LATENT SELF-EXCITING POINT PROCESS MODEL FOR SPATIAL-TEMPORAL NETWORKS

Yoon-Sik Cho

USC Information Sciences Institute
Marina del Rey, CA 90292, USA

Aram Galstyan[1], P. Jeffrey Brantingham[2] and George Tita[3]

[1]USC Information Sciences Institute, USA
[2]University of California, Los Angeles, USA
[3]University of California, Irvine, USA

Abstract. We propose a latent self-exciting point process model that describes geographically distributed interactions between pairs of entities. In contrast to most existing approaches that assume fully observable interactions, here we consider a scenario where certain interaction events lack information about participants. Instead, this information needs to be inferred from the available observations. We develop an efficient approximate algorithm based on variational expectation-maximization to infer unknown participants in an event given the location and the time of the event. We validate the model on synthetic as well as real-world data, and obtain very promising results on the identity-inference task. We also use our model to predict the timing and participants of future events, and demonstrate that it compares favorably with baseline approaches.

1. **Introduction.** In recent years there has been a considerable interest in understanding dynamic social networks. Traditionally, longitudinal analysis of social network data has been limited to relatively small amounts of data collected from manual and time-consuming surveys. Recent development of various sensing technologies, online communication services, and location-based social networks has made it possible to gather time-stamped and geo-coded data on social interactions at an unprecedented scale. Such data can potentially facilitate better and more nuanced understanding of geo-temporal patterns in social interactions. To harness this potential, it is imperative to have efficient computational models that can deal with spatial-temporal social networks.

One of the main challenges in social network analysis is handling missing data. Indeed, most social network data are generally incomplete, with missing information about links [15, 20, 21], nodes [10] or both [19]. In repeated interaction networks studied here, there is another source of data ambiguity that comes from limited observability of certain interaction events. Namely, even when interactions are recorded, information about participants might be missing or only partially known. A real-world problem that highlights the latter scenario is concerned with inter-gang
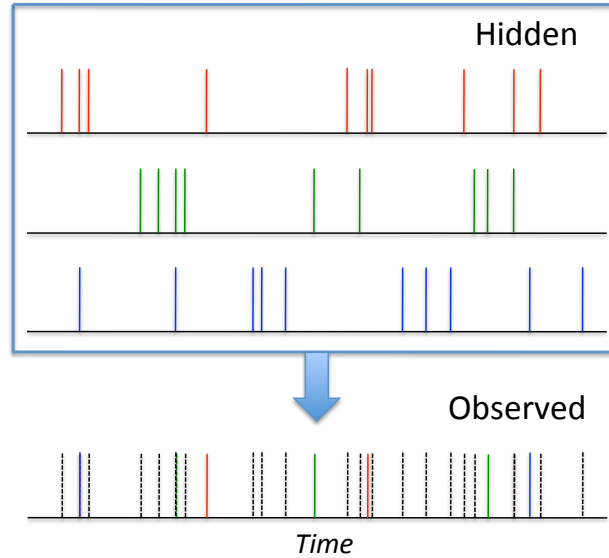

FIGURE 1. Schematic demonstration of the missing label problem for temporal point processes. The dashed lines represent events for which the generating process is unknown.

rivalry network in Los Angeles, where the records of violent events between rival gangs might lack information about one or both participants [33].

Here we formalize the missing label problem for spatial-temporal networks by introducing Latent Point Process Model, or LPPM, to describe geographically distributed interaction events between pairs of entities. LPPM assumes that interaction between each pair is governed by a spatial-temporal point process. In contrast to existing models [16, 26, 29, 30], however, it allows a non-trivial generalization where certain attributes of those events are not fully observed. Instead, they need to be inferred from available observations. To illustrate the problem, consider a sequence of events generated by M temporal point processes; see Figure 1. Each sequence is generated via a non-homogenous and possibly history-dependent point process. The combined time series is a *marked* point process, where the mark, or the label, describes the component that generates the event. The observed data consists of the recorded events. However we assume that those labels are only partially observable, and need to be inferred from the observations.

How well can one identify the label of a specific event based on limited observations? The answer depends on the nature of the process generating the events. For instance, if the events in Figure 1 are generated by a set of independent and homogenous Poisson processes with intensities $\lambda_1 < \lambda_2 < \lambda_3$, then the identification accuracy is limited by $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}$, i.e., all the unlabeled events are attributed to the process with the highest intensity. Luckily, most real-world processes describing human interactions demonstrate highly non-homogenous and history-dependent temporal patterns, suggesting that interaction events are not statistically independent, but exhibit non-trivial correlations [3, 32]. To account for temporal correlations, here we augment LPPM with a model of *self-exciting point process* known as Hawkes

process, which has been used previously in a number of applications. Furthermore, we use interaction-specific mixture distributions of spatial patterns of interactions to inform the inference problem.

Learning and inference with LPPM constitutes inferring missing labels, predicting the timing and/or the source of the next event, and so on. Due to missing observations, exact inference and learning is intractable for even moderately large datasets. Toward this end, we develop an efficient algorithm for learning and inference based on the variational EM approach [4]. We validate our model for both synthetic and real-world data. For the latter, we use two distinctly different datasets (1) data on inter-gang violence from Los Angeles Police Department; (2) User check-in data from Gowalla, which is a location based social networking service. Our results indicate that LPPM is better than baselines in both inference and prediction tasks.

The rest of the paper is organized as follows: After reviewing some relevant work in Section 2, we define our latent point process model in Section 3. In Section 4 we describe variational EM approach for efficient learning and inference with LPPM. We present results of experiments with both synthetic and real-world data in Sec. 5 and 6, and provide some concluding remarks in Section 7.

