topics as in the Dynamic Topic Model (Blei & Lafferty, 2006b).

We choose the current approach as it is more robust in the presence of limited influence information. As each event may only trigger a limited number of events, there is not enough information to recover the influence paths if document diffusion is modeled via the topic hyperparameters. Otherwise, the influence can be also modeled on the word level (Dietz et al., 2007). Our approach yields a simpler model since documents are utilized to recover the influence pathways.

3. Inference

Exact inference for the HawkesTopic model is clearly intractable. Thus, we derive a joint variational inference algorithm based on the mean-field approximation. We apply the full mean-field approximation for the posteriors distribution $P(\boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{P} | E, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{A}, \boldsymbol{\mu})$ as

$$Q(\boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{P}) = \prod_{e \in E} \left[q(\eta_e | \hat{\eta}_e) q(P_e | r_e) \prod_{n=1}^{N_e} q(z_{e,n} | \phi_{e,n}) \right].$$

Since the Correlated Topic Model is a building block in our HawkesTopic model, our model is not conjugate. We adopt the Laplace Variational Inference method in (Wang & Blei, 2013) to handle the non-conjugate variable $q(\boldsymbol{\eta})$. The variational distribution for η_e is assumed to be a Gaussian distribution with its mean as the parameter to infer, $\eta_e \sim N(\hat{\eta}_e, \hat{\sigma}^2 I)$. The choice of using the same simple covariance matrix is to limit the complexity of our model. The variational distributions for the remaining variables are: $z_{e,n} \sim \text{Discrete}(\phi_{e,n})$ and $P_e \sim \text{Discrete}(r_e)$.

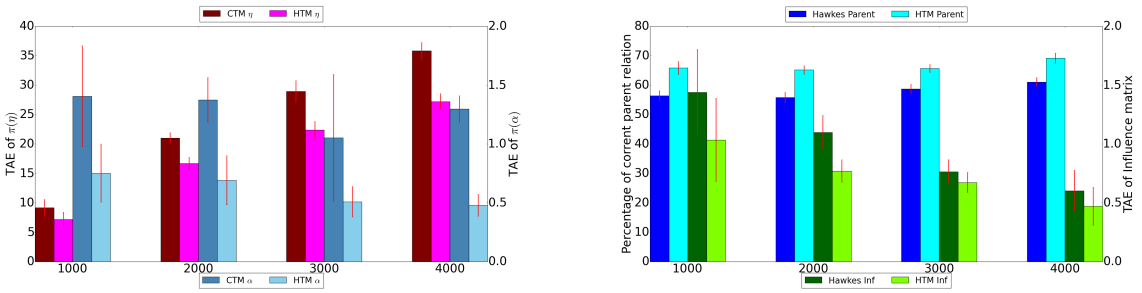
Under the standard variational theory, the inference task becomes minimizing the KL divergence between $Q(\boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{P})$ and $P(\boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{P} | E, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{A}, \boldsymbol{\mu})$. This is equivalent to maximizing a lower bound $\mathcal{L}(Q)$ on the log marginal likelihood. The full expression of the complete likelihood $P(E, \boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{P} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{A}, \boldsymbol{\mu})$ and the lower bound $\mathcal{L}(Q)$ are included in the supplementary material due to space constraints.

We only present the update for the variational distribution of the parent relationship $q(\boldsymbol{P})$ as it is unique in our model and depends on both time and content information. The update for other variational distributions and model parameters are included in the supplementary material.

