Learning Root Source with Marked Multivariate Hawkes Processes

Wei Zhang\textsuperscript{*}, Fan Bu\textsuperscript{†}, Derek Owens-Oas\textsuperscript{†}, Katherine Heller\textsuperscript{†}, and Xiaojin Zhu\textsuperscript{*}

\textsuperscript{*}Department of Computer Science, University of Wisconsin-Madison
\textsuperscript{†}Department of Statistical Science, Duke University

Abstract

In group conversations, people make new comments or reply to previous comments. Each root comment can start a new discussion, and many additional comments follow. The result is a structure of multiple comment trees. The question of interest in this paper is “who started it?” We develop a term, root source, to describe the speaker or author of a root comment. With a generative model for comment trees, based on marked multivariate Hawkes processes, we leverage word inheritance, temporal information, and source identity to estimate tree structure. Parameters are inferred using variational expectation maximization with empirical Bayes priors. We then propose a fast algorithm, based on dynamic programming, which computes root source probabilities for each comment.

1 Introduction

Group conversations are complex social phenomena, with individuals making original comments and responding to previous comments from others. In many situations, for example, online forums, a single comment prompts an entire novel conversation topic. On the internet, people have developed a name for the person who starts a new topic, calling the person the OP, for original poster. We generalize the term to accommodate verbal conversations, calling this person the root source.

Knowing the root source is useful for two main reasons, credit attribution and inferring power relations Danescu-Niculescu-Mizil et al. [2012]. In some settings, like Twitter and Facebook, replying and re-posting form a tree-like structure of posts which is observed, and the root source can be found by tracing the tree structure down to its root. However, in other settings, like chat rooms, court room transcriptions, or group conversations, it is not explicit who is responding to whom, and the structure is unobserved. A model to uncover the root source is therefore of substantial value. We call this task root source identification.

To accomplish such task, we jointly model three aspects of information in a group conversation setting: temporal, reciprocal, and textual. The first aspect is often readily available from the timestamp of each comment. The second aspect refers to the mutual influence between speakers in a group, where a comment from person $i$ is more likely to inspire a responding comment from person $j$ if $i$ has stronger influence on $j$, but such network-like power structure is usually latent and needs to be inferred. The third aspect comes from the concept linguistic accommodation, a phenomenon that a speaker tends to borrow vocabulary from the person he/she is responding to. We develop a model based on mutually exciting point processes with textual marks to leverage these three aspects.
1.1 Related Work

In recent years, there has been a growing interest in mutually-exciting point processes to infer latent network structure and study reciprocity. [Blundell et al. [2012]] consider a doubly Hawkes process to model reciprocity between groups of individuals and associate events on co-dependent network edges. [Tan et al. [2016]] extend Blundell et al. [2012] with Gaussian processes to model message significance as well as reciprocity. [Linderman and Adams [2014]] combine Poisson processes and Hawkes processes with random graph models to analyze latent network structures. [Zhou et al. [2013], Du et al. [2013]] model events on social networks as multivariate Hawkes processes to estimate network node influence. [Farajtabar et al. [2015a], Rong et al. [2015]] utilize Hawkes processes to infer the most probable propagation route of events on social networks, and [Farajtabar et al. [2015b]] use interwoven stochastic processes to predict network evolution and event traces.

The stochastic processes approach can also incorporate the text content associated with each event to uncover linguistic patterns. [Guo et al. [2015]] present a Bayesian generative model based on discretized Hawkes processes that captures latent textual influence patterns. [He et al. [2015]] incorporate topic modeling into Hawkes processes to simultaneously characterize textual information and reason about the diffusion pathways in text cascades on social networks. [Farajtabar et al. [2015b]] jointly models information diffusion and network evolution in online social networks using interwoven stochastic processes and generates predictive event traces with learned parameters from history.

Our work differs from previous related works because we make explicit inference about unobserved tree structure. Specifically, we propose a new concept, root source probability, which quantifies uncertainty in assignment of original source for each event on a social network.

The rest of this paper is organized as follows: Section 2 introduces background knowledge about multivariate Hawkes processes and the branching structure formulation, Section 3 formally defines root source probability and states an algorithm for its computation, Section 4 formulates a model for text cascades and develops the parameter estimation algorithm, Section 5 presents experiment results and analysis on synthetic and real data, and Section 6 concludes the paper, pointing out several possible extensions.

2 Background

2.1 Marked Multivariate Hawkes Processes

A $S$-dimensional multivariate Hawkes process (MHP) [Hawkes [1971a], Embrechts et al. [2011]] is a coupling of $S$ counting processes $N(t) = \{N^{(1)}(t), \ldots, N^{(S)}(t)\}$. It can be viewed as a sequence of events $e_1, e_2, \ldots$, and each event $e_i$ consists of a timestamp $t_i$ and a dimension label $s_i$, indicating when and at which dimension the event occurs. The conditional intensity $\lambda(t) = \{\lambda^{(1)}(t), \ldots, \lambda^{(S)}(t)\}$ of MHP takes the form

$$\lambda^{(s)}(t | H_{t-}) = \mu^{(s)}(t) + \sum_{i_{s} < t} \lambda^{(s)}_{i}(t), \quad s = 1, \ldots, S,$$

where $\mu^{(s)}(t)$ and $\lambda^{(s)}_{i}(t)$ are the base intensity and intensity component attributed to previous event $e_i$ for dimension $s$, respectively, and $H_{t-} = \{e_i = \{t_i, s_i\} : t_i < t\}$ is the set of events that occur before $t$.

