Time varying Topics for Modeling Content Diffusion over Social Network

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFIMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

Master of Engineering

IN

System Science and Automation

BY

Pushkar Nagar

Electrical Department
Indian Institute of Science
Bangalore – 560 012 (INDIA)

May, 2021

Declaration of Originality

I, **Pushkar Nagar**, with SR No. **04-03-02-10-41-15-1-12371** hereby declare that the material presented in the thesis titled

Time varying Topics for Modeling Content Diffusion over Social Network

represents original work carried out by me in the **Department of Electrical Engineering** at **Indian Institute of Science** during the years **2015-2017**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discusions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Prof. Chiranjib Bhattacharyya Advisor Signature

DEDICATED TO

Family and Friends

Acknowledgements

I would like to express my sincere gratitude to my project advisor, Prof. Chiranjib Bhattacharyya for giving me an opportunity to work on this project. I am thankful to him for his valuable guidance and moral support. I feel lucky to be able to work under his supervision.

I also sincerely thank my lab mates for constant motivation and support to put all the ideas into reality. I am overwhelmed to acknowledge their humbleness.

I would also like to thank the Department of Electrical Engineering for providing excellent study environment. The learning experience has been really wonderful here.

Finally I would like to thank all IISc staff, my family and friends for helping me at critical junctures and making this project possible.

Abstract

Clustering in document streams, such as online news articles or continuous stream documents can be induced by their textual contents, as well as by the temporal dynamics of their arriving patterns. We consider the clustering of short length documents by taking into account the temporal dynamics of clusters(topics) and how these topics are influencing each other i.e. also obtaining the Topic Influence Network.

Topic Influence Network is helpful in understanding the dynamics of information diffusion through cluster network at broader level, which helps in analyzing the generation of clusters which are being affected by generation of other clusters. We consider a situation which is analogous to finding the topic influence network where the users are considered as source of posting documents thus forming a network. Also we study how information diffusion occurs across the network which eventually helps in finding the user level clusters and topic level clusters.

Contents

List of Figures

Chapter 1

Introduction

1.1 Real World Document Streams

The increasing prevalence of mobile devices and social apps have facilitated the collection of news and information by people connected to internet. Massive amount of user generated content, mostly in the form of short documents are clustered around real life events. It is important to effectively organize this data according to their contents such that online users can digest them easily. Besides textual information, temporal information also helps in clustering of online documents mainly short documents based on some notion. For an instance, weather reports, forecasts and warnings of the blizzard in New York City this year appeared online before the snowstorm actually started. As the blizzard conditions gradually intensified, more subsequent blogs, tweets were triggered around this event in various online social media. Since in successive time period, large documents are being generated which are centered around this event and this event may eventually influence other events as well as the source of event will influence other source of events. From now we will assume that finding topic influence network is similar to finding the influence network over the various source of documents, since both the problems are analogous. In some other situations, relevant news articles or tweets show different temporal dynamics, for example documents on emergency may rise and fall quickly with some intensity while some rumours may have far reaching influence. Such self-excitation phenomenon often leads to many closely related articles within a short period of time. Hawkes processes[8, 2] can be used to capture above mentioned phenomena since posting of documents with respect to time can be considered as events. Poisson processes are generally used in modelling temporal point patterns and Hawkes process is special kind of Poisson process which has many benefits.

Using Hawkes Process helps in many advantages like, preferential attachment i.e. larger the intensity for a Hawkes Process, the more likely the next event is from that cluster(topic). Also Hawkes Process is suitable to model temporal behavior because it can capture self excitation of clusters and temporal decays i.e. a particular document belonging to a cluster temporally decays over time.

1.2 Temporal Clustering and its review

Clustering of documents by using the time information that how documents are arriving temporally is known as Temporal Clustering. There are models present which have considered the dynamic evolution of topics over time, but very less work has been done for clustering documents which leverages time information attached with document and there is no work done for considering how one topic influences other topics or how once source of documents influence other source of documents. In Dynamic topics models[3] we divide the data by time slices, for example by months. We model the documents of each time slice with a K-component topic model, where the topics associated with time t evolve from the topics associated with time t-1. As we discussed, temporal nature of topic evolution varies across different time periods so it is not able to capture long tail temporal dynamics of arriving of documents. Thus clustering of documents can be improved by taking into account the underlying heterogeneous temporal dynamics. Different temporal dynamics for different clusters will also help in to track their popularity and to predict future topics and finding how topics are temporally influencing the generation of other topics.

1.3 Information Diffusion

Understanding the information diffusion in social networks and social media at broader level requires modeling the topic diffusion process, and how temporal evolution of one topic affects the temporal evolution of other topic. This notion is defined by us as **Topic Influence Network** which is not been explored yet. Such kind of networks play a variety of roles in various domains, like monitoring the spread of news, evaluating the effects of network in marketing etc. Although there is a lot of work done in understanding the diffusion of information in social networks and social media at user level, like Hawkes Topic Model[5] which combines Hawkes Processes and topic modeling to simultaneously reason about the information diffusion pathways and the topics characterizing the observed textual information, but they have considered the users as nodes and none of them has utilized both time and content of documents which actually provides better insights of real life scenarios.

1.4 Our Contributions

People have worked on discovering latent influence in online social activities[12] , where we find latent influence network of users i.e. how one user influences other user, but they do not take into account content information i.e. how information spreads through the network.

- We extended the notion of Dirichlet-Hawkes process to introduce the multidimensional Hawkes process that can jointly model the time and topics of events that comes from multiple dependent sources.
- We can easily extend our model in which a document comes from multiple topics not a single topic.
- We can do multiple analysis of data, such as clustering of users based on similar behavior and clustering of topic of events which help in finding the users interest and current popular topics too.
- We have done experiments on both real Dataset and Synthetic Dataset.

1.5 Structure of the Thesis

- In introduction chapter we have talk about the time varying clustering and information diffusion in social network and brief description of our area related to it.
- In problem definition chapter we have given precise description of problem current drawbacks and finally we have given the formal model definition.
- In background chapter we have explain the tools we required for further understanding the model.
- In related work chapter we have talk about some of the work which has jointly model the time and content of document.
- In Dirichlet Hawkes Process Model chapter we have explained the model currently present in literature work.
- Generalized Proposed Mode chapter we have generalized the dirichlet hawkes process model which can be applicable in different real world scenarios.