Let $f_\Delta(\Delta_t)$ be the pdf for the delay distribution and $f_N(x | \mu, \Sigma)$ be the pdf for the Gaussian distribution with mean μ and covariance matrix Σ. The derivative of $\mathcal{L}(Q)$ with respect to r_e gives the following update equations:

$$r_{e,0} \propto \mu_{v_e} f_N(\hat{\eta}_e | \alpha_{v_e}, \hat{\sigma}^2 I)$$
$$r_{e,e'} \propto A_{v_{e'}, v_e} f_N(\hat{\eta}_e | \hat{\eta}_{e'}, \hat{\sigma}^2 I) f_\Delta(t_e - t_{e'}).$$

Intuitively, we combine three aspects in our joint HawkesTopic model to decide the parent relationship for each event: (i) $A_{v_{e'}, v_e}$ captures the influence between nodes, (ii) $f_N(\hat{\eta}_e | \hat{\eta}_{e'}, \hat{\sigma}^2 I)$ considers the similarity between event documents, and (iii) $f_\Delta(t_e - t_{e'})$ models the proximity of events in time. In contrast, the traditional MHP model uses only the time proximity and node influences to determine an event's parent.



(a) Performance of text modeling

(b) Performance of network inference

Figure 2. Results on synthetic datasets.

4. Empirical Evaluation

We present an empirical evaluation of the HawkesTopic model (HTM). The main questions we seek to address are: (1) how effective HTM is at inferring diffusion networks and (2) how well HTM can detect the topics associated with the event documents and their diffusion.

We empirically study these questions using both real and synthetic datasets. First, we apply HTM on synthetically generated data. Since the true diffusion network and topics are controlled, this set of experiments serves as proof of concept for the variational inference algorithm introduced in Section 3. Then, we provide an extensive evaluation of HTM on diverse real-world datasets and compare its performance against several baselines showing that HTM achieves superior performance in both tasks.

4.1. Synthetic Data

We first evaluate HTM on synthetic datasets and focus mainly on our variational inference algorithm.

Data generation: We generate a collection of datasets following the generative model assumed by HTM with a circular diffusion network G consisting of five nodes. We set the number of topics K to five. The specification of the node topic priors, the node influence matrix and other parameters are provided in the supplementary material. We vary the observation window length from 1000 to 4000. For each length, we generate five different datasets and report the mean and standard deviation for the evaluation measures below.

Results: First, we compare the true values of the topic distribution parameters η_e and node topic prior parameters α_v with their inferred equivalents $\hat{\eta}_e$ and $\hat{\alpha}_v$ respectively. The total absolute error (TAE) for the two parameters is computed as: $\text{TAE}(\pi(\eta_e)) = \sum_{e \in E} |\pi(\eta_e) - \pi(\hat{\eta}_e)|_1, \text{TAE}(\pi(\alpha_v)) = \sum_{v \in V} |\pi(\alpha_v) - \pi(\hat{\alpha}_v)|_1$. The corresponding errors are shown in Figure 2(a) together with the performance of the Correlated Topic Model (CTM) (Blei & Lafferty, 2006a). Our HTM exhibits improved performance yielding an error-reduction of up to 85% for α_v and up to 25% for η_e. The TAE corresponding to $\boldsymbol{\eta}$ increases for larger window length as the number of observed events increases, while the error of $\boldsymbol{\alpha}$ is rather stable as it is independent of the number of events.