An important extension of MHP is marked MHP, which introduces a mark $x_i$ to event $e_i$. The mark $x_i$ is often assumed to be drawn after $t_i$ and $s_i$ are drawn, according to a mark density $P(\cdot | t_i, s_i, H_{t-})$. As the notation suggests, in the most general case, mark density could depend on timestamp and dimension label of the current event and all historical events.
2.2 Branching Structure

MHP can be equivalently viewed as a Poisson clustering process [Kingman, 1993, Rasmussen, 2013].
Initially, there are $S$ clusters and each of them is associated with an inhomogeneous Poisson process with base intensity $\mu^{(s)}(t)$. Once the generation procedure starts, events begin to occur at different clusters by the corresponding base intensity $\mu^{(s)}(t)$. These events are called immigrants. Whenever an immigrant event $e_i$ is generated, it adds to every cluster $s$ a new inhomogeneous Poisson process with intensity $\lambda^{(s)}_{ij}(t)$. The new Poisson processes will also generate events, and the events generated by any $\lambda^{(s)}_{ij}(t)$ are called offsprings. Eventually, the event sample is the aggregation of all immigrants and offsprings generated at all clusters.

This Poisson clustering point of view introduces a hidden branching structure for a sample of MHP, defined by a birth-death relationship of each event. Specifically, the parent variable $z_i = (z_{ij})_{j=0,...,n}$ for event $e_i$ is defined as a one-hot vector, such that $z_{i0} = 1$ suggests that $e_i$ is an immigrant from $\mu^{(s)}(t)$ and $z_{ij} = 1$ indicates that it is an offsprin from $\lambda^{(s)}_{ij}(t)$. In addition, based on the superposition property of Poisson processes, the probability distribution of $z_i$ conditioned on $t_i$, $s_i$ and history $\mathcal{H}_{t_{i-}}$ is:

$$P(z_i|t_i, s_i, \mathcal{H}_{t_{i-}}) = \begin{cases} \mu^{(s)}(t_i)/\lambda^{(s)}_{ij}(t_i|\mathcal{H}_{t_{i-}}) & z_{i0} = 1, \\ \lambda^{(s)}_{ij}(t_i)/\lambda^{(s)}_{ij}(t_i|\mathcal{H}_{t_{i-}}) & z_{ij} = 1, \\ 0 & \text{o.w.} \end{cases}$$

(2)

Given the branch structure, a simple form of mark density $P(\cdot|t_i, s_i, \mathcal{H}_{t_{i-}})$ is often chosen with the aim of emphasizing the direct connections of the marks between events and their parent (Rasmussen, 2013). Depending which previous event the “parent” of $e_i$ is, mark $x_i$ is drawn from two cases, i.e.

$$x_i \sim P(\cdot|t_i, s_i, \mathcal{H}_{t_{i-}}) = \begin{cases} f(\cdot|t_i, s_i) & z_{i0} = 1, \\ f(\cdot|t_i, s_i, e_j) & z_{ij} = 1, \end{cases}$$

(3)

where $f(\cdot|t_i, s_i)$ and $f(\cdot|t_i, s_i, e_j)$ are two types of parameterized probability densities. Note that by combining (2) and (3) and marginalizing out parent variable $z_i$, we have

$$P(x_i|t_i, s_i, \mathcal{H}_{t_{i-}}) = \frac{\mu^{(s)}(t_i)}{\lambda^{(s)}_{ij}(t_i)} f(x_i|t_i, s_i) + \sum_{j < i} \frac{\lambda^{(s)}_{ij}(t_i)}{\lambda^{(s)}_{ij}(t_i|\mathcal{H}_{t_{i-}})} f(x_i|t_i, s_i, e_j).$$

(4)

Thus, this mark density is in fact a mixture whose weights are proportional to the different intensity components.

Unless stated otherwise, in the remaining paper we shall focus on the marked MHP whose conditional intensity and mark density are defined in (1) and (4), respectively. The generative procedure within a maximum observation window $T$ is summarized as follows:

For $i = 1, 2, \ldots$ do

1. Draw $t_i$ and $s_i$ together from inhomogeneous Poisson processes with intensity $\lambda^{(1)}(t|\mathcal{H}_{t_{i-}}), \ldots, \lambda^{(S)}(t|\mathcal{H}_{t_{i-}})$.
2. Draw the parent variable $z_i$ according to $P(\cdot|t_i, s_i, \mathcal{H}_{t_{i-}})$.
3. Draw mark $x_i$:
   - (a) If $z_i$ corresponds to base intensity $\mu^{(s)}$, draw $x_i \sim f(\cdot|t_i, s_i)$.
   - (b) If $z_i$ corresponds to a previous event $e_j$, draw $x_i \sim f(\cdot|t_i, s_i, e_j)$.
4. If $t_i > T$, then terminate and return sample $\mathcal{H}_T = \{e_i : t_i \leq T\}$. 

3
3 Root Source Probability for Multivariate Hawkes Processes

One important property of a branching structure is that it can be viewed as a forest with $S$ trees whose roots correspond to baseline intensities $(\mu^{(s)}(t))_{i=1,\ldots,S}$ because the “parent” of each event $e_i$ is either an earlier event $e_j$ or the base intensity $\mu^{(s)}$. As a result, given a branching structure, for each event $e_i$, we can trace back to its “root” using the intrinsic parenthood relationship. Figure 1 illustrates such property.

Often, we are interested in inferring the root of each event and we would like to compute

$$P(\text{root of } e_i \text{ is on dimension } s | \mathcal{H}_T),$$

where $\mathcal{H}_t$ represents all the events occurring before and at time $t$.

Formally, suppose $n$ events are observed in time window $[0,T]$. Let the collection of parent variables $z_I \triangleq (z_i)_{i \in I}$ indicate the branching structure over events indexed by $I \subseteq [n]$, and let $\mathcal{Z}(I)$ be the space of all possible $z_I$’s. Let $\mathcal{Z}_i^{(1)}, \mathcal{Z}_i^{(2)}, \ldots, \mathcal{Z}_i^{(S)}$ be a partition for $\mathcal{Z}(I)$ with respect to event $e_i$, where

$$\mathcal{Z}_i^{(s)}(I) \triangleq \{z_I \in \mathcal{Z}(I) : e \text{ roots from } \mu^{(s)}(t) \text{ according to } z_I\}.$$  

We now define root source probability.