2. **Related work.** Modeling temporal social networks has attracted considerable interest in recent years. Both discrete-time [17] and continuous-time [11, 31, 37] models have been proposed to study longitudinal networks. In particular, Perry and Wolfe [26] suggested point process models for describing repeated interactions among a set of nodes. They used a Cox hazard model and allowed the interaction intensity to depend on the history of interactions as well as on node attributes. In contrast to our work, however, Ref. [26] assumes that all the interactions are perfectly observable. Different continuos time models such as Poisson Cascades [30], Poisson Networks [29], and Piecewise-Constant Conditional Intensity Model [16], have also been used to describe temporal dependencies between events.

In a related line of research, a number of authors have addressed the problem of uncovering hidden networks that facilitate information diffusion and/or activation cascades, based on time-resolved traces of such cascades. Most of the existing approaches rely on temporal information only [12, 13, 14, 8, 35], although several other methods also utilize additional features, such as prior knowledge about the network structure [23], or the content diffusing through the network [38, 36].

Self-exciting point process was originally suggested in seismology to model aftershock of earthquakes [24]. Self-exciting models have since been used in a number of diverse applications, such as assessing financial portfolio credit risk [9], detecting terrorist activities [28], predicting corporate defaults [2]. Recently, Mohler et al. [22] used spatial-temporal self-exciting process to model urban crime. Their model, however, studies a different problem and does not assume any missing information. In particular, they consider a univariate point process, as opposed to multi-variate model used here, which is needed to describe interactions among different entities.

Stomakhin et al. [33] studied the temporal data reconstruction problem in a very similar settings. Their approach, however, assumes known model parameters, which is impractical in real-world scenario, thus limiting their experiments to synthetically generated data only. In contrast, LPPM learns the model parameters directly from the data labeled or unlabeled. More recently, Hegemann et al. [18] proposed a method that does not assume known parameters but learns those parameters using an iterative scheme. The main difference between LPPM and Ref. [18] is that the

former is a generative probabilistic model, which allows to estimate the *posterior probability* that a certain pair is involved in a given event based on the observations. Ref. [18], on the other hand, calculates heuristic score-functions that can be used to rank different pairs' involvement as more or less likely, but those scores cannot be interpreted as probabilities. Furthermore, in contrast to Ref. [18] , here we consider both temporal and spatial components, and use the generative model for event prediction.

3. **Spatial-temporal model of relationship network.** Consider N individuals forming M pairs that are engaged in pairwise interactions with each other. Generally M would be the total number of undirected edges that N has, which is $N(N-1)/2$. However, for some cases (i.e., the network structure is given or some pairs are out of our consideration) total number of pairs M could be fixed to the size of our interest for efficient computation. We observe a sequence of interaction events (called events hereafter) given as $\mathcal{H} = \{h_k\}_{k=1}^n$, where each event is a tuple $h_k = (t_k, \mathbf{x}_k, z_k)$. Here $t_k \in \mathbb{R}^+$ and $\mathbf{x}_k \in \mathbb{R}^2$ are the time and the location of the event, while z_k is the symmetric interaction matrix for the event k: $z_k^{ij} = 1$ the agents i and j are involved in the k-th event, and $z_k^{ij} = 0$ otherwise. Since each event involves only one pair of agents, we have $\sum_{i<j} z_k^{ij} = 1$. Without loss of generality, we assume $t_1 = 0$ and $t_n = T$.

Let $\mathcal{H}_t$ denote the history of events up to time t as the set of all the events that have occurred before that time, $\mathcal{H}_t = \{h_k\}_{t_k < t}$. We assume that the interactions between the pairs are point processes with spatial-temporal conditional intensity function $S_{ij}(t, \mathbf{x}|\mathcal{H}_t)$, so that the probability that the agents i and j will interact within a time window $(t, t+d)]$ and location $(\mathbf{x}, \mathbf{x} + d\mathbf{x})$ is simply $S_{ij}(t, \mathbf{x}|\mathcal{H}_t)dtd\mathbf{x}$. Note that the intensity function is conditioned on the history of past events. Here we assume that the above intensity function can be factorized into temporal and spatial components as follows:

$$S_{ij}(t, \mathbf{x}|\mathcal{H}_t) = \lambda_{ij}(t|\mathcal{H}_t)r_{ij}(\mathbf{x}) \tag{1}$$

The intensity function $S_{ij}(\cdot)$ is based on the separability of spatio-temporal covariance functions [7] assuming that the temporal evolution proceeds independently at each spatial location, but rather on the history of its own. Note that the temporal conditional intensity $\lambda_{ij}(t|\mathcal{H}_t)$ is history-dependent, whereas the spatial component is not. The scope of our research is not the influence of spatial preference between nodes, but rather the spatial activities of pairs. In this regard, we assume that the pair's preference of location stays the same over time. Let us define

$$\Lambda_{ij}^T = \int_0^T \lambda_{ij}(\tau|\mathcal{H}_\tau)d\tau \tag{2}$$

The likelihood of an observed sequence of interactions under the above model is given as

$$\mathbf{p}(\mathcal{H}; \Theta) = \prod_k \prod_{i<j} \underbrace{[\lambda_{ij}(t_k)]^{z_k^{ij}} e^{-\Lambda_{ij}^T}}_{\text{temporal process}} \underbrace{[r_{ij}(\mathbf{x}_k)]^{z_k^{ij}}}_{\text{spatial process}} \tag{3}$$

where the products are over all the events and the pairs, respectively. Here Θ encodes all the hyperparameters of the model (to be specified below). From this point, we simplify the intensity expression to $\lambda_{ij}(t)$ omitting $\mathcal{H}_t$.

So far our description has been rather general. Next we have to specify a concrete parametric form of the temporal and spatial conditional intensity functions. As

stated above, the consideration of non-trivial temporal correlations between the events suggest that it is not realistic to use Poisson process with constant intensity. Instead, here we will use a Hawkes Process, which is a variant of a self-exciting process [24].