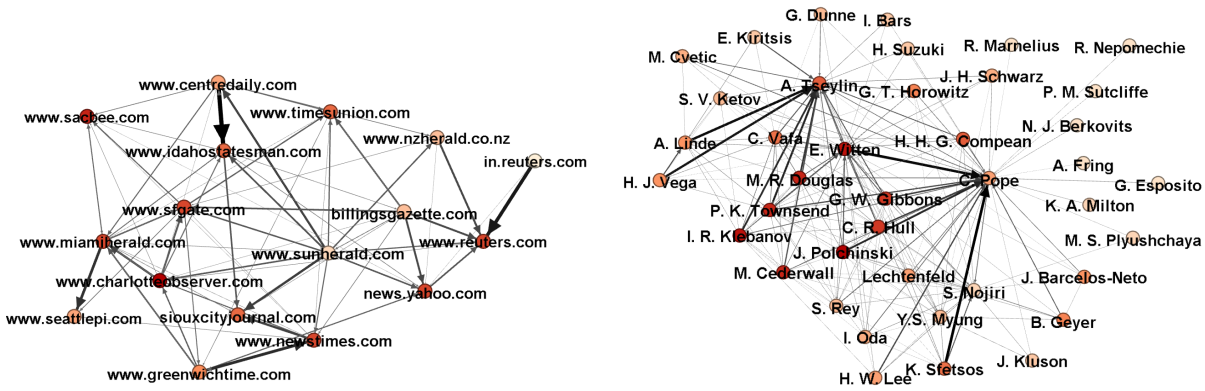
Next, we evaluate HTM at inferring the structure of the underlying diffusion network. We measure the accuracy of the inferred network using two metrics: (i) the percentage of correctly identified parent relations for the observed events, and (ii) the sum of absolute differences between the true node influence matrix A and the estimated matrix $\hat{A}$. The results are shown in Figure 2(b). We compare HTM against a Hawkes process model that does not consider the available textual information. Our HTM yields an increased accuracy of around 19% at identifying the event parent relationships and a decreased error of up to 28% for inferring the overall influence matrix. The decreasing error trend for the latter is due to the increased number of observed events for larger window length.

4.2. Real Data

We further evaluate HTM on two diverse real-world datasets. The first dataset corresponds to articles from news media over a time window of four months extracted from EventRegistry,[2] an online aggregator of news articles, while the second corresponds to papers published in ArXiv over a period of 12 years. We apply HTM on these seeking to: (i) identify the hidden topics associated with the documents in each dataset and (ii) infer the hidden diffusion network of opinions and ideas respectively.

EventRegistry Dataset: We collected news articles from EventRegistry by crawling all articles with keyword "Ebola" from 2014/07/01 to 2014/11/01. The dataset contains 9180 articles from 330 distinct news media sites. News media sites are treated as nodes in the diffusion network, and published articles as events in our model. We preprocessed the articles to remove stop words and words that appear less than ten times. Since the true diffusion network is not available, we use the the available *copying information* across news media, i.e., identical news articles published in multiple sites, to approximate the true diffusion network. More

[2] eventregistry.org



(a) EventRegistry (b) Arxiv

Figure 3. Inferred diffusion network from the EventRegistry and Arxiv datasets. The colors of nodes represent the out-degree of the source. (the darker the color, the higher the out-degree) The edge width represents the strength of influence.

precisely, if one article appears in multiple media sites, we consider the site that publishes the article first as the true *source* and add an edge to all sites who publish the article at a later time point. We use the three largest connected components in the induced graph as three separate datasets in our experiments. To extract the delay distribution for the Hawkes processes in our model, we fit an empirical delay distribution based on the delay times observed for duplicate articles.

ArXiv Dataset: We also use the ArXiv high-energy physics theory citation network data from Stanford's SNAP.[3] The dataset includes all papers published in ArXiv high-energy physics theory section from 1992 to 2003. We treat each author as a node and each publication as an event. For each paper, we use its abstract instead of the full paper as the document associated with the event. We use the articles of top 50/100/200 authors in terms of number of publications as our datasets. The citation network is used as ground truth to evaluate the quality of the inferred diffusion network. Similarly to EventRegistry, we fit an empirical delay distribution for the Hawkes processes based on the observed citation delays.

Algorithms: We compare the topic modeling and network inference capabilities of the following algorithms:

- **HTM:** Our HawkesTopic model. We set the number of topics K to 50, except that we use $K = 100$ for the ArXiv dataset with 200 authors as it contains more documents. We normalize the observation interval to a time window of length 5000 and fix the base intensity for all Hawkes processes to 0.02. The variance $\hat{\sigma}^2$ for topic diffusion is set to 0.5.
- **LDA:** Latent Dirichlet allocation with collapsed Gibbs sampling. The hyper-parameter α for the document topic distributions is set to 0.1 and the hyper-parameter for the topic word distributions is set to 0.02. The number of topics is set as in HTM. This is a baseline against the topic modeling component of HTM.
- **CTM:** Correlated topic model with variational inference. CTM serves as another baseline for the topic modeling of HTM.
- **Hawkes:** Network inference based on Hawkes processes considering only time using the same empirical delay distribution as in HTM and setting the base intensity same as in HTM. This is a baseline against the diffusion network inference component of HTM.
- **Hawkes-LDA:** Hawkes-LDA is a two-step approach that first infers the topics of each document with LDA and then uses those as marks for each event in the Hawkes processes. We compare this algorithm with HTM in terms of network inference accuracy.
- **Hawkes-CTM:** Similar to Hawkes-LDA with CTM being used instead.

Hawkes-LDA and Hawkes-CTM are the equivalent point process version of the TopicCascade algorithm (Du et al., 2013), a state-of-the-art baseline for inferring the diffusion network with textual information (Section 5).

Evaluation Metrics: We compare HTM to baseline methods both with respect to the quality of the discovered topics as well as the accuracy of the inferred diffusion network. To measure the quality of the discovered topics, we use the document completion likelihood (Wallach et al., 2009) via sampling for all algorithms. We use the area under the ROC curve (AUC) to evaluate the accuracy of network inference for all algorithms.

Results: The results for the three EventRegistry datasets with respect to the topic modeling performance and network inference performance are shown in Table 2 and Table 3. We see that our HTM outperforms the baseline algorithms in both text modeling and network inference for

[3] snap.stanford.edu/data/cit-HepTh.html

Table 2. EventRegistry: text modeling result (log document completion likelihood)

	LDA	CTM	HTM
Cmp. 1	-42945.60	-42458.89	**-42325.16**
Cmp. 2	-22558.75	-22181.76	-22164.05
Cmp. 3	-17574.70	-17574.30	-17571.56

Table 3. EventRegistry: network inference result (AUC)

	Hawkes	Hawkes-LDA	Hawkes-CTM	HTM
Cmp. 1	0.622	0.669	0.673	**0.697**
Cmp. 2	0.670	0.704	0.716	**0.730**
Cmp. 3	0.666	0.665	0.669	**0.700**

all three. While the improvement is marginal, our algorithm consistently performs better compared to baselines. We conjecture that the low AUC near 0.7 for all algorithms is due to the noisy ground truth network. Moreover, as promptness is essential for news sites, time information plays a more important role in this diffusion scenario. This explains the fact that HTM has similar performance with only near 10% improvement over Hawkes.

In Figure 3(a), we visualize the diffusion network for the third component. From the graph, we can clearly see that some news sites, like local news papers or local editions of Reuters, are the early bird in reporting stories. For example, the three red nodes, "sunherald.com", "miamiherald.com" and "billingsgazette.com" correspond to local news papers while "in.reuters.com" corresponds to the Indian edition of Reuters. On the other hand, bigger news agencies, such as "reuters.com" are strongly influenced by other sites. This is to be expected, as it is common for news agencies to gather reports from local papers and redistribute them to other news portals.[4]

The results for the ArXiv datasets are shown in Table 4 and Table 5. HTM consistently performs better than CTM and LDA in text modeling. This indicates that HTM discovers topics of higher quality by utilizing the cascade of information. In terms of network inference, our HTM model achieves more than 40% improvement in the accuracy compared to the Hawkes process by incorporating the textual information associated with each event. This result validates our claims that in many domains, timing information alone is not sufficient to infer the diffusion network. Moreover, our method outperforms the strong baselines Hawkes-LDA and Hawkes-CTM, suggesting that joint modeling of topics and information cascades is necessary and the information of diffusion pathways and the content information can benefit from each other vastly. The performance drops as the number of nodes increases since the dataset for top 100 authors has a limited number of publications. We believe that the reason is the increasing sparsity which makes the inference problem harder. Additionally, we carry out experiments on different observation time lengths on the ArXiv dataset with the top 50 authors. Namely, we train

Table 4. Arxiv: text modeling result (log document completion likelihood)