**Define Root Source Probability.** Given an $S$-dimensional (marked) multivariate Hawkes process with sample $\mathcal{H}_T$ of size $n$, the $s$-root probability of event $e_i$ is defined as

$$r_i^{(s)} \triangleq \sum_{z_{[n]} \in \mathcal{Z}_i^{(s)}([n])} P(z_{[n]} | \mathcal{H}_T).$$

In addition, vector $r_i \triangleq [r_i^{(1)}, \ldots, r_i^{(S)}]$ is defined as the root source probability of event $e_i$.

At first glance, the computation of root source probabilities should be difficult, as it requires summing over all the posterior probability of all branching structures in $\mathcal{Z}_i^{(s)}([n])$ whose size grows exponentially with the increase of $n$. However, the marked MHP satisfies several independence properties and consequently there is an efficient algorithm for computing the root source probabilities for all events.

The marked MHP has the following two independence properties:

1. All $z_i$’s are mutually independent conditioned on $\mathcal{H}_T$.
2. $z_{[i]}$ and $\mathcal{H}_T \setminus \mathcal{H}_{t_i}$, i.e. the future events, are independent conditioned on $\mathcal{H}_{t_i}$.

These two properties can be easily verified from the generative procedure of the process, which immediately gives the following statement: for any $i \in [n]$ and $t \geq t_i$, $P(z_{[i]} | \mathcal{H}_t) = P(z_{[i]} | \mathcal{H}_{t_i})$.

We then derive the result below for computing the root source probabilities.
Finally, since with (2) (3) (4), we have

\[ r_i^{(s)} \propto \delta_{s_i,s} \mu^{(s)}(t_i)f(x_i|t_i, s_i) + \sum_{j<i} r_j^{(s)} \lambda_j^{(s)}(t_i)f(x_i|t_i, s_i, \epsilon_j). \]  

(6)

Now we give a brief proof for the claim above.

First, the space \( Z_i^{(s)}([n]) \) can be factorized as \( Z_i^{(s)}([i]) \times Z([i+1, \ldots n]) \). By independence, we can marginalize out \( z_{i+1, \ldots n} \) and simplify \( r_i^{(s)} \) as

\[ r_i^{(s)} = \sum_{z_{[i]} \in Z_i^{(s)}([i])} P(z_{[i]}|H_{t_i}). \]  

(7)

For any \( z_{[i]} \in Z_i^{(s)}([i]) \), it can only be one of the following two cases:

1. \( z_{i0} = 1 \) and \( s_i = i \).
2. \( z_{ij} = 1 \) and \( z_{[i-1]} \in Z_j^{(s)}([i-1]) \) for some \( j < i \).

Thus we can rewrite the right hand side of (7) as

\[ \delta_{s_i,s} \sum_{z_{[i-1]} \in Z^{(s)}([i-1])} P(z_{[i-1]}|H_{t_{i-1}}) P(z_{i0} = 1|H_{t_i}) + \sum_{j<i} \sum_{z_{[i-1]} \in Z_j^{(s)}([i-1])} P(z_{[i-1]}|H_{t_{i-1}}) P(z_{ij} = 1|H_{t_i}). \]

Note that \( \sum_{z_{[i-1]} \in Z_j^{(s)}([i-1])} P(z_{[i-1]}|H_{t_{i-1}}) \) is indeed \( r_j^{(s)} \), so we have

\[ r_i^{(s)} = \delta_{s_i,s} P(z_{i0} = 1|H_{t_i}) + \sum_{j<i} r_j^{(s)} P(z_{ij} = 1|H_{t_i}). \]

Finally, since

\[ P(z_{[i]}|H_{t_i}) \propto P(z_{[i]}|t_i, s_i, H_{t_{i-1}}) P(x_i|z_i, t_i, s_i, H_{t_{i-1}}), \]

with (2) (3) (4), we have

\[ r_i^{(s)} \propto \delta_{s_i,s} P(z_{i0} = 1|t_i, s_i, H_{t_{i-1}}) f(x_i|t_i, s_i) + \sum_{j<i} r_j^{(s)} P(z_{ij} = 1|t_i, s_i, H_{t_{i-1}}) f(x_i|t_i, s_i, \epsilon_j) \]

\[ \propto \delta_{s_i,s} \mu^{(s)}(t_i)f(x_i|t_i, s_i) + \sum_{j<i} r_j^{(s)} \lambda_j^{(s)}(t_i)f(x_i|t_i, s_i, \epsilon_j). \]

4 Model Formulation and Inference

Now we turn to an application of the root source probability framework we developed in the last section: inferring the root cause of each event in multiple related text cascades. A text cascade is defined as a sequence of posting events observed within a single information source, for example, social media, news websites, blogs, and forums. It consists of a sequence of posting events and each event is defined as a triplet \((t_i, s_i, x_i)\) meaning that a post is made on source \(s_i\) at time \(t_i\) with text \(x_i\). Our goal is to learn the root source for each event.