- In experiments chapter we have shown some of the results on both synthetic data and real world data.
- In conclusion and future work we have precisely describe our work and its future aspects.

Chapter 2

Problem definition

2.1 Problem Setting

Temporal clustering refers to time varying clustering of documents. In sequential streaming, data comes time-wise, so we can model the arriving of documents as temporal point patterns. Also due to self exciting and temporal decay property we can use a particular type of point process called Hawkes Process. We are addressing a particular scenario in which users post an article or documents which may trigger other users also to post the same document that belong to same topic index or may start a new topic which actually results in forming a network of users. In this network, users can influence other users in adapting the same trends of documents and such a scenario has not been explored before. Such scenarios are useful in many ways, we can predict each user's interest over time and how information diffusion process occurs in network of users.

2.2 Drawbacks

Recent paper by Nan du[2] in this direction has not taken into account such scenarios and their model is not applicable in our setting. Hence they are not able to get insights of users' network formed by influential posting of documents. For capturing the influence of users we are using extension of univariate Hawkes process called Multi-variate Hawkes process[8],[10] which works well in such cases. Multi-variate Hawkes process gives insights about how information diffusion occurs in user's network and clustering of user's which follows the same behavior in adapting the interest over time.

2.3 Formal Model Definition

In nutshell, the problem we are addressing is to find the time varying topics of documents by considering the temporal dynamics of evolution of topics and also how information diffusion occurs at broader level i.e. how temporal evolution of one topic affects the temporal evolution of other topics. In short we are also interested in finding the Topic Influence Network. The main idea is to use multi-variate point process to model the time of events and non-parametric model to model the marks of events. The setting where we are considering the topic influence network is analogous to setting where users are forming the influence network. So we are considering users setting due to availability of public dataset. The data consists of D(t) $=e_{i}{}_{i=1}^{N(t)}$ which denotes the set of events observed until time t, where event e_i is a triple (t_i, u_i, d_i) which indicates that at time t_i, user u_i shares document d_i and users will influence each other instead of influence of topics. Since users of network influence each other, the events in a network are mutually exciting, i.e. each event triggers some new events in the network. Each event e_i has an associated latent factor called topic index θ_i. We have assumed that event which has triggered preceding event has same topic index.

Chapter 3

Background

A temporal point pattern can be explained as a sequence of times of events. Temporal point patterns can be used to model many real life phenomena such as sequence of arrivals of requests at a server, sequence of earthquakes, etc. Usually complex mechanisms drive these seemingly random phenomena therefore the time and location of occurrence of these events is usually unknown in advance. For modeling these special mechanisms, especially in predicting future events, a powerful tool can be used which is a stochastic process -'a temporal point process'.

Given, all the times of previously occurred events, a point process can be defined by specifying a stochastic model for the time of the next event. This is referred to as an evolutionary point process. A point process is an ordered list of times $t_1, ..., t_N$ at which an N sequence of events $E_1, ..., E_N$ occur. Let $N(t)$ denote the number of points (i.e. occurrences of events) in $(\infty, t]$ and $H_t = \{E|t_E < t\}$ be the history of events up to but not including t. Let the conditional density function of the time of the next event t_{n+1} given the history of previous events $(..., t_{n-1}, t_n)$ be denoted by $f^*(t) = f(t|H_t)$. The density function $f^*(t)$ is used to specify the distribution of all interevent times i.e. the lengths of the time intervals between subsequent events, starting in the past. The distribution of all events is given by the joint density

$$f(.., t_1, t_2, ..) = \prod_i f(t_i|..., t_{i-2}, t_{i-1}) = \prod_i f^*(t_i) \tag{3.1}$$

The conditional intensity function also called the hazard function is defined as

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)} \tag{3.2}$$

Alternately, it can be defined as

$$\lambda(t|H_t) = \lim_{\Delta t->0} \frac{\mathbb{E}[N(t+\Delta t)|H_t]}{\Delta t} = \frac{\mathbb{E}[dN(t)|H_t]}{dt} \tag{3.3}$$

where $\mathbb{E}[dN(t)|H_t]$ is the expected number of events occuring in the interval $(t, t+dt]$ given the past historical observations H_t . Thus it is easy to see that the conditional intensity function $\lambda^*(t)$ represents the expected instantaneous rate of events at time t.

3.1 Hawkes process

The Hawkes process is a special class of point process models that is self or mutually exciting (Hawkes, 1971)[8] i.e. happening of previous events triggers the occurrence of same type of events in near future. This can also be termed as self-exciting Hawkes process. A single variable (univariate) Hawkes process is defined as

$$\lambda^*(t) = \mu(t) + \alpha \sum_{t_k<t} \gamma(t - t_k; \beta) \tag{3.4}$$

where $\mu : \mathbb{R} \longrightarrow \mathbb{R}^+$ is a deterministic base intensity for the occurrence of events, $\gamma(t;\beta)$ represents a density function on $(0,\infty)$ depending on parameter β which exhibits the time-decay effect and α is a positive parameter. If & μ is considered a constant and the density is taken to be exponential , the equation becomes :

$$\lambda^*(t) = \mu + \alpha \sum_{t_k<t} \exp(-(t - t_k)) \tag{3.5}$$

Looking through the model we can see that when a new point arrives in the process, the conditional intensity grows by α and then decreases exponentially back towards μ. Alternately, we can say this type of point process enhances the chance of getting other points immediately after, thus representing a model for clustered point patterns (self-excitation). Therefore *self-exciting* nature of Hawkes Process is very well utilised in modeling phenomena in varied fields such as crimes (Mohler et al.,2011)[9], financial contagions (Bacrya et al., 2012)[10] and earthquake aftershocks (Ogata, 1981)[**?**]. A simulation of Hawkes process with parameters $(\mu, \alpha, \beta) = (0.13, 0.023, 0.11)$ as well as its conditional intensity is shown in Figure 3.1.