	LDA	CTM	HTM
Top 50	-11074.36	-10769.11	**-10708.96**
Top 100	-15711.53	-15477.24	**-15252.47**
Top 200	-27757.71	-27629.87	**-27443.41**

Table 5. Arxiv: network inference result (AUC)

	Hawkes	Hawkes-LDA	Hawkes-CTM	HTM
Top 50	0.594	0.656	0.645	**0.807**
Top 100	0.588	0.589	0.614	**0.687**
Top 200	0.618	0.630	0.629	**0.659**

our models using the papers published in the first three, six, and nine years, and the complete data set. The AUC of network inference accuracy is shown in Table 6. The results show that our model is capable of inferring the diffusion network accurately with only limited observations.

Figure 3(b) shows the hidden network among the top 50 authors. From the figure, we can see that the diffusion network has a core-peripheral structure. Influential authors such as Edward Witten, Michael R. Douglas, Joseph Polchinski almost form a clique in middle left of Figure 3. The common characteristics of these authors is that they do not publish the most, however, on average, each of their paper receives the largest number of citations. For example, Edward Witten has published 397 papers but has received more than 40000 citations. As another example, Joseph Polchinski has received near 9000 citations with only 190 publications. They serve as the core of the influence network, suggesting they may be the innovators in their corresponding fields. Influenced by the core authors, researchers such as Christopher Pope and Arkady Tseylin with intermediate number of both in-coming and out-going edges, further pass the influence to other authors. Overall, their works also receive a lot of citations, however, they publish more papers than the core authors with less average citations for each paper. For example, Christopher Pope received 6898 citation with 563 papers. This suggests that they can be considered as the mediator in the diffusion networks. Most of other authors, lying in the outside part in Figure 3, have few out-going edges, suggesting them perform more as the receiver of the new scientific ideas.

Besides inferring the influence relationship between authors, our model is also able to discover the research topics of the authors accurately. We list the inferred top-three topics for two authors together with the top-three words in each topic in Table 4.2. For HTM, we simply select the topics with largest value in $\hat{\alpha}_v$. For LDA and CTM model, we average over the topics of the papers published by the author. We compare the learnt topics to the research interests listed by the authors in their website. One of Andrei Linde's major research areas is the study of inflation.[5] Only our HTM model discovers it among the top-three topics of

[4] en.wikipedia.org/wiki/News_agency#Commercial_services

[5] physics.stanford.edu/people/faculty/andrei-linde

Table 7. Inferred topics for authors Andrei Linde and Arkady Tseytlin under LDA, CTM and our HTM model.

Author	LDA	CTM	HTM
Andrei Linde	black, hole, holes	black, holes, entropy	black, holes, hole
	supersymmetry, supersymmetric, solutions	supersymmetric, supersymmetry, superspace	universe, inflation, may
linde@physics.stanford.edu	universe, cosmological, cosmology	metrics, holonomy, spaces	supersymmetry, supersymmetric, breaking
Arkady Tseytlin	magnetic, field, conformal	solutions, solution, x	string, theory, type
	type, iib, theory	action, effective, background	action, actions, duality
a.tseytlin@ic.ac.uk	action, superstring, actions	type, iib, iia	bound, configurations, states

Table 6. Arxiv with different observation time length.

	1992-1995	1992-1998	1992-2001	1992-2004
AUC	0.614	0.747	0.789	0.807

the authors. LDA and CTM fail completely in discovering this topic. Arkady Tseylin reports string theory as his main research area in his webpage.[6] HTM successfully lists the string theory topic first, while CTM and LDA both leave this topic out of the top-three topics of the author. Our model accurately detects the topics because it can distinguish between spontaneous and triggered events. It infers authors preferences based only on spontaneous publications, while baseline models infer those using all publications.