This problem is solved in two stages: an \(S\)-dimensional marked MHP is firstly trained to model interactions between different sources, and then the root source probability \(r_i\) for each event can be directly computed using the learned model parameters with the formula in (6). We describe the formulation of the model in Section 4.1 and explain the procedure of estimating parameters in Section 4.2.
4.1 Formulation

To concretize the Hawkes model, we first propose the following factorized parameterization for the components of conditional intensities in (1):

\[
\begin{align*}
\mu^{(s)}(t) &= \rho_s \tilde{\mu}^{(s)}(t) \\
\lambda_t^{(s)}(t) &= \alpha_{s,s_i} \beta(x_i) \kappa^{(s)}(t_i, t).
\end{align*}
\]

where \( \rho_s \) and \( \tilde{\mu}^{(s)}(\cdot) \) are the multiplying scalar and the shape function for the base intensity of dimension \( s \), \( \mathbf{A} = (\alpha_{ss'})_{s,s'} \) characterizes the strength of mutual excitation between dimension pairs, function \( \beta(\cdot) \) measures the impact of marks, and \( \kappa^{(s)}(t, t') \) for \( s = 1, \ldots, S \) are decay kernels satisfying \( \int_t^\infty \kappa^{(s)}(t, t') dt' = 1 \).

Next, we propose the following generative procedure for two cases of mark densities \( f(x_i|t_i, s_i) \) and \( f(x_i|t_i, s_i, e_j) \):

1. Draw the text length \( L_i \sim \text{Poi}(|d^{(s_i)}) \).
2. For \( f(x_i|t_i, s_i) \), draw \( x_i \sim \text{Multinomial}(|L_i, \theta^{(s_i)}) \); Otherwise for \( f(x_i|t_i, s_i, e_j) \), draw \( x_i \sim \text{Multinomial}(|L_i, (1-\gamma)\theta^{(s_i)} + \gamma T) \)

Here \( \text{Poi}(d) \) and \( \text{Multinomial}(m, \theta) \) refer to Poisson and Multinomial distribution with parameters \( d \) and \( m, \theta \) respectively, and \( \bar{x} = x/\sum x \) is the normalized bag-of-words vector. Note that the case for \( f(x_i|t_i, s_i, e_j) \) actually corresponds to a word-level Multinomial mixture, which can be equivalently viewed as drawing each word independently with probability \( 1-\gamma \) from \( \text{Categorical}(\theta^{(s_i)}) \) and with probability \( \gamma \) from \( \text{Categorical}(\bar{x}_j) \).

The graphical model of this marked Hawkes process is included in the supplementary material.

4.2 Parameter Estimation

We estimate the major parameters \( \Theta = \{\rho, \mathbf{A}, \theta, \gamma, d\} \) by maximizing the marginal likelihood on \( \mathcal{H}_T \):

\[
\max_\Theta P(\mathcal{H}_T|\Theta)
\]

However, \( P(\mathcal{H}_T|\Theta) \) is intractable due to exponential number of hidden \( z_{[n]} \). To solve this, we adopt the variational EM method [Bernardo et al., 2003], a variation of EM methods, which approximates \( P(z_{[n]}|\mathcal{H}_T, \Theta) \) with a proposed distribution \( Q \) and maximizes a lower bound surrogate \( \log P(\mathcal{H}_T|\Theta) \) over \( \Theta \) and \( Q \) alternatively. In details, given the event sample \( \mathcal{H}_T \) and branching structure \( z_{[n]} \), the complete likelihood is [Rasmussen, 2013]

\[
P(\mathcal{H}_T, z_{[n]}|\Theta) = \exp \left( - \int_0^T \sum_s \lambda^{(s)}(t|\mathcal{H}_T dt) \right)
\times \prod_{i=1}^n \left( \mu^{(s_i)}(t_i) f(x_i|t_i, s_i) \right)^{z_{i0}} \times \prod_{i=1}^n \prod_{j<i} \left( \lambda^{(s_i)}(t_i) f(x_i|t_i, s_i, e_j) \right)^{z_{ij}}
\]

Utilizing a fully-factorized proposed distribution

\[
Q(z_{[n]}) = \prod_{i=1}^n \text{Multinomial}(z_i|\eta_i),
\]

we are able to construct a lower-bound surrogate of \( \log P(\mathcal{H}_T|\Theta) \) using evidence lower bound (ELBO)

\[
\log P(\mathcal{H}_T|\Theta) \geq \mathbb{E}_Q[\log P(\mathcal{H}_T, z_{[n]}|\Theta)] - \mathbb{E}_Q[\log Q(z_{[n]})] \\
\triangleq \mathcal{L}(\Theta, Q).
\]


The detailed expression of $\mathcal{L}(\Theta, Q)$ can be found in the supplementary material. Note that similar tricks have been adopted in previous stochastic process literatures [He et al. 2015, Yang and Zha 2013, Hoffman et al. 2013].

**Update $\rho, A$** By maximizing [12] with respect to $\rho$ and $A$, we obtain the following closed-form updates:

$$
\rho_s = \sum_{i=1}^n \delta_{s_i,s} \eta_{0} \int_0^T \bar{\mu}(s')(t') dt',
$$

(13)

$$
\alpha_{ss'} = \frac{\sum_{i=1}^n \delta_{s_i,s} \delta_{s_j,s'} \eta_{ij}}{\sum_{i=1}^n \delta_{s_i,s} \beta(s_i) \int_{t_i}^T \kappa(s_i)(t,r) dr}.
$$

(14)

**Update $\eta$** Similarly, the updates for $Q$, i.e. $\eta$, by maximizing the relevant parts in [12], are

$$
\eta_{0} \propto \mu(s_i(t_i)|x_i, s_i),
$$

(15)

$$
\eta_{ij} \propto \lambda(s_i(t_i)|x_i, s_i, e_j).
$$

(16)

**Update $\theta$ and $\gamma$** The updates for $\theta$ and $\gamma$ are given by solving the following sub-optimization problem:

$$
\max_{\theta, \gamma} \sum_{s=1}^S \sum_{i=1}^n \sum_{v=1}^V \delta_{s_i,s} \theta_{s_i,v}(\theta, \gamma) \quad \text{s.t.} \quad \sum_{v=1}^V \theta_{s_i,v} = 1 \quad \forall s,
$$