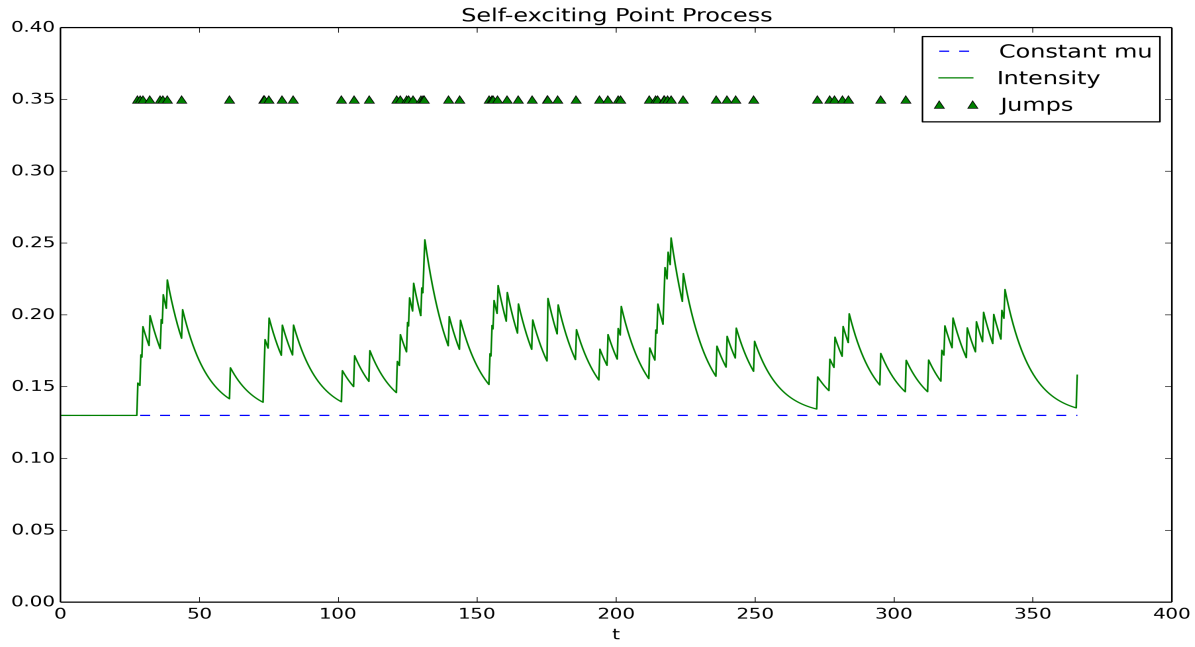

Figure 3.1: Simulation of Hawkes process

A simulation of the Hawkes process with parameter $(\mu, \alpha, \beta) = (0.13, 0.023, 0.11)$ is seen at the bottom of this plot,and the corresponding conditional intensity function can be seen in the top.Image Credits[9].

3.2 Multi-Variate Hawkes process

The multivariate Hawkes process is a multi-dimensional extension of single variable Hawkes Process (Hawkes, 1971; Liniger, 2009[11]). For this type of process, a dimension or a mark is associated with each point corresponding to the dimension in which it occurred i.e. we can say each t_l has a dimension i associated with it. Hence $T = \{t_1^{i_1}, t_2^{i_2}, ...t_n^{i_n}\}$ where each i_j comes from a finite set I of dimensions, 'D'. These are characterized by conditional intensity functions $\lambda_i^*(t)$ for each dimension $i \in I$. The intensity function $\lambda^* = [\lambda_1^*, ..., \lambda_D^*]^T$ is defined by

$$\lambda_i^*(t) = \mu_i + \sum_{t_l^j < t} \alpha_{ji} \kappa(t - t_l^j) \tag{3.6}$$

where $\mu_i \in \mathbb{R}^+$ is the baseline or intrinsic intensity in dimension i, $\kappa : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is a time-decaying triggering kernel. Each α_{ij} entry comes from an asymmetric infectivity matrix $\alpha \in \mathbb{R}^+_{D \times D}$ which characterizes the structure of the network where each element of matrix α_{ij} expresses the degree of influence from a node i to another node j. The expression $\sum_{t_l^j < t} \alpha_{ji} \kappa(t-$

t_l^j) quantifies the influence of historical events on the instantaneous rate of event at time t in dimension i.

3.3 Dirichlet Process

Dirichlet process is a Bayesian Non-Parametric process. Bayesian non-parametric processes are infinite parameter processes which are widely used in probabilistic graphical models. It is parametrized by concentration parameter α and a base distribution $G_0(\theta)$ over a given space $\theta \in \Theta$. Thus a sample drawn from Dirichlet Process $G \sim DP(\alpha, G_0)$ is a discrete distribution where Base distribution can also be continuous while the concentration parameter controls the level of descretization in G. If $\alpha \to 0$, distribution is concentrated on single value, while if $\alpha \to \infty$, then the distribution becomes continuous and for α in between 0 and ∞ we get discrete distributions with less concentration as α increases.

From DP(Dirichlet process) we get a Distribution which can be used as a parameter to draw samples from it namely $\theta_{1:n}$ and use these samples as parameters for models of clusters. Now we will show that these two stage process can be simulated as follows:

1. Draw θ_1 from G_0
2. For n > 1:
   (a) With probability $\frac{\alpha}{\alpha+n-1}$ draw θ_n from G_0
   (b) With probability $\frac{m_k}{\alpha+n-1}$ reuse θ_k for θ_n , where m_k is the number of previous samples with value θ_k , where θ_k is the set of distinct values in $\theta_{1:n}$

The above simulation process also called Chinese Restaurant Process, which captures the "rich get richer" or preferential attachment phenomenon. Thus CRP(Chinese Restaurant Process) has property to generate infinite number of clusters , since there is always some probability to make new cluster adapted to increasing complexity of data.

Chapter 4

Related Work

In this section we will talk about some of the related work which jointly models time and content of document.