5. Related Work

The prior work related to the techniques proposed in this paper can be placed mainly in three categories; we describe each of them in turn:

Diffusion Network Inference: There has been a significant amount of work on inferring the information diffusion network where either *cascade models* (Gomez-Rodriguez et al., 2012; 2011; Praneeth & Sujay, 2012) or point processes (Yang & Zha, 2013; Zhou et al., 2013) are used. Both approaches infer the diffusion networks utilizing only the observed time when nodes post or publicize the information. Finally, recent work has considered inferring heterogeneous diffusion networks (Du et al., 2012). While effective for context agnostic tasks (e.g., product adoption), these techniques fail to capture the complex context interdependencies.

Text-Content Cascades: While most of the previous work utilizes only the timing information to infer the diffusion networks, a different line of work has considered analyzing the available textual information and use *text-based cascades* (Dietz et al., 2007; Foulds & Smyth, 2013; Du et al., 2013). However, the work by Dietz et al. and Foulds et al. assumes that the influence paths are known and Du et al. assume that the topics characterizing the information are given in advance. Our proposed approach is different in that it does not make any restricting assumption on knowing the underlying diffusion network or the information topics in advance, thus being applicable in domains like news media where the underlying influence network is unknown and the contents vary significantly over time.

Diffusion Networks and Text-Content Cascades: Finally, a recent line of work focuses on joint modeling of diffusion networks and text-based cascades (Yang & Zha, 2013; Liu et al., 2010). Liu et al. extend the basic text-based cascades model in (Dietz et al., 2007) such that the diffusion paths also need to be inferred. However, the proposed approach is agnostic to time. The most relevant model to HTM is the MMHP model proposed in (Yang & Zha, 2013). Nevertheless, HTM is fundamentally different in the following aspects: (1) MMHP utilizes the textual information to cluster activations into different cascades, while, HTM leverages text to improve the prediction of a single cascade, and vice-versa, by modeling the evolution of textual information and event times jointly. (2) MMHP uses a simple language model and assumes that documents are drawn independently without considering the source of the influence. Instead, HTM models the evolution of textual information with CTM through the cascade of topics, which is essential in text diffusion processes. To our knowledge, there are only two papers that combine Hawkes processes and topic modeling (Li et al., 2014; Guo et al., 2015). Li et al. use the model to identify and label search tasks, while Guo et al. focus on studying conversational influence. In contrast, we combine the above for modeling text-based cascades.

6. Conclusion

In this paper, we studied the problem of analyzing text-based cascades and introduced a novel model, HawkesTopic, that combines Hawkes processes with correlated topic models to jointly infer the topics of available textual information and the information diffusion pathways. Since the inference task at hand is not tractable, we introduced a new variational inference algorithm. Our new model exploits the diffusion of topics to infer more accurate diffusion network. Our experimental results show that our techniques exhibit significant accuracy improvements when inferring the hidden structure of the diffusion network and are capable of discovering higher quality topics than several state-of-the-art baselines.

[6] www.imperial.ac.uk/people/a.tseytlin

Acknowledgments

This work was supported by NSF grant IIS1218488 and IIS1254206, IARPA via DoI/NBC contract number D12PC00337 and DARPA SMISC program with agreement number W911NF-12-1-0034. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DARPA, DoI/NBC, or the U.S. Government.

References

Adar, E., Zhang, L., Adamic, L.A., and Lukose, R.M. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, volume 13, 2004.

Blei, D. and Lafferty, J. Correlated topic models. In *Proc. 18th Advances in Neural Information Processing Systems*, pp. 147–152, 2006a.

Blei, David M. and Lafferty, John D. Dynamic topic models. In *Proc. 23rd Intl. Conf. on Machine Learning*, pp. 113–120, 2006b.

Dietz, Laura, Bickel, Steffen, and Scheffer, Tobias. Unsupervised prediction of citation influences. In *Proc. 24th Intl. Conf. on Machine Learning*, pp. 233–240, 2007.

Domingos, Pedro and Richardson, Matt. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 57–66, 2001.