(17)

where

$$
g_{s_i,v}(\theta, \gamma) = \eta_{0} x_{i,v} \log \theta_{s_i,v} + \sum_{j<i} \eta_{ij} x_{j,v} \log \left((1 - \gamma)\theta_{s_i,v} + \gamma \bar{x}_{j,v}\right)
$$

(18)

Our overall strategy for optimizing [17] is to first apply Jensen’s inequality to each logarithm term in the summation of [18] with a coefficient $\xi_{s_i,j,v} \in (0,1)$ so that a lower bound can be constructed for each $g_{s_i,v}(\theta, \gamma)$, and then to optimized the summed lower bounds instead. Specifically, let $\hat{\theta}_{s_i,v}$ and $\hat{\gamma}$ be the current estimate of $\theta_{s_i,v}$ and $\gamma$, respectively. One can show that with a choice of $\xi_{s_i,j,v}$ to be

$$
\xi_{s_i,j,v} = \frac{\hat{\gamma} \bar{x}_{j,v}}{(1 - \hat{\gamma})\hat{\theta}_{s_i,v} + \hat{\gamma} \bar{x}_{j,v}},
$$

(19)

this optimization procedure yields a “Minorization-Maximization” algorithm, and the correspondent closed-form updates are

$$
\theta_{s_i,v} \propto \sum_i \delta_{s_i,s} [\eta_{0} x_{i,v} + \sum_{j<i} \eta_{ij} (1 - \xi_{s_i,j,v}) x_{j,v}],
$$

(20)

$$
\gamma = \frac{\sum_{i=1}^n \sum_{j<i} \sum_{v=1}^V \eta_{ij} x_{j,v} \xi_{s_i,j,v}}{\sum_{i=1}^n \sum_{j<i} \sum_{v=1}^V \eta_{ij} x_{i,v}}.
$$

(21)

The lower bound of $g_{s_i,v}(\theta, \gamma)$ and other details are in the appendix.
4.3 Empirical Bayes

In real-world settings where experts can provide intuition about the data, it is advantageous to include prior information in parameter estimation. Specifically, Bayesian priors $\pi(\rho_s)$ and $\pi(\alpha_{s,s'})$ on each source’s baseline event-rate and offspring event-rate improve model performance significantly.

We recommend independent Gamma priors $\rho_s \sim \text{Gamma}(a_s, b_\rho)$ and $\alpha_{s,s'} \sim \text{Gamma}(a_\alpha, b_\alpha)$.

The updates for $\rho_s$ and $\alpha_{s,s'}$ are then:

$$\rho_s = \frac{a_{\rho}^{(s)} - 1 + \sum_{i=1}^{n} \delta_{s_i,s} \eta_{0i}}{b_\rho + \int_{0}^{T} \bar{\mu}^{(s)}(t') dt'}$$

$$\alpha_{s,s'} = \frac{a_{\alpha}^{(s)} - 1 + \sum_{i=1}^{n} \sum_{j<i} \delta_{s_i,s} \delta_{s_j,s'} \eta_{ij}}{b_\alpha + \sum_{i=1}^{n} \delta_{s_i,s'} \beta(x_i) \int_{t_i}^{T} k^{(s)}(t_i,r) dr}.$$ (23)

5 Experiments

5.1 Synthetic Data

We first evaluate our model on a synthetic dataset to validate our parameter estimation algorithm and the ability of root source probabilities in identifying meaningful root sources. The dataset has $S = 5$ dimensions with base intensities $\mu^{(s)}(t) = 1$ and $\rho_s = 0.1$, and excitation kernel functions $k^{(s)}(t,t') = \frac{1}{10} \exp\left\{-\frac{1}{10}(t'-t)\right\}$ for all $s$. The influence matrix $\mathbf{A}$ has diagonal entries of 0.4 and off-diagonal entries of 0.1, and the mark impact function $\beta(x)$ is set to constant 1. The five dimensions have average text lengths of 10, 20, 30, 40 and 50, respectively, a constant vocabulary mixing rate of $\gamma = 0.3$, and word distribution parameters $\theta^{(s)} \sim \text{Dir}(1)$. The total vocabulary size is $V = 5000$.

5.1.1 Parameter Recovery

Figure 2a shows relative mean square errors (RMSE) for each source vocabulary parameter $\theta^{(s)}$. RMSEs decrease as the sample size increases. Also, RMSEs are smaller for dimensions with longer text lengths.

RMSE is also used to quantify the accuracy of parameter recovery for $\mathbf{A}$, the mutual excitation matrix:

$$\epsilon(\hat{\mathbf{A}}) = \|\hat{\mathbf{A}} - \mathbf{A}\|_F^2 / \|\mathbf{A}\|_F^2$$

where $\|\cdot\|_F$ is the Frobenius norm for matrices. In Figure 2b we verify that $\epsilon(\hat{\mathbf{A}})$ drops with more training events.

Parameter recovery results are robust to multiple runs of our inference algorithm.

5.1.2 Root Source Identification

To our knowledge, there is no existing literature providing root source predictions in settings where tree structure is unobserved. Therefore, we compare the ability of our computation algorithm ($\text{RP}_\text{FIT}$) for root source probabilities in recovering the sample root sources against the following baselines:

- Sub-model baselines: root source probability with temporal info only ($\text{RP}_\text{TEMP}_\text{FIT}$) and with textual info only ($\text{RP}_\text{MARK}_\text{FIT}$), whose root source probabilities are calculated using the follow-
We use the comparison against these two sub-model baselines as an ablation study for our model.

- Supervised baselines: Naive Bayes multi-class classifier trained with $c \in (0, 1)$ proportion of the data. We set $c = 0.5$ (NB\_0.5) or $0.7$ (NB\_0.7).

- Heuristic running window baselines: use the normalized source counts over the $M$ most recent events as an estimate of root source probabilities. We set $M = 1$ (RW\_1), 10 (RW\_10), or $+\infty$ (RW\_inf).