- Xinran He et al.[5] proposed a joint model for Network Inference and Topic Modeling from Text-Based Cascades which focused on inferring the diffusion of information together with the topics characterizing the information.It is more useful to infer the hidden diffusion network and tracking of thematic content of documents as they propagate through the diffusion network. They consider a text content cascades among a set of nodes V = $1, 2, ...|V|$. They have considered that each node may belong to one user in social network, one researcher in research community etc. They consider a tuple say a^i = (t^i,v^i,x^i) , this means node v^i posts document x^i at time t^i. The model consist of two components. Firstly, modeling the posting time of activity by a user with the help of Multivariate Hawkes Process[5] and secondly modeling the documents. In MHP(Multivariate Hawkes Process) model, each node is associated with a point process and the occurrence of one activity can lead to 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 mutual exciting property makes the MHP model a perfect fit for this type of setting.

- Shangsong Liang et. al.[6] proposed a model for Dynamic clustering of short Documents applicable for social network data, twitter dataset , blogs etc. In this they have proposed two dynamic topic models: one for short term dependency of the current inference of the topics and another for long term dependency. Based on these topic models they have captured the dynamic changes in clusters. But their models do not capture the long tail influence of generation of topics.

- D. M. Blei et. al.[3] proposed model where the parameters at time t comes from $t-1$, but it does not capture the changes in temporal dynamics of clusters. This is the first model which takes into account of time factor in topic modeling literature.
- Nan Du et. al.[11] proposed a time series model based on Multi Variate Hawkes Process, but it does not take into account content. They consider the events as sequence of time events and creates a diffusion network.
- S. Hosseini et. al. [13] proposed a joint model of time and content, in which they consider the influence matrix consisting of network of users. We have used the same dataset namely Event Registry.

Chapter 5

Dirichlet Hawkes Process Model

5.1 Dirichlet Hawkes Process Model

Here we are defining the recently proposed model based on Hawkes Process and Dirichlet Process by Nan Du[2]. He proposed Dirichlet-Hawkes Process as a prior for modeling the continuous time documents. It captures long tail temporal dynamics of clusters easily, the set of $\{\theta_i\}$ sampled from Dirichlet Hawkes Process will be used as parameters for document content modelling whereas, each θ_i(latent variables) belongs to one cluster, while different clusters follows different temporal dynamics. A sample drawn from Dirichlet Hawkes Process, which shows two Hawkes Process separately for θ_1 and θ_2 (cluster indicator variables also called hidden variables) is shown in Figure 5.1. The sample shown from Dirichlet Hawkes Process, the two initial events t_1 and t_2 generate a Hawkes Process of their own, with events at time { t_2, t_3, t_5, t_9} and { t_6, t_7, t_8 } , respectively. We will define now the overall generative process of the model.

1. Draw t_1 from Poisson(λ_0) and θ_1 from Dir($\theta|\theta_0$)
2. For each word v in document 1 : $w_1^v \sim$ Multi(θ_1)
3. For n > 1:
   (a) Draw $t_n > t_{n-1}$ from Poisson($\lambda_0 + \sum_{i=1}^{n-1} \gamma_{\theta_i}(t_n, t_i)$, where $\gamma_{\theta_i}(t_n , t_i)$ is kernel for Hawkes Process
   (b) Draw θ_n from Dir($\theta|\theta_0$) with probability $\frac{\lambda_0}{\lambda_0 + \sum_{i=1}^{n-1} \gamma_{\theta_i}(t_n, t_i)}$, and draw α_{θ_n} from Dir($\alpha|\alpha_0$)
   (c) Reuse previous θ_k for θ_n with probability $\frac{\lambda_{\theta_k}(t_n)}{\lambda_0 + \sum_{i=1}^{n-1} \gamma_{\theta_i}(t_n, t_i)}$, where $\lambda_{\theta_k}(t_n) = \sum_{i=1}^{n-1} \gamma_{\theta_i}(t_n, t_i) I[\theta_i = \theta_k]$

(d) For each word v in document n: $w_n^v \sim Multi(\theta_n)$

In above model different clusters can have different temporal dynamics. Triggering kernel funtion of Hawkes process is represented as a non-negative combination of K base kernel functions i.e

$$\gamma_{\theta_i}(t_i, t_j) = \sum_{l=1}^{K} \alpha_\theta^l \ . \ \kappa(\tau_l, t_i - t_j)$$

where α_θ draws from Dirichlet Distribution, which helps in tracking the different evolving temporal dynamics of clusters , $t_j < t_i$, τ_i is the typical reference time points. Figure 5.2 shows the overview of Dirichlet Hawkes Process.

5.2 Inference

In above model to calculate the posterior distribution, the inference algorithm alternates between two subroutines. The first subroutine samples the latent cluster membership for the current document d_n by Sequential Monte Carlo [10,11]; and then, the second subroutine updates the parameters of learned triggering kernels of the respective cluster.

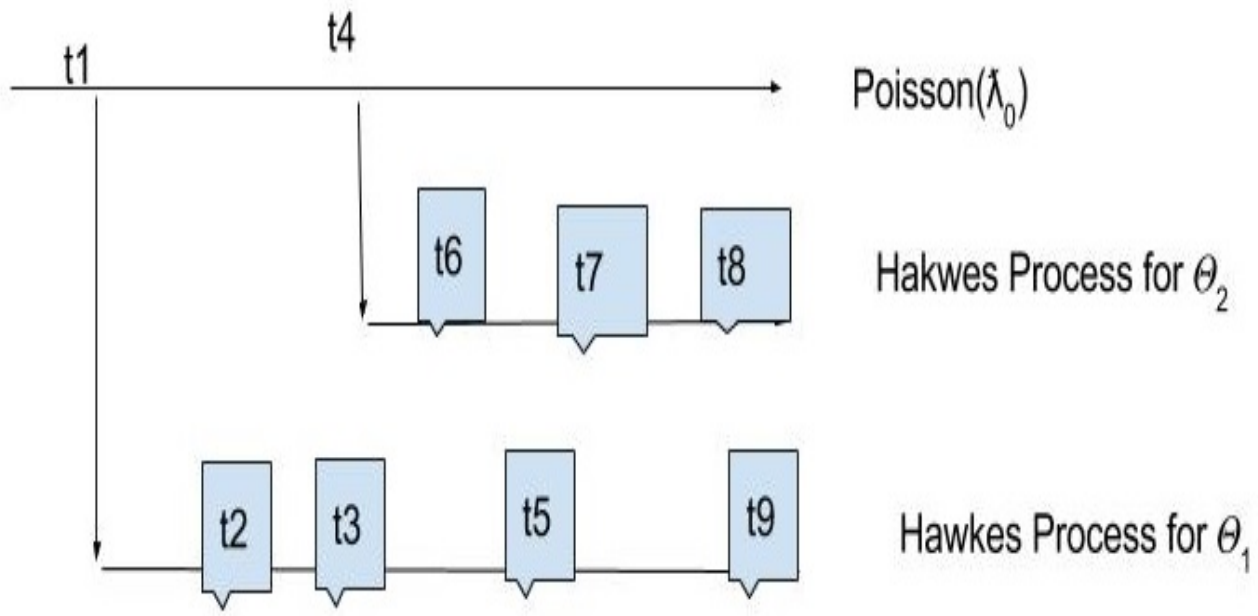

Figure 5.1: A sample from Dirichlet Hawkes Process. Simulation of two Hawkes Process for two different clusters respectively θ_1 and θ_2