Dong, Xin Luna, Berti-Equille, Laure, and Srivastava, Divesh. Truth discovery and copying detection in a dynamic world. *Proc. VLDB Endow.*, 2(1), August 2009.

Dong, Xin Luna, Berti-Equille, Laure, Hu, Yifan, and Srivastava, Divesh. Global detection of complex copying relationships between sources. *Proc. VLDB Endow.*, 3 (1-2), September 2010.

Du, Nan, Song, Le, Yuan, Song, and Smola, Alex J. Learning networks of heterogeneous influence. In *Proc. 24th Advances in Neural Information Processing Systems*, pp. 2780–2788. 2012.

Du, Nan, Song, Le, Woo, Hyenkyun, and Zha, Hongyuan. Uncover topic-sensitive information diffusion networks. In *Proc. 16th Intl. Conf. on Artificial Intelligence and Statistics*, 2013.

Foulds, James and Smyth, Padhraic. Modeling scientific impact with topical influence regression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 113–123, 2013.

Gomez-Rodriguez, Manuel, Balduzzi, David, and Schölkopf, Bernhard. Uncovering the temporal dynamics of diffusion networks. In *Proc. 28th Intl. Conf. on Machine Learning*, pp. 561–568, 2011.

Gomez-Rodriguez, Manuel, Leskovec, Jure, and Krause, Andrease. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data*, 5(4), 2012.

Gruhl, Daniel, Guha, R., Liben-Nowell, David, and Tomkins, Andrew. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pp. 491–501, 2004.

Guo, Fangjian, Blundell, Charles, Wallach, Hanna, and Heller, Katherine. The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Proc. 18th Intl. Conf. on Artificial Intelligence and Statistics*, 2015.

Kempe, David, Kleinberg, Jon, and Tardos, Éva. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 137–146, 2003.

Leskovec, Jure, Kleinberg, Jon, and Faloutsos, Christos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pp. 177–187, 2005.

Leskovec, Jure, Adamic, Lada A., and Huberman, Bernardo A. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

Li, Liangda, Deng, Hongbo, Dong, Anlei, Chang, Yi, and Zha, Hongyuan. Identifying and labeling search tasks via query-based Hawkes processes. In *Proc. 20th Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 731–740, 2014.

Liniger, Thomas Josef. *Multivariate Hawkes processes*. PhD thesis, ETH Zurich University, 2009.

Liu, Lu, Tang, Jie, Han, Jiawei, Jiang, Meng, and Yang, Shiqiang. Mining topic-level influence in heterogeneous networks. In *Proc. 19th Intl. Conf. on Information and Knowledge Management*, pp. 199–208, 2010.

Praneeth, Netrapalli and Sujay, Sanghavi. Learning the graph of epidemic cascades. In *Proc. 12th ACM Sigmetrics Conf. on Measurement and Modeling of Computer Systems*, pp. 211–222, 2012.

Simma, Aleksandr. *Modeling Events in Time Using Cascades Of Poisson Processes*. PhD thesis, EECS Department, University of California, Berkeley, Jul 2010.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *Proc. 26th Intl. Conf. on Machine Learning*, pp. 1105–1112, 2009.

Wang, Chong and Blei, David M. Variational inference in nonconjugate models. *J. Mach. Learn. Res.*, 14(1), April 2013.

Wang, Senzhang, Hu, Xia, Yu, Philip S., and Li, Zhoujun. MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proc. 20th Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 1246–1255, 2014.

Wang, Yu, Cong, Gao, Song, Guojie, and Xie, Kunqing. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proc. 16th Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 1039–1048, 2010.

Yang, Shuang-Hong and Zha, Hongyuan. Mixture of mutually exciting processes for viral diffusion. In *Proc. 30th Intl. Conf. on Machine Learning*, 2013.

Zhou, Ke, Zha, Hongyuan, and Song, Le. Learning social infectivity in sparse low-rank network using multi-dimensional Hawkes processes. In *Proc. 30th Intl. Conf. on Machine Learning*, 2013.