3.1. **Hawkes process.** We assume that the intensity of events involving the pair (i, j) at time t is given as follows:

$$\lambda_{ij}(t) = \mu_{ij} + \sum_{p:t_p<t} g_{ij}(t - t_p) \tag{4}$$

where the summation in the second term is over all the events that have happened up to time t. In Equation 4, μ_{ij} describes the background rate of event occurrence that is time-independent, whereas the second term describes the self-excitation part, so that the events in the past increase the probability of observing another event in the (near) future. We will use a two-parameter family for the self-excitation term:

$$g_{ij}(t - t_p) = \beta_{ij}\omega_{ij} \exp\{-\omega_{ij}(t - t_p)\} \tag{5}$$

Here β_{ij} describes the weight of the self-excitation term (compared to the background rate), while ω_{ij} describes the decay rate of the excitation.

3.2. **Spatial Gaussian Mixture Model (GMM).** To model the spatial aspect of the interactions, we assume that different pairs might have different geo-profiles of interactions. Namely, we assume that the interaction of specific pair is spatially distributed according to a pair-specific Gaussian mixture model:

$$r_{ij}(\mathbf{x}) = \sum_{c=1}^{C} w_{ij}^c \mathcal{N}(\mathbf{x}; \mathbf{m}_{ij}^c, \Sigma_{ij}^c) \tag{6}$$

In Equation 6, C is the number of components, $\mathcal{N}(\mathbf{x}; \mathbf{m}_{ij}^c, \Sigma_{ij}^c)$ denotes for 2-D multi-variate normal distribution with mean $\mathbf{m}_{ij}^c$ and covariance Σ_{ij}^c, and w_{ij}^c is the weight of c-th component for pair i, j. The number of components C was obtained using Ref. [25], where BIC scores were used to optimize the number of components. More weights on specific component on space means more chances of appearance within the cluster of the component. For simplicity, the dynamics of the weights over time has been ignored.

We would like to note that the use of Gaussian mixtures rather than a single Gaussian model is justified by the observation that interactions among the same pair might have different modalities (e.g., school, or movies, etc.). In this sense, the model borrows from the mixed membership stochastic block model [1], which assumes that the agents can interact while assuming different roles.

Equations 2-6 complete the definition of our latent point process model. Next we describe our approach for efficient learning and inference with LPPM.

4. **Learning and inference.** As mentioned in the introduction, we are interested in scenario where the actual participants of the events are not observed directly, and need to be inferred, together with the model parameters (i.e., pair-specific parameters of the Hawkes process model and the Gaussian mixture model). For the latter, we employ maximum likelihood (ML) estimation. ML selects the parameters that maximize the likelihood of observations, which consist of the timing and the location of the events, and participant information for some of the events.

Due to the missing labels, some of the interaction matrices z_k are unobserved (or *latent*) for some k. Therefore, there is no closed-form expression for the likelihood of the observed sequence of events. Instead, one has to resort to approximate

techniques for learning and inference, which is described next. Here we use a variational EM approach [4] by positing a simpler distribution $Q(Z)$ over the latent variables with free parameters. The free parameters are selected to minimize the Kullback-Leibler (KL) divergence between the variational and the true posterior distributions. Recall that the KL divergence between two distribution Q and P is defined as

$$D_{KL}(Q||P) = \int_Z Q(Z) \log \frac{Q(Z)}{P(Z,Y)} dZ \tag{7}$$

where Z is the hidden variables, and Y is the observed variables. In our case, Z is the hidden identity of interaction where some of the portion is known, whereas Y describes the location and the time of the incident.

We introduce the following variational multinomial distribution:

$$Q(\mathcal{Z}_n|\Phi) = \prod_k \prod_{i<j} q(z_k^{ij}|\phi_k) \tag{8}$$

where $\mathcal{Z}_k = \{z_l\}_{l=1}^k$ denotes the set of interaction matrices for events up to the k-th event, and $q(\cdot|\phi_k)$ being the multinomial distribution with parameter ϕ_k. The matrix ϕ_k consists of the free variational parameters ϕ_k^{ij} describing the probability that the agents i and j are involved in the k-th event. Note that the present choice of the variational distribution discards correlations between past and future incidents, thus making the calculation tractable.

The variational parameters are determined by maximizing the following lower bound for the log-likelihood [4]:

$$\begin{aligned}
\mathcal{L}_\Phi(Q,\Theta) &= E_Q\Big[\log \prod_k \prod_{i<j}[\lambda_{ij}(t_k)]^{z_k^{ij}} e^{-\Lambda_{ij}^T}\Big] \\
&+ E_Q\Big[\log \prod_k \prod_{i<j}[r_{ij}(\mathbf{x}_k)]^{z_k^{ij}}\Big] \\
&- E_Q\Big[\log \prod_k \prod_{i<j} q(z_k^{ij}|\phi_k)\Big]
\end{aligned} \tag{9}$$

where Φ is the set of variational parameters, and Θ is the set of all the model parameters. The above equation can be rewritten as follows:

$$\begin{aligned}
\mathcal{L}_\Phi(Q,\Theta) &= E_Q\Big[\sum_k \sum_{i<j} z_k^{ij} \log[\lambda_{ij}(t_k)] - \Lambda_{ij}^T\Big] \\
&+ \sum_k \sum_{i<j} \phi_k^{ij} \log[r_{ij}(\mathbf{x}_k)] \\
&- \sum_k \sum_{i<j} \phi_k^{ij} \log \phi_k^{ij}
\end{aligned} \tag{10}$$

where in the last two terms we have explicitly performed the averaging over the multinomial variational distribution defined in Equation 8.