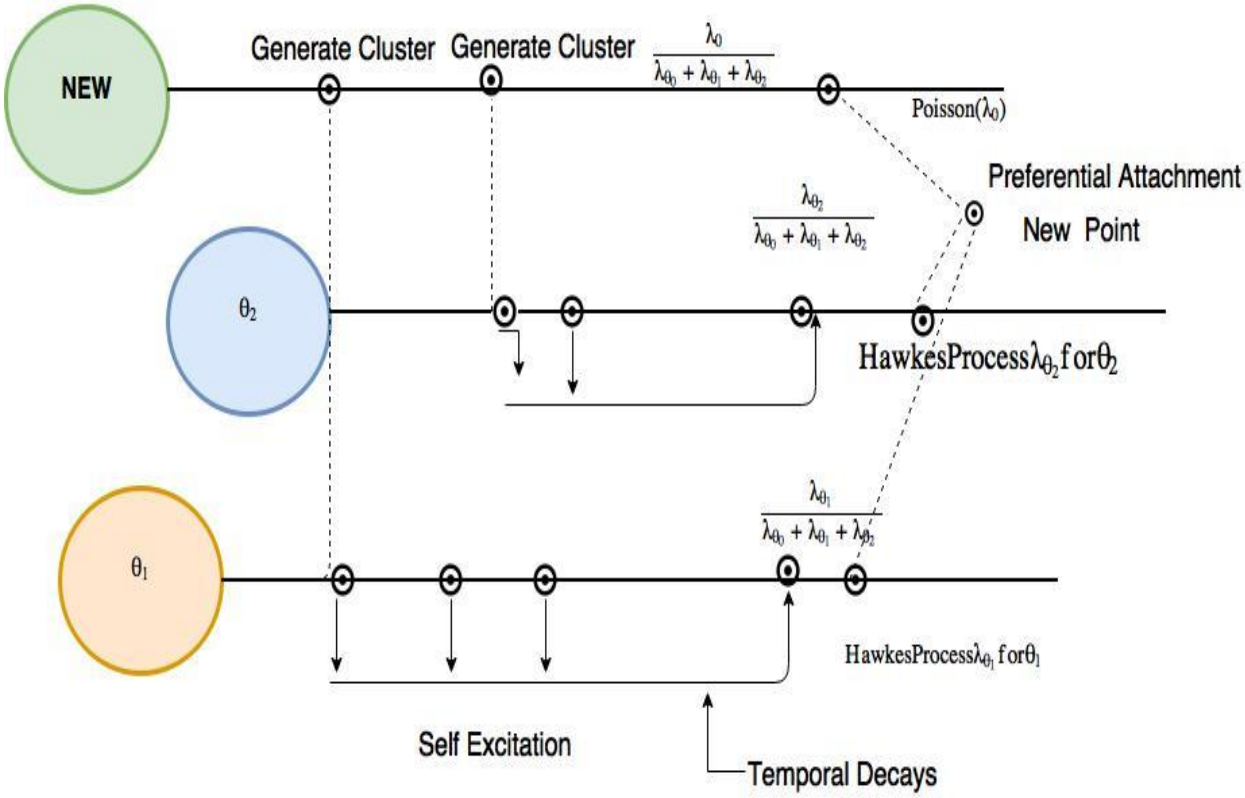

Figure 5.2: Overview of Dirichlet Hawkes Process

Chapter 6

Generalized Proposed Model

As we discussed, the above Generative model based on Dirichlet Hawkes Process captures the different temporal dynamics of clusters, but it does not capture how temporal evolution of cluster affects other clusters which is very obvious in real life scenario. Thus we can include the influence matrix in Multi Variate Hawkes Process(MHP) to capture the influence of one cluster on another cluster , and also to find the topic cluster influence network. In this direction for MHP, the intensity process λ_θ(t/H) takes the following form:

$$\lambda_\theta(\text{t/H}) = \mu_\theta + \sum_{e:t_e<t} \kappa_e\ (t, \theta)\ ,$$

where μ_θ is the base intensity of the process, while each previous event e adds a non-negative impulse response $\kappa_e\ (t, \theta)$ to the intensity, increasing the likelihood of the future events. Impulse response drives two factors: firstly it captures the influence between clusters and secondly how the diffusion occurs in topic cluster network.

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

Here, A= { $A_{u,w}$ } is non-negative matrix modeling the influence between clusters. $A_{u,w}$ is the expected number of events that a single event at cluster u can trigger the process of cluster w. The larger the $A_{u,w}$, the stronger the influence cluster u has on cluster w. $f_\Delta(.)$ is the probability density function for delay distribution which captures how long it takes for an event(generation of documents that belong to a particular topic) at one cluster to influence other node and how this influence will last.