We also compute the root source probabilities with ground truth parameters known and take it as a new baseline (RP\_TRUE).

We compare the identified roots against true roots to measure how accurately each method recovers the real root sources. For NB methods, we only consider the accuracy on remaining events since the labels in the training set are given. Figure 3 shows the accuracies of different methods. The RP-based methods outperform all other baselines in all cases, and as the sample size increases RP\_FIT converges to RP\_TRUE, the Bayes Optimal.

Figure 3: Accuracy of root source identification.
5.1.3 Inference Algorithm Scalability

Algorithm running time is recorded in the synthetic data experiments. As shown in Figure 4, model training time scales linearly with the number of events.

![Figure 4: Training time scales linearly with the number of events.](image)

5.2 Real Data Experiments

We then apply our model to two real-world datasets:

- **Reddit**: Comments collected from `reddit.com`, under the political article titled “2016 Election Day Returns Megathread”. Each comment is either an original comment, or a response to a previous comment. The root source of each comment is defined as the author of its first level ancestral comment.

- **12 Angry Men**: A transcript from the 1957 legal-themed film, *12 Angry Men*, where the first approximately 600 utterances in the movie are considered. The root source of each utterance is defined as the speaker (juror in the film) of the utterance that initiated the thread of conversation that leads to this utterance.

In the Reddit dataset, the branching structure among comments is already known but omitted in the model training stage, and is used for model evaluation. In contrast, *12 Angry Men* dataset does not have explicit replying structure to serve as ground-truth. We manually labeled the data as a human intuition reference against model outputs.

5.2.1 Hyper-parameter Specification

We choose the baseline intensity shape function as \( \bar{\mu}^{(s)}(t) \equiv 1 \) and excitation kernel \( \kappa^{(s)}(t, t') = \frac{1}{\nu} \exp\left(-\frac{1}{\nu}(t' - t)\right) \), where \( \nu = 450 \) for Reddit and \( \nu = 8 \) for 12 Angry Men, based on our observation on the decaying replying rate in each dataset. For the empirical Bayes priors, we set \( a_p^{(s)} = N_s, b_p = T/c \), \( a_\alpha^{(s)} = N_s \) and \( b_\alpha = T/(1 - c) \), where \( T \) is the total time, \( N_s \) is the total event count on source \( s \), and \( c = 1/10 \), based on our observation that approximately 1/10 of the utterances are original comments.

5.2.2 Quantitative Model Evaluation

For the Reddit data, where the replying structure is available for calculating root source identification accuracies (such structure is ignored in model training), we can evaluate model performance through log conditional probabilities, accuracies, and top-10 accuracies by comparing the identified root sources with the actual authors of root comments. The results are presented in Table 1. To our knowledge, no existing works directly identify root sources. Therefore, we compare the model with the three
running window baselines introduced in Section 5.1.2 along with a Uniform baseline, which assigns

equal probabilities to all root sources and guesses one uniformly at random.

Our model, RP\_FIT, attains the highest log probability and highest accuracy. The baseline RW\_1

has high accuracy only because for a large number of comments the author is the actual root source,

but this baseline is outperformed by RP\_FIT in the other two assessments.

| Method       | Log Prob. | Acc.  | Top-10 Acc. |
|--------------|-----------|-------|-------------|
| RP\_FIT      | -852.15   | 0.74  | 0.79        |
| RW\_inf      | -1807.84  | 0.04  | 0.30        |
| RW\_10       | -2807.71  | 0.11  | 0.77        |
| RW\_1        | -2300.28  | 0.74  | 0.77        |
| Uniform      | -2111.18  | 0.00  | 0.07        |

Table 1: Performance of various methods

5.2.3 Qualitative Model Evaluation

In most of the real-world datasets for group conversation, unfortunately, neither the direct response

relationship nor the root sources are available for constructing the root source labels as ground-truth.

Therefore, it is also critical to examine the goodness-of-fit of the model through comprehensive quali-

tative analysis. In this part, we provide three types of qualitative analysis for our model, to examine

the “power” of users, the reciprocal relationship among users, and word-inheritance captured by our

model.

The primary outputs of the model are the comment specific root source probabilities \(\{r_i^1, ..., r_i^{(S)}\}_{i=1}^N\),

giving the predicted probabilities that each speaker \(s\) is the root source for each comment \(i\). The sum

over all events of the root source probabilities attributed to source \(s\), i.e. \(\sum_{i=1}^N r_i^{(s)}\), quantifies the

ability of source \(s\) to initiate conversations, providing a natural measure of influence exerted by a

speaker in a group conversation.

In Table 2a, we sort speakers in 12 Angry Men using this measure. Similar to the findings in the

influence matrix, Juror 8 is identified as the most active conversation initiator, who in fact proposed an

evidence inspection and multiple in-depth discussions in the film. The model also identified arguably

the second most influential character, Juror 3, the antagonist who is last to change his vote to “not

guilty”.

For the Reddit data, we also rank the forum users by this metric developed from root source

probabilities. Table 2b presents the 5 most influential Redditors ranked by such metric and compares

it with the rankings using total Reddit Gold (approval votes received by a Reddit user). To protect

privacy, we have replaced usernames with code names. The most influential users by our metric also

receive relatively high rankings by total Reddit Gold. This shows root source probabilities can act as

a proxy for the popularity of an author in an online social network. Root source probability provides

additional value because it quantifies an author’s potential of exciting future comments, measuring

the indirect impact of that author on the social network, while Reddit Gold only addresses the direct

impact caused by the author’s posts.