Variational EM algorithm works by iterating between the E–step of calculating the expectation value using the variational distribution, and the M–step of updating the model (hyper)parameters so that the data likelihood is locally maximized. The overall pseudo-algorithm is shown in Algorithm 1. The details of update equations used in both E–step and M–step are provided in the appendix.

Algorithm 1 Variational EM

Size: consider total of n events, $M = \frac{N(N-1)}{2}$ pairs
Input: data $\mathbf{x}_{1:n}$, $t_{1:n}$, z_k of complete events
Start with initial guess for hyper parameters.
Fix all $\phi_k = z_k$ for labeled events.
repeat
    Initialize all components of ϕ_k corresponding to unknown pairs or event k to $\frac{1}{M}$
    repeat
      for $k = 1$ **to** n **do**
        if the pairs of k-th event is unknown **then**
          Update ϕ_k using Eq. 18
        end if
      end for
    until convergence across all time steps
    Update hyper parameters.
until convergence in hyper parameters

---

5. **Experiments with synthetic data.** We first report our experiments with synthetically generated data for six pairs of agents. The sequence of interaction events was generated according to the LPPM process as follows:

1. For each pair, sample the first time of the incident using an exponential distribution with rate parameter μ.
2. For each pair, sample the duration of time until the next incident using Poisson thinning. Since we are dealing with non-homogeneous Poisson process, we use the so called thinning algorithm [27] to sample the next time of the event. By repeating step 2, we obtain the timestamps of incidents for each pair.
3. For every timestamp of a given pair we sample the location of the incident.

To compare the performance of our algorithm with previous approaches, we follow the experimental set-up proposed in [33], where the authors used temporal-only information for reconstructing missing information in synthetically generated data. In addition to ML estimation, Ref. [33] also used an alternative objective function over relaxed continuous variables, and performed constrained optimization of the new objective function using l^2 regularization. Although their method does not assign proper probabilities to the various timelines, it can provide a ranking of most likely participants.

Following Ref. [33], we consider 40 events, and assume that for 10% (4 events) we do not have participant information. Table 1 shows the overall performance of different approaches. To make the comparison meaningful, we omit the spatial information in our model, and focus on the temporal part only. For our algorithm the results are averaged over 1000 runs.

Throughout this paper we measure the accuracy (expressed as a percentage) by counting the number of correctly identified events divided by the total number of hidden events. Table 1 indicates that all four methods perform almost identically. In particular, all four methods have significantly better accuracy than the simple baseline value 1/6, where each pair is selected randomly. Also, we note that while our methods does slightly worse, it is important to remember that the other methods

TABLE 1. Model evaluation for total of n = 40 events between 6 pairs. Only 4 events have unknown participants. The parameters are $\mu = 10^{-2}$days^{-1}, $\omega = 10^{-1}$days^{-1} and $\beta = 0.5$. The accuracy of top three method is from Ref. [33], and Variational EM is our result using LPPM. The results are averaged over 1000 trials.

METHOD	ACCURACY
EXACT ML	47.3 %
MAX l^1	47%
MAX l^2	47.1%
VARIATIONAL EM	46.9%

assume known value of the parameters, whereas LPPM learns the parameters from the data.

In the next set of experiments we examine the relative importance of spatial and temporal parts by comparing three variants of our algorithm that use 1. Temporal only data, 2. Spatial-only data, and 3. Combined spatial and temporal data. For the spatial component of the data, we use six multivariate normal distributions with the center of each on the vertex point of hexagon (for all 6 pairs). Here we use simple Gaussian for each pair for the spatial process. As in Figure 2, we fix the side length of the hexagon to 1, and analyzed how varying the width of the normal distribution affects the overall performance. Specifically, we varied the covariance matrix Σ from 0.25$\mathbf{I}$ to 4$\mathbf{I}$. Again, the results are averaged over 100 runs. The accuracy was computed by averaging the number of correct estimates divided by the number of unknown incidents.

As expected, the relative importance of the spatial information decreases when increasing σ. In the limit when σ is very large, location of an event does not contain any useful information about the participants, so that the accuracy based on spatial information only should converge to the random baseline 1/6. On the other hand, for small values of Σ, the spatial information helps to increase accuracy.

In the last set of experiments with synthetic data, we examine the performance of LPPM by varying the fraction of unknown incident labels. We compare the performance of LPPM to two baseline methods.

- Baseline I (B1): This method uses self-exciting Hawkes process model using *labeled* data only. We perform MLE to estimate the model parameters of Hawkes process by only considering the events that are labeled. In other words, we discard the events which misses the label: the information on the pairs.
- Baseline II (B2): This method uses homogenous Poisson process model with constant intensity using *both* labeled and unlabeled data. This method is similar to our method except that the temporal process is based on Poisson process. This can be treated as a special case of Hawkes process with $\beta = 0$.

We note that both LPPM and the baseline methods use the spatial component, so any differences in their performance should come from the temporal part of the model only.