Chapter 7

Experiments

7.1 Problem Setting Description

We will consider a scenario where we will apply our model. Let D(t) $=e_{ii=1}^{N(t)}$ denotes the set of events observed until time t, where event e_i is a triple (t_i, u_i, d_i) which indicates that at time t_i, user u_i shares document d_i and user's will influence each other instead of topic. Since users of network influence each other, the events in a network are mutually exciting, i.e. each event triggers some new events in the network. Each event e_i has a associated latent factor called topic index θ_i. We have assumed that event which has triggered preceding event has same topic index. Each topic is a distribution over words of dictionary and for simplicity we have considered every document belongs to one of the topics.
We will consider the following notations. $\{\phi_k\}_1^{K(t)}$ denotes set of topics over network until time t. $\{\psi_{uk}\}_{i=1}^{K_u(t)}$ denotes topics belonging to interest of user u. $K_u(t)$ denotes number of topic user u is interested in at time t. $N(t)$ represents the number of events until time t. $D_{uk}(t)$ denotes the set of events triggered by event s generated by user u until time t with topic ϕ_k. Let $D^0(t)$ be the set of events generated due to base intensity of user until time t. We will use dot notation to represent union over the dotted variable. $D_{u.}(t)$ represent the events of user u before time t with any topic, and $D_{uk}(t)$ represent the events of all users except u, before time t, with topic k. Let Z_i be the index of the topic of the i_{th} event.
According to our model, we assume that the time at which user u publishes documents follows a Hawkes process with intensity function:

$$\lambda_u(t) = \mu_u + \sum_{s=1}^{N(t)-1} \lambda_u(t, s)$$

where μ_u is the base intensity with which user generate new events. $N(t)$ represents the number

of events until time t. $\lambda_u(t,s)$ is the amount of intensity of user u at time t that is caused by event s. $\lambda_u(t,s)$ is defined as $\alpha_{u_s,u}\kappa_{z_s}(t,t_s)$ where $\alpha_{u_s,u}$ is the influence of event $s's$ user on u. $\kappa_{z_s}(t,t_s)$ is a kernel function which determines the diffusion rate of events associated with topic z_s. For simplicity we will use exponential kernel:

$$\kappa_{z_s}(t,t_s) = \exp\left(-\beta_k(t-t_s)\right)$$

7.2 Generative Process

Generative Process: For all events e_i:

1. User u publishes documents which follows a Hawkes process with intensity function: $\lambda_u(t) = \mu_u + \sum_{s=1}^{N(t)-1} \lambda_u(t,s)$
2. $\lambda_u(t,s) = \alpha_{u_s,u}\kappa_{z_s}(t,t_s)$ and $\kappa_{z_s}(t,t_s) = \exp\left(-\beta_k(t-t_s)\right)$
3. User u draws reused topic ψ_{uj} with probability $\frac{n_{uj}(t)}{n_{u:}(t)+\gamma}$, where $\mathrm{n}_{uk}(t) = \mu_u + \sum_{e \epsilon D_u^0(t)} \exp\left(-\nu(t-t_e)\right)$ $\psi_{uk}]$
4. User u draws new topic $\psi_{u,new}$ with probability $\frac{\gamma}{n_{u:}(t)+\gamma}$.
5. Draw $\psi_{u,new}$ with probability $\frac{m_k(t)}{n_{u:}(t)+\zeta}$ or new topic with probability $\frac{\zeta}{m_{.}(t)+\zeta}$ where $\mathrm{m}_k(t) = \sum_{e \epsilon D_u^0(t)} \exp\left(-\nu_k(t-t_e)\right) I[\theta_e = \phi, l_e = 1]$
6. $\mathrm{w}_{e_i}^v \sim Multi(\phi_k)$

The inference procedure is same is we described in section(5). We evaluate the performance of model by using both synthetic and real data. Experiments on synthetic data are primarily used to evaluate the inference algorithm. While using real dataset we evaluate the performance of model in inferring the topics as well as temporal behaviour of topics.

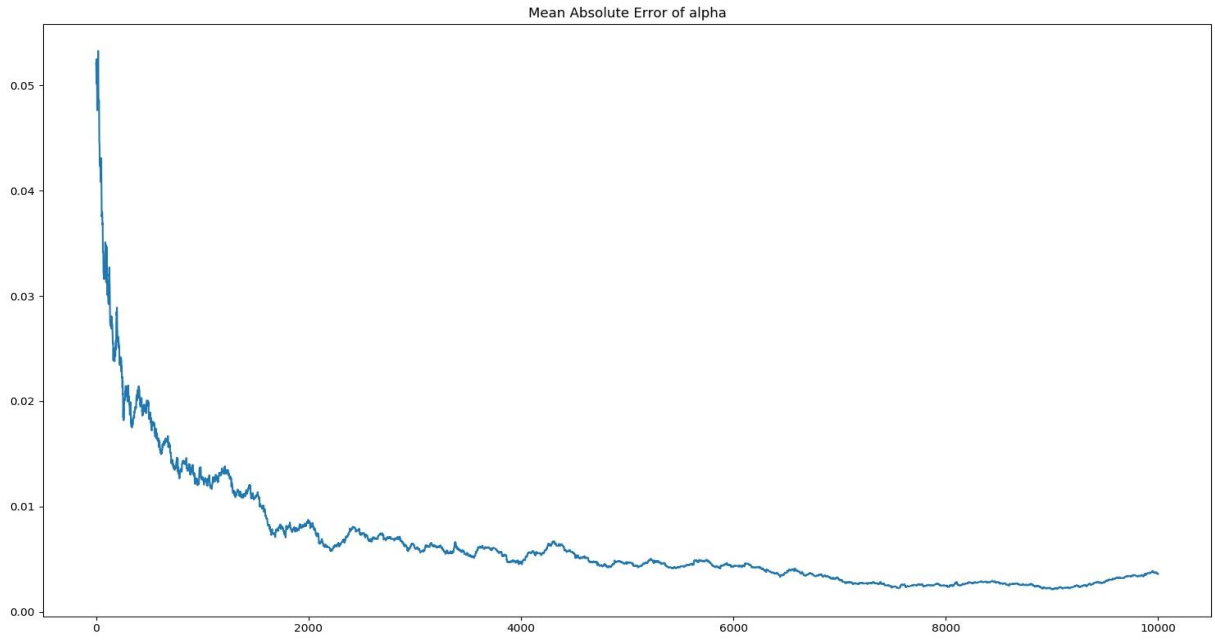

Figure 7.1: Mean Absolute error of α(Influence matrix of users) vs Number of events