Another meaningful output of the analysis is the influence matrix \(A = \{\alpha_{ss'}\}_{ss'}\), quantifying the

impact strength of a comment by source \(s'\) on the comment rate of source \(s\). We threshold and plot

the resulting network of interactions between sources for 12 Angry Men in Figure 5. Edge darkness is

proportional to influence, and vertex size is proportional to baseline comment rate. Influential authors

have multiple outgoing edges. The model successfully identified Juror 8, the movie’s protagonist, as

an influential person. In the movie, he is the only juror initially voting “not guilty”, and by the end,

all other jurors change their votes.
Table 2: Five most influential users in two real-world datasets, where “Role” explains the stance each juror holds in the film, and “Gold” refers to approval votes received by a Reddit user.

(a) 12 Angry Men

| Rank | Source | Influence | Role          |
|------|--------|-----------|---------------|
| 1    | Juror 8 | 269.89    | insists “not guilty” |
| 2    | Juror 3 | 53.34     | insists “guilty”    |
| 3    | Juror 7 | 40.29     |                |
| 4    | Juror 1 | 36.99     | serves as Foreman |
| 5    | Juror 10| 35.78     |                |

(b) Reddit

| Rank | Username | Influence | Gold | Gold Rank |
|------|----------|-----------|------|-----------|
| 1    | User E   | 15.94     | 385  | 2         |
| 2    | User S1  | 11.66     | 72   | 17        |
| 3    | User S2  | 10.38     | 44   | 28        |
| 4    | User S3  | 10.32     | 73   | 16        |
| 5    | User R   | 10.20     | 73   | 15        |

Figure 5: Influence network of 12 Angry Men from the influence matrix.

Figure 6 shows word inheritance in comments whose identified root sources agree with our manual labels in the 12 Angry Men data. We notice inheritance of the word “witnesses”, the phrase “could they be wrong”, and the word “people”, verifying the model assumption that a response comment may borrow vocabulary from the parent comment. A similar word inheritance plot is contructed on the Reddit data in the appendix.

6 Conclusion

We introduced a novel model for group conversations, based on a multivariate Hawkes process. The main practical applications are two-fold. First, the model learns a source-pair influence matrix which can be used to infer directed power relations among sources and create network visualizations. The second contribution is an algorithm to estimate root source probabilities for each event, given event time-stamps, source labels, and text marks.

We experiment with the model on synthetic and real world data. On synthetic data, the model clearly outperforms its subset baselines and a naive Bayes classifier in terms of accuracy for identifying root sources. On a real world data set from reddit.com where ground truth root sources are used for calculating accuracies, the final model attains the highest log probability and highest accuracy when compared with some baselines. This performance suggests the value of our model in settings where structure is unobserved. We apply the model to a group conversation transcription from the movie 12
Angry Men, and find the model identifies root sources which agree with human intuition in exchanges where linguistic accommodation occurs.

Beyond this work, a possible extension considers the entire ancestral path, over all authors of parent comments, back to the root author. One variation on our algorithm is to learn a probability distribution over root-comments, rather than root-sources. An alternative to our unsupervised approach is to explicitly model structural information (i.e. branching structure), to train a model in a supervised manner, and to predict missing structural data. A final possible extension explicitly models the changing character of conversation as it evolves, accounting for decaying comment rates as the conversation becomes outdated.

References

J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. David, D. Heckerman, A. F. M. Smith, M. West, and Others. The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures. *Bayesian Statistics*, 7:453–464, 2003.

C. Blundell, J. Beck, and K. A. Heller. Modelling Reciprocating Relationships with Hawkes Processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012. ISBN 9781627480031.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012. ISBN 9781450312295. doi:10.1145/2187836.2187931.

N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable Influence Estimation in Continuous-Time Diffusion Networks. *Advances in Neural Information Processing Systems*, pages 3147–3155, 2013. ISSN 10495258.

P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes Processes: An Application to Financial Data. *Journal of Applied Probability*, 48(A):367–378, 2011.
M. Farajtabar, M. Gomez-Rodriguez, M. Zamani, N. Du, H. Zha, and L. Song. Back to the Past: Source Identification in Diffusion Networks from Partially Observed Cascades. In AISTATS, volume 38, 2015a.

M. Farajtabar, Y. Wang, M. G. Rodriguez, S. Li, H. Zha, and L. Song. COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution. Advances in Neural Information Processing Systems, pages 1–20, jul 2015b. ISSN 10495258.

F. Guo, C. Blundell, H. Wallach, and K. Heller. The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, volume 38, page 14, nov 2015. URL [http://arxiv.org/abs/1411.2674](http://arxiv.org/abs/1411.2674).

A. G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. Biometrika, 58(1):83–90, 1971a. ISSN 00063444. doi: 10.1093/biomet/58.1.83.

A. G. Hawkes. Point Spectra of Some Mutually Exciting Point Processes. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):83–90, 1971b.

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. Proceedings of the 32nd International Conference on Machine Learning, 37, 2015.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic Variational Inference. The Journal of Machine Learning Research, 14(1):1303–1347, 2013.

J. F. C. Kingman. Poisson Processes (Oxford studies in probability; 3). Clarendon Press, 1993.

S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. ICML, 32:1413–1421, feb 2014.

J. G. Rasmussen. Bayesian Inference for Hawkes Processes. Methodology and Computing in Applied Probability, 15(3):623–642, dec 2013. ISSN 13875841. doi: 10.1007/s11009-011-9272-5.

Y. Rong, H. Cheng, and Z. Mo. Why It Happened: Identifying and Modeling the Reasons of the Happening of Social Events. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15, pages 1015–1024, 2015. ISBN 9781450336642. doi: 10.1145/2783258.2783305.

X. Tan, S. A. Z. Naqvi, A. Y. Y. Y. Qi, K. A. Heller, and V. Rao. Content-based Modeling of Reciprocal Relationships using Hawkes and Gaussian Processes. In UAI, page 2016, 2016. ISBN 9781510827806.

S. Yang and H. Zha. Mixture of Mutually Exciting Processes for Viral Diffusion. In Journal of Machine Learning Research, volume 28, pages 1–9, 2013.