The results of our comparative studies are shown in Figure 3. It can be seen that LPPM outperforms both baselines by a significant margin, which increases as the data becomes more noisy. Thus, LPPM is a much better choice when the amount
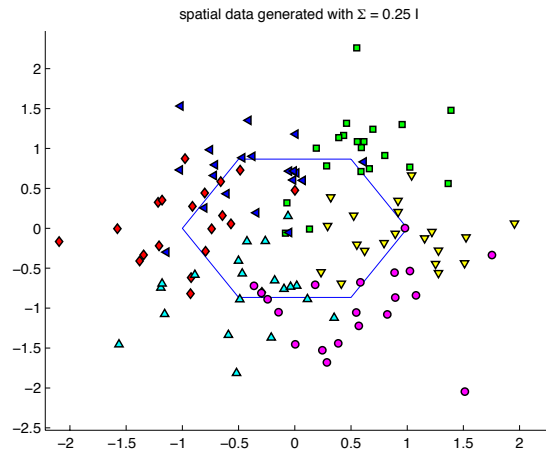
(a)
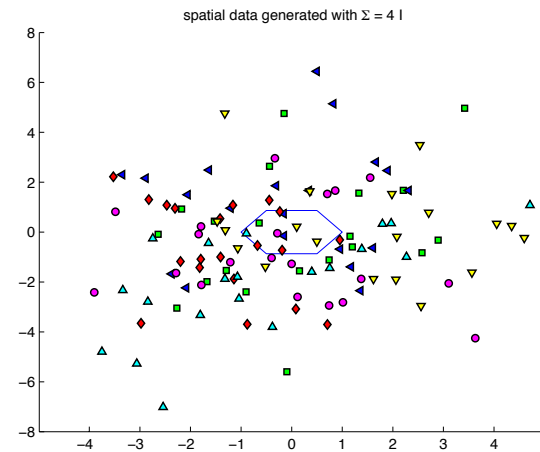
(b)

FIGURE 2. The spatial data generated varying the covariance matrix from 0.25**I** (a) to 4**I** (b). Each color and symbol represents the pairs. 6 centers are on the vertex point of hexagon with side length 1, while the covariance matrix is being controlled

of missing information is significant. The result also reflects that learning model parameters only with the labeled data is not sufficient for inferring missing labels.

6. **Experiments with real–world data.** In this section we report on our experiments using two distinctly different real-world datasets. The first dataset describes gang-rivalry networks in the Hollenbeck police division of Los Angeles [34], and the second dataset is from a popular location-based social networking service *Gowalla* [6]. The rest of the section is organized as follows: We next describe
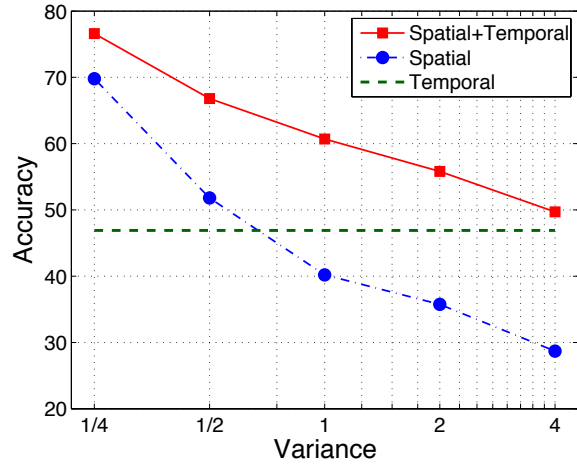
(a)
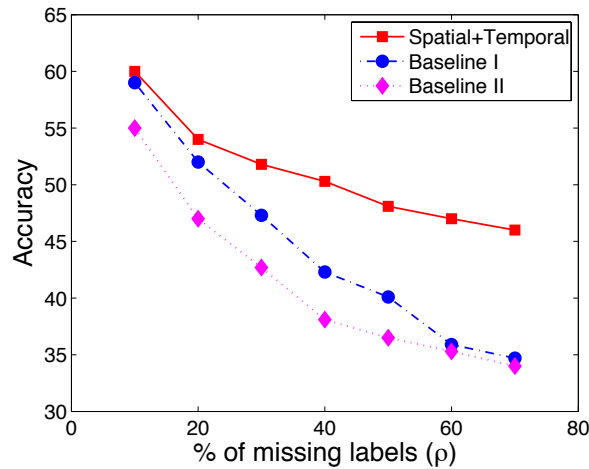
(b)

FIGURE 3. (a) Accuracy of inference using spatial data only, temporal data only, and spatial-temporal data, for different settings of the standard deviation of the spatial Gaussian model. The results are averaged over 100 trials; (b) Average accuracy (over 20 trials) plotted against the percentage of missing labels. Spatial data was generated based on Gaussian with standard deviation 1.

both datasets; Then we conduct experiments on identity-inference problems in Section 6.2. Finally, we evaluate LPPM for event prediction problem in Section 6.3

6.1. **Data description.**
LAPD dataset. Hollenbeck is a 15.2 square mile (39.4 km2) policing division of the Los Angeles Police Department (LAPD), located on the eastern edge of the City of Los Angeles, with approximately 220,000 residents. Overall, 31 active criminal

street gangs were identified in Hollenbeck between 1999-2002 [34]. These gangs formed at least 40 unique rivalries, which are responsible for the vast majority of violent exchanges observed between gangs. Between November 14, 1999 and September 28, 2002 (1049 days), there were 1208 violent crimes attributed to criminal street gangs in the area. Of these, 1132 crimes explicitly identify the gang affiliation of the suspect, victim, or both. The remaining events include crimes such as 'shots fired' which are known to be gang related, but the intended victim and suspect gang is not clear. For each violent crime, the collected information includes the street address where the crime occurred as well as the date and time of the event [34], allowing examination of the spatial-temporal dynamics of gang violence. In Figure 4 we show temporal and spatial distribution of interactions between three most active gangs. For this dataset, we found that each pair is characterized by a simple Gaussian. This could be treated as a special case of GMM with $C = 1$.

Gowalla dataset. Gowalla is a location-based social networking website where users share their locations by checking-in [6]. We used the top 20 nodes who actively check-in to places. The network consists of 196,591 nodes and 950,327 undirected edges. 6,442,890 check-ins of these users were gathered from Feb. 2009 - Oct. 2010. Each check-in not only has its latitude and longitude coordinates but also has a given location ID provided by Gowalla. The location ID is very useful in that it enables us to verify the co-occurrence of a pair at a given location even though the location of latitude and longitude has some error or has a multi-story building at the given coordinates. Gowalla also has a list of friends, where the edge between them is undirected. We looked into every check-in of the friends of 20 nodes and assumed they have interacted each other if the check-in of the two at same location was within 10,000 seconds. The venues of popular places such as airport and stations has been removed to rule out the unexpected coincidence between users. Out of 20 active nodes, we were able to collect 3 groups: one from Stockholm, Tokyo, and San Francisco.