7.3 Synthetic Data

We generate synthetic dataset by using the generative process through simulation by the famous Ogata's Thinning algorithm[1]. For this we provided fixed prior parameters, which consists of base intensity μ_u for each user and β as an exponential kernel parameter already defined in Section 7.1. An event is defined to be "user posting a document" and so we generated 10^4 events. The event consists of user representing the dimension, the time of event as well as document content posted by user which comes from a vocabulary of size 20. We have evaluated the inference procedure performance on the basis of Mean Absolute Error of matrix α and Mean Absolute error(MAE) of base intensity μ_u of all user learned by it. Mean Absolute error is average absolute error between the estimated parameters obtained by inference procedure and original parameters that are used for generating the data. Being precise in case of matrix, it is the sum of absolute difference of matrix entries obtained from inference and the original matrix averaged over number of matrix elements and in case of μ_u it is the squared difference of the base intensities averaged over all . As shown in Figure 7.1 and Figure 7.2 our model does not make improvements , but as events are increasing the MAE decreases. Figure 7.1 shows the MAE of estimated α and Figure 7.2 shows the MAE of estimated μ_u.

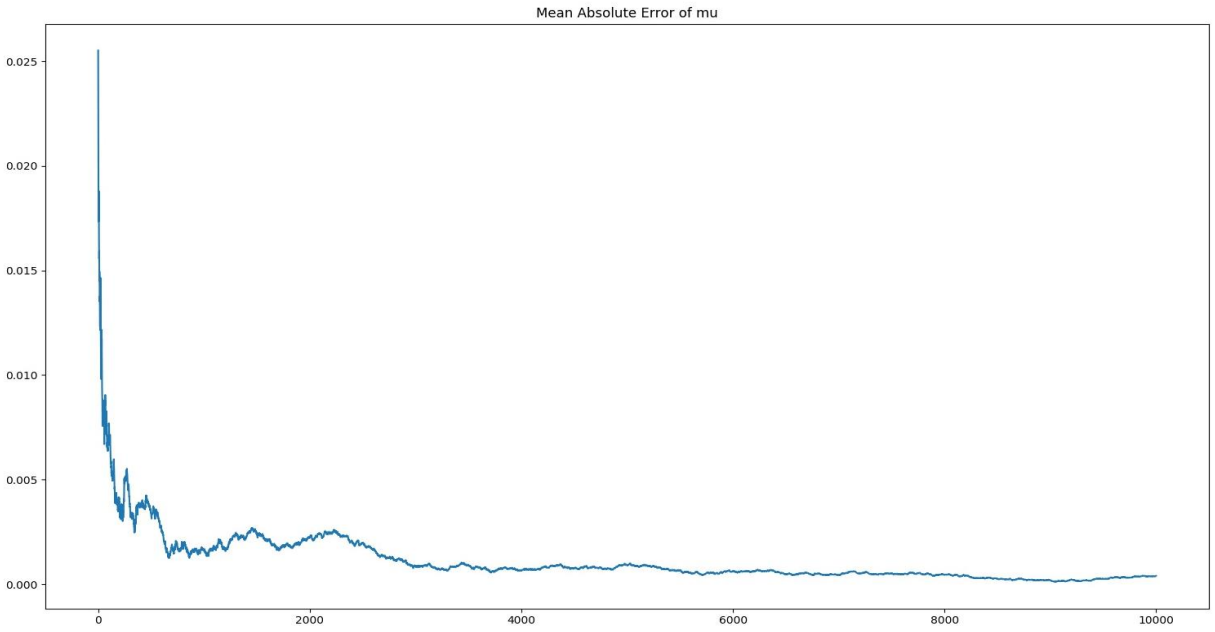

Figure 7.2: Mean Absolute error of μ_us (user's intensity) vs Number of events

7.4 Real Dataset

We also evaluate the performance of the model on real dataset , gathered from EventRegistry[1]. We evaluate the performance on the basis of two things, how different topics are being generated content-wise and how well the model captures the temporal dynamics of topics.

- **Data Description**: Data set extracted from EventRegistry consist of news articles around the real life events of world. Articles consist of mainly 3 different tags; FIFA, Iran-Sanctions, and Paris-Attack from 2015/11/01 to 2016/01/13. Collected data consist of 100 news sites. The news sites are treated as nodes which we consider them to be as users, since users are posting the articles along with other information. We preprocessed the articles and extracted the bag of words model for each article.

- **Content Analysis**: We also show how our model is capturing different topics, for that we depicted the word distribution of top 3 main topics : Paris-attack, Football cup, Iran-sanction deal. Figure(7.4,7.6,7.8) shows the word cloud of top frequent words in 3 main topics. To analyze the temporal behavior of topics, we depicted the intensity function of top 3 topics against time, which is the representative of popularity of topics over time. Figure(7.3,7.5,7.7) represents the intensity function of 3 different detected topics over