K. Zhou, H. Zha, and L. Song. Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes. In Artificial Intelligence and Statistics, pages 641–649, 2013. ISBN 9783662448502. doi: 10.1007/978-3-662-44851-9_11.
Appendices

A Graphical Model for Marked Multivariate Hawkes Processes

B Details for Parameter Estimation

B.1 Log-likelihood and Variational Lower Bound

The complete likelihood is

\[
P(H_T, z[n] | \Theta) = \exp \left( - \int_0^T \sum_s \lambda^{(s)}(t|H_{t-}) dt \right) \prod_{i=1}^n \lambda^{(s_i)}(t_i|H_{t_i-}) \times \prod_{i=1}^n \left\{ \left( \frac{\mu^{(s_i)}(t_i)}{\lambda^{(s_i)}(t_i|H_{t_i-})} f(x_i|t_i, s_i) \right)^{z_{i0}} \times \prod_{j<i} \left( \frac{\lambda^{(s_i)}(t_i)}{\lambda^{(s_j)}(t_i|H_{t_i-})} f(x_i|t_i, s_i, e_j) \right)^{z_{ij}} \right\}.
\]

Then complete log likelihood with the parameterization specified in the paper becomes:

\[
\log P(H_T, z[n] | \Theta) = - \sum_s \rho_s \int_0^T \bar{\mu}^{(s)}(t) dt
- \sum_s \sum_{i=1}^n \alpha_{s,s_i} \beta(x_i) \int_{t_i}^T \kappa(t, t) dt
+ \sum_{i=1}^n z_{i0} \log \left[ \rho_s \bar{\mu}^{(s_i)}(t_i) f(x_i|t_i, s_i) \right]
+ \sum_{i=1}^n \sum_{j<i} z_{ij} \log \left[ \alpha_{s_i,s_j} \kappa(t_j, t_i) f(x_i|t_i, s_i, e_j) \right].
\]
The variational lower bound \( \mathcal{L}(\Theta, Q) \) is
\[
\mathcal{L}(\Theta, Q) \triangleq \mathbb{E}_Q[\log P(H_T, z_{[n]} | \Theta)] - \mathbb{E}_Q[\log Q(z_{[n]})] \\
= - \sum_s \rho_s \int_0^T \bar{\mu}^{(s)}(t) dt \\
- \sum_s \sum_{i=1}^n \alpha_{s, s_i} \beta(x_i) \int_{t_i}^T \kappa(t, t_i) dt \\
+ \sum_{i=1}^n \eta_{i0} \log \left[ \rho_s \bar{\mu}^{(s)}(t_i) f(x_i | t_i, s_i) \right] \\
+ \sum_{i=1}^n \sum_{j<i} \eta_{ij} \log \left[ \alpha_{s_i, s_j} \kappa(t_j, t_i) f(x_i | t_i, s_i, e_j) \right] \\
- \sum_{i=1}^n \left( \eta_{i0} \log \eta_{i0} + \sum_{j<i} \eta_{ij} \log \eta_{ij} \right). \tag{26}
\]

**B.2 Derivation of Updates with Empirical Bayes Priors**

The modified lower bound using Gamma priors \( \rho_s \sim \text{Gam}(a^{(s)}_\rho, b^{(s)}_\rho) \) and \( \alpha_{s, s'} \sim \text{Gam}(a^{(s)}_\alpha, b^{(s)}_\alpha) \) is
\[
\tilde{\mathcal{L}}(\Theta, Q) \\
= - \sum_s \rho_s \int_0^T \bar{\mu}^{(s)}(t) dt \\
- \sum_s \sum_{i=1}^n \alpha_{s, s_i} \beta(x_i) \int_{t_i}^T \kappa(t, t_i) dt \\
+ \sum_{i=1}^n \eta_{i0} \log \left[ \rho_s \bar{\mu}^{(s)}(t_i) f(x_i | t_i, s_i) \right] \\
+ \sum_{i=1}^n \sum_{j<i} \eta_{ij} \log \left[ \alpha_{s_i, s_j} \kappa(t_j, t_i) f(x_i | t_i, s_i, e_j) \right] \\
- \sum_{i=1}^n \left( \eta_{i0} \log \eta_{i0} + \sum_{j<i} \eta_{ij} \log \eta_{ij} \right) \\
+ \sum_s \left[ (a^{(s)}_\rho - 1) \log \rho_s - b^{(s)}_\rho \rho_s \right] \\
+ \sum_s \sum_{s'} \left[ (a^{(s)}_\alpha - 1) \log \alpha_{s, s'} - b^{(s)}_\alpha \alpha_{s, s'} \right].
\]

Taking first order derivatives regarding \( \rho_s \) and \( \alpha_{s, s'} \) gives the modified updates
\[
\rho_s = \frac{a^{(s)}_\rho - 1 + \sum_{i=1}^n \delta_{s, s_i} \eta_{i0}}{b^{(s)}_\rho + \int_0^T \bar{\mu}^{(s)}(t) dt} \tag{27}
\]
\[
\alpha_{s, s'} = \frac{a^{(s)}_\alpha - 1 + \sum_{i=1}^n \sum_{j<i} \delta_{s_i, s_j} \delta_{s, s'} \eta_{ij}}{b^{(s)}_\alpha + \sum_{i=1}^n \delta_{s, s'} \beta(x_i) \int_{t_i}^T \kappa^{(s)}(t, t_i) dt} \tag{28}
\]
C More Experiment Results on Reddit Data

Figure 8 shows word inheritance in the text of Reddit comments. We visualize comments whose root sources are correctly identified by our model. Note inheritance of the word “DNC” in the first branch and the statement “Hillary is president” in the second branch.

Figure 8: Word inheritance in the Election Day Megathread Comments.