6.2. **Inferring event participants.** As we mentioned earlier, most social network data is noisy and incomplete with missing information about nodes and/or interactions. In this section, we consider a scenario where one has the timing and location of interaction events, but only partial information about event participants. A specific real-world problem described within this scenario is inter-gang violence, where one has a record of reported violent inter-gang events, but where either the perpetrator gang, the victim gang, or both, are unknown. Thus, the problem is to infer the unknown participants based on available information. The naive solution would be to discard the missing data, learn the model parameters based on fully observed events only, and then use the learned model for inferring participants of partially labeled events. However, below we show that the naive approach is sub-optimal. Instead, by taking into account missing data via the expectation-maximization framework, one achieves better accuracy in the participant identification task.

6.2.1. *Experiments with LAPD dataset.* As described above, the LAPD dataset contains the time stamp and the location of incidents between pairs of gangs. Approximately 31% of the records contain information about both participants in the event. Furthermore, 62% of the records contain information about one of the participants, but not the other. Finally, 7% do not have any information about the participants. For better understanding of gang-rivalries, it is important to recover missing information on those 70% of the whole data. Since this research is not the
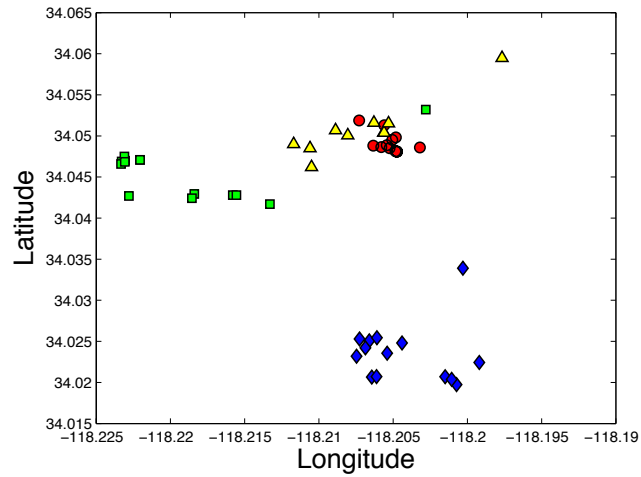
(a)
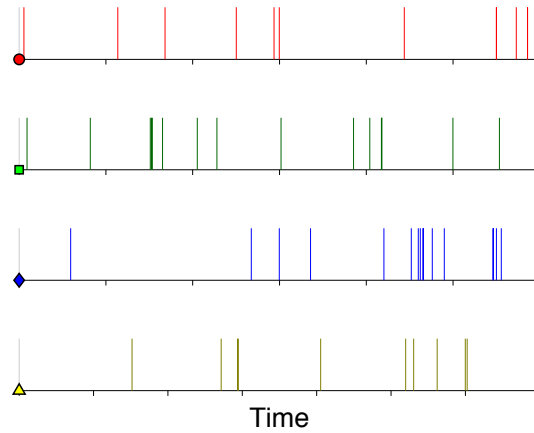
(b)

FIGURE 4. Spatial (a) and temporal (b) description of the events involving four active gang rivalries. Different colors represent different pairs. In (b) each spike represents the time of the event.

studies of the actual rivalries in Hollenbeck but to verify how well our algorithm performs on inference, in the experiments below, we discard the latter portion of the data.This way we could validate our inference and by comparing it with actual given label. In the remaining data, we focused on 31 active gangs which were involved in at least 4 incidents within the time period. Furthermore, out of all possible pairs, we use 40 pairs which had more than one reported incident between each other.

In the first set of experiments, we focused on the portion of the data that contains information about both participants. We randomly select a fraction ρ of the incidents, and then hide the identity of the participants for those incidents. Next,
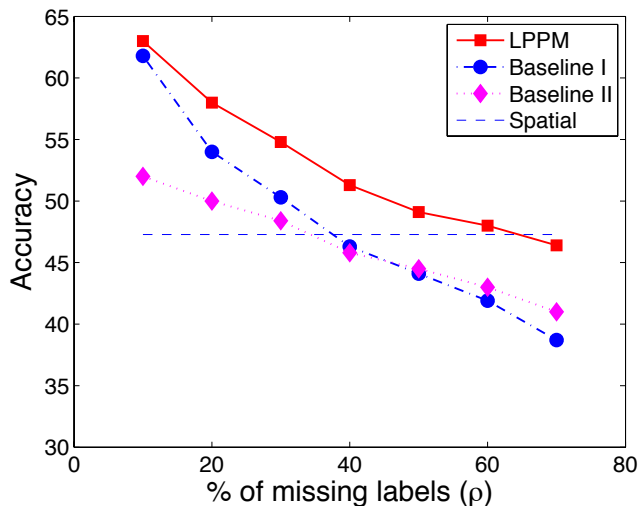
FIGURE 5. Average accuracy for varying fraction of missing labels. Baseline I and Baseline II are defined in Section 5. The horizontal line corresponds to inference using spatial data only.

we use LPPM to see how well it can reconstruct the hidden identities by varying ρ. We compared the results to the same two baseline methods outlined in Section 5. In addition, we add another baseline that uses all existing labels to learn a spatial-only model. The accuracy is defined as the fraction of events for which the algorithm correctly recovers **both** participants. The results were averaged over 20 different runs. The center of clusters were initialized with the mean location of labeled data.