[1]http://eventregistry.org/

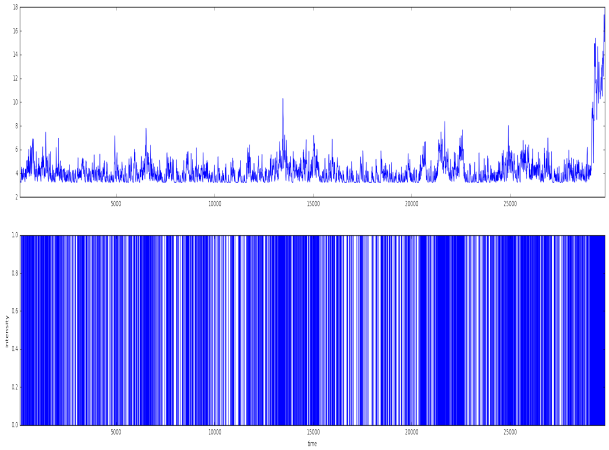

Figure 7.3: Intensity graph for top first topic vs Time

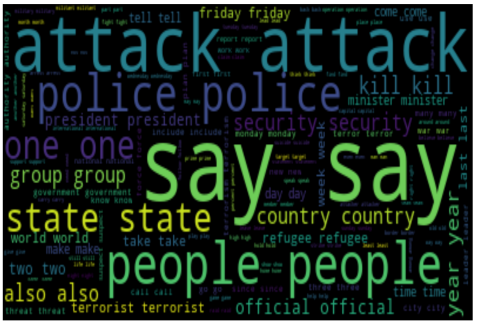

Figure 7.4: Content analysis graph for topic 1

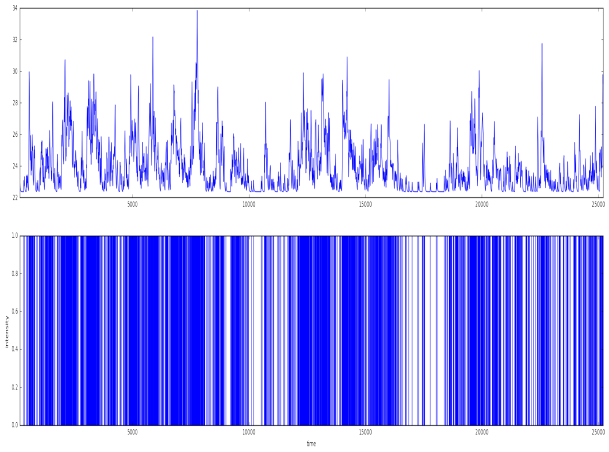

Figure 7.5: Intensity graph for top second topic vs Time

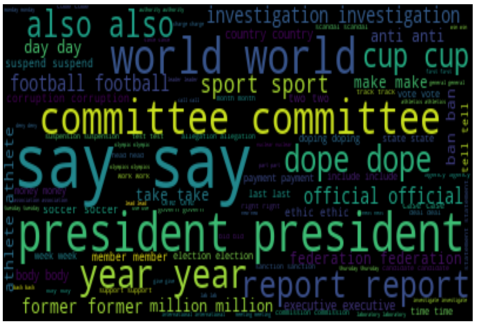

Figure 7.6: Content analysis graph for topic 2

time. Since Football cup is discussed all over the time, its intensity is also distributed over entire time space in figure(7.5,7.6). As we can see in figure(7.7,7.8) Iran-Sanctions topic is a periodical popularity, it is desirable that its popularity rises just after these negotiations. Thus results shows that model is able to capture temporal patterns of evolution of topics over time.

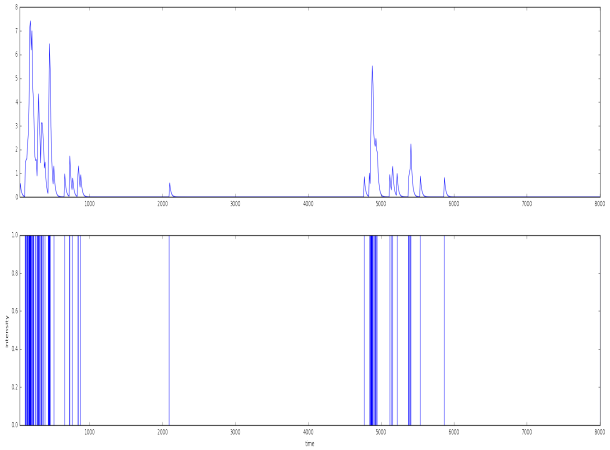

Figure 7.7: Intensity graph for top third topic vs Time

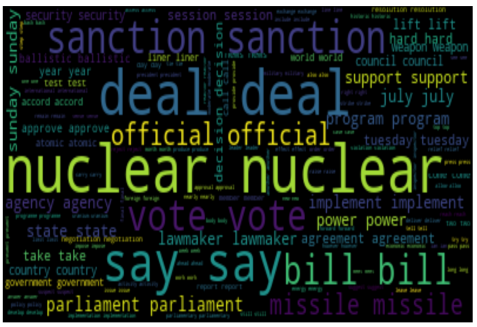

Figure 7.8: Content analysis graph for topic 3

Chapter 8

Conclusion and Future Work

We studied how dependent groups of temporal events can be modelled with long tail dependencies. We could successfully create a joint model for the time and marks of events which adapts itself to the complexity of data. It provides the clustering capabilities at different level such as user level clustering, topic level clustering. In nutshell we used this approach for modeling the content diffusion over social media and the clustering allowed us to infer the source of events and also the hot topic evolving over the network. We also studied different Topic modeling techniques which captures the temporal dynamics, of which Multi-variate Hawkes process came out to be best suited for modeling the time of events as it better captures the influence occurring among users of network. It utilizes the non-parametric methods for modeling mark of events. Due to these capabilities it allows our model to adapt its temporal and topical complexity according to the complexity of data, which is suitable in many real world scenarios. We have shown the performance of model in users network setting on both synthetic and real word dataset using Sequential Monte Carlo method which can efficiently infer parameters of the model. As a future work we can extend this notion of joint model of content and time of events by using neural network architectures like LSTM's.

Bibliography

[1] Ogata Y. On Lewis' simulation method for point processes. IEEE Transactions on Information Theory, 27(1):2331, 1981. 19

[2] N .Du, M. Frajtbar, A. Ahmed, A. J. Smola and L. Song. Dirichlet Hakwes processes with applications to clustering continous document streams KDD, 2015. 1, 5, 13

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models ICML, 2006. 2, 12

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation, Journal of Machine Learning Research, January 2003.

[5] 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 ICML, 2015. 2, 11

[6] S. Liang, E. Yilmaz, E. Kanoulas. Dynamci Clustering of Streaming Short Documents KDD, 2016. 11

[7] L.M. Aiello, Georgios, Sensing Trending Topics in Twitter, IEEE 2013

[8] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes.Biometrika, 58(I):8390, 1971. 1, 5, 8

[9] Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., and Tita, G. E. Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493):100108, 2011. 8

[10] Bacrya, E., Delattreb, S., Hoffmannc, M. and Muzyd,J.F. Modeling microstructure noise with mutually exciting point processes. Quantitative Finance,2012. 5, 8

[11] Liniger, T. Multivariate hawkes processes. ETH Doctoral Dissertation No. 18403, 2009. 9, 12

[12] T.Iwata, A. Shah, Z. Ghahramani Discovering Latent Influence in Online Social Activites via Shared Cascade Poiosson Processes 3

[13] S Hosseini, S.A., Khodadadi, A., Arabzade, S., Rabiee, H.R.: HNP3: A hierarchical non-parametric point process for modeling content diffusion over social media. 12