Figure 5 demonstrates our results. One can see that the LPPM does consistently better than B1 and B2. For only 10% of missing label information, the accuracy of LPPM and B1 are fairly close. This is to be expected, since for vanishing ρ those algorithms become identical – they learn the same model using the same data. However, LPPM performs much better than B1 when ρ increases. Another interesting observation is that B2 performs better than B1 when ρ is sufficiently large. This suggests that for large ρ it is better to use a *simpler* (and presumably wrong) model using both missing and labelled data, than learn a more elaborate model using labelled data only.

We also note LPPM does better than the spatial-only baseline even when half of the events are hidden. This is significant since the spatial model uses all the label information that is not available to LPPM. Although the spatial model performs better when ρ increases further, LPPM remains very competitive even when 70% of the events are hidden, which is the same condition (i.e., fraction of unknown) of LAPD gang related crime data.

6.2.2. *Experiments with Gowalla dataset.* Next, we perform experiments on the participant-inference task using the Gowalla data. Note that while the participant information is generally available in this data, it still provides an interesting benchmark for validating LPPM.

Out of 20 most active users in Gowalla network, we focus on three users that have high interaction frequency with their friends. [1] Coincidently, three users were from different city (Tokyo, Stockholm, and San Francisco). We found that some of the check-in locations were repeated by the same pairs. Strictly speaking, this suggests that the spatial component is not a point process. However, this detail has little bearing on our model, as the spatial interactions can still be modeled via the Gaussian mixture model.

Spatial analysis of the dataset reveals that the interaction are multi-modal in the sense that the same pair of users interact at different locations. This is different from the crime dataset, and necessitates using more than one component for the spatial mixture model. In the experiments, we used 4 components of GMM for two of the pairs (Stockholm and San Francisco), and three components for the other pairs (Tokyo).

The results of the experiments are shown in Figure 6. Due to limited space, we present the result of simulation using users in San Francisco. Since the two baseline methods perform similarly, here we show the comparison only with B2, which learns a homogenous Poisson point process model using both labeled and unlabeled data. Again, the results suggest that LPPM is consistently better than the baseline for all of the pairs. The gap between LPPM and the baseline is not significant as before which is mainly due to the active pairs which dominates the interactions. When there are dominant active pairs, Poisson process could distinguish the users by comparing the rate between the pairs. Moreover, there were some active pairs which have checked into the exact same location repeatedly leading to higher accuracy.
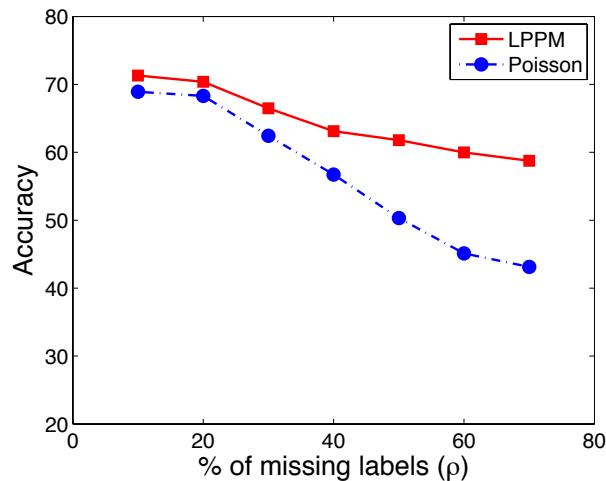

FIGURE 6. Average accuracy of participant-inference task for the user in San Francisco. The fraction of missing labels is varied between 10% and 70%.

6.3. **Event prediction with LPPM.** LPPM can be used not only for inferring missing information but also predicting future events, which can be potentially

---

[1]Recall that for this dataset, an interaction between two users is determined by near-simultaneous check-ins; see the description of the dataset

useful for many applications. For instance, in the context of proactive policing, the predictions can be used to anticipate the participants/timing/location of the next event, and properly assign resources for patrol, etc. Related to friendship network, one can predict the spatial-temporal movement patterns by predicting the hot clusters involving given pairs. This kind of prediction can be also very useful in epidemiology, i.e., by predicting diffusion patterns of an infectious disease.

In this section, we use learned LPPM models for two different prediction tasks:
(1) Predicting the timing of the next interaction event.
(2) Predicting the pair that will have the next interaction.

Let us first discuss the timing prediction problem. Given the history of events up to the k-th event, our goal is to predict the timing of the $(k+1)$-th event. Note that the prediction can be either pair-specific, or across all pairs. Here we select the latter option.

The estimated waiting time until the next incident is given by

$$\int_0^L t\lambda_{\mathcal{S}}(t)\exp(-\int_0^t \lambda_{\mathcal{S}}(\tau)d\tau)dt \tag{11}$$

where L is fairly a large number, and $\lambda_{\mathcal{S}}(t) = \sum_{(ij)}\lambda_{ij}(t)$ is the sum of the conditional intensity function across all the pairs. Below we compare the prediction performance of LPPM with the B2 defined in Section 5, which employs homogenous Poisson processes. According to this baseline, the expected waiting time to the next event is simply $\frac{1}{\sum_{(ij)}\lambda_{ij}^*}$ ($\lambda_{ij}(t) \equiv \lambda_{ij}^*$), where λ_{ij}^* is the time-independent intensity for the pair (i,j).

The prediction accuracy is measured using the *mean absolute percentage error* (MAPE) score, which measure the relative error of the predicted waiting time: MAPE $= |\frac{A_n-F_n}{A_n}|$, where A_n is actual waiting time until the next incident, and F_n is our predicted value. Note that more accurate prediction corresponds to lower MAPE score, MAPE $= 0$ for perfect prediction.

We measure the MAPE score for LPPM prediction on the LAPD and Gowalla datasets. For the former, we use LPPM to predict the timing of the last 50 incidents among top 40 pairs. As for the latter dataset, we focus on only one of the users (in Tokyo), and use the last 10 events (out of 40 total) for prediction. For both datasets, LPPM provides significantly more accurate prediction than the baseline for most of the incidents. LAPD dataset had **2.7502** for LPPM compared to **11.0434** for B2; Gowalla dataset had **1.2236** for LPPM compared to **5.9350** for B2. A possible explanation of the poor performance of the Poisson model is that it fails to accurately predict the timing of highly correlated events that are clustered in time, whereas LPPM is able to capture such correlations. When the next event is highly influenced by the previous event, Poisson model is limited in that it considers the triggered event as a random event.

For the prediction task (2), we used LPPM to find the conditional intensity of interactions between different pairs based on all the events up to event k that happens at time t_k. We then predict that the pair with the highest conditional intensity to have an interaction event at a time $t > t_k$, assuming that no other interaction has taken place in time interval $[t_k, t]$. Note that the homogeneous Poisson process model (Baseline II) simply selects the pair that has been the most active in the past. For this particular task, we also use another prediction method (Baseline III) which predicts that the pair that had the last event will also participate in the follow-up

event. In addition to the top pair, we also predict the second and third best predictions. We performed experiments with the crime dataset, for which 14 incidents out of 100 were predicted correctly by LPPM. Baseline II correctly predicted only 8 incidents, whereas Baseline III did considerably better with 13 correct predictions. Furthermore, LPPM outperforms both methods in predicting top 2 and top 3 users, as shown in Table 2.

TABLE 2. Prediction accuracy of top-K choices for K=1,2,3.

METHOD	BASELINE II	BASLINE III	LPPM
TOP 1	8%	13%	14%
TOP 2	16%	20%	26%
TOP 3	23%	22%	37%

7. **Conclusion.** We suggested a latent point process model to describe spatial-temporal interaction networks. In contrast to existing continuous time models of temporal networks, here we assume that interactions along the network links are only partially observable. We describe an efficient variational EM approach for learning and inference with such models, and demonstrated a good performance in our experiments with both synthetic and real-world data.

We note that while our work was motivated by modeling spatial-temporal interaction networks, the latent point process suggested here is much more general and can be used for modeling scenarios where one deals with latent mixture of arbitrary point processes. For instance, LPPM can be generalized to describe geographically distributed sequence of arbitrary events of multiple pairs even with the events which misses the pair information.

There are several ways to generalize the model further. For instance we have assumed a homogenous background rate, whereas in certain scenarios one might need to introduce cyclic activity patterns. Furthermore, the assumption that the process intensity is factorized into temporal and spatial components might not work well for certain types of processes, where the location component might depend on the event time.

Appendix A. **Variational E-step.** In the variational E-step, we maximize $\mathcal{L}_\Phi$ over the variational parameters. Note that the variational parameters shoud satisfy the normalization constraint $\sum_{i<j} \phi_p^{ij} = 1$. By introdcuing Lagrange multipliers γ_p to enforce those constraints, and taking the derivative of Equation 10 with respect to the variational parameters yields

$$
\begin{aligned}
0 \;=\; & \frac{\partial}{\partial \phi_p^{ij}} E_Q \Big[\sum_k z_k^{ij} \log[\lambda_{ij}(t_k)] - \Lambda_{ij}^T \Big] \\
+ \; & \log[r_{ij}(\mathbf{x}_p)] \\
- \; & \log \phi_k^{ij} - 1 + \gamma_p
\end{aligned}
\tag{12}
$$

Solving the constrained optimization problem with Lagrange multipliers, we have the update equation for variational parameter ϕ_p^{ij} as below:

$$\phi_p^{ij} = \frac{1}{C_p} \exp\left\{ \frac{\partial}{\partial \phi_p^{ij}} E_Q \left[\sum_k z_k^{ij} \log \lambda_{ij}(t_k) - \Lambda_{ij}^T \right] \right\} [r_{ij}(\mathbf{x}_p)] \tag{13}$$

where the Lagrange multiplier has been absorbed in the normalization constant C_p.

For the evaluating the derivative of the expectation of $\log \lambda_{ij}(t_k)$ with respect to ϕ_p^{ij} in the above equation, we separate into two cases when $k > p$ and $k = p$. Before expressing the derivatives for two cases, we introduce a new function for a simpler expression.

$$\mathcal{M}_{ij}(\mathcal{Z}_k) = \prod_{l=1}^{k} \phi_l^{ij}{}^{z_l^{ij}} \left(1 - \phi_l^{ij}\right)^{\left(1 - z_l^{ij}\right)} \tag{14}$$

which is a joint probability of given scenario from the beginning up to event k.

First, for the case when $k = p$, we have

$$\frac{\partial}{\partial \phi_p^{ij}} E_Q[\log \lambda_{ij}(t_p)] = \sum_{\mathcal{Z}_{p-1}} \mathcal{M}_{ij}(\mathcal{Z}_{p-1}) \log\left[\mu_{ij} + \sum_{l=1}^{p-1} z_l^{ij} g_{ij}(t_p - t_l) \right] \tag{15}$$

In the right hand side of Equation 15, the sum is over all the possible configurations of the latent variables up to the event $p - 1$, $Z_{k=1}^{p-1}$. Similarly, we can derive the derivative with respect to ϕ_p^{ij} for the terms with $k > p$. For steps when k is greater than p,

$$\frac{\partial}{\partial \phi_p^{ij}} E_Q[\log \lambda_{ij}(t_k)] = \sum_{\tilde{\mathcal{Z}}_{k-1}^p} \phi_k^{ij} \tilde{\mathcal{M}}_{ij}^p(\tilde{\mathcal{Z}}_{k-1}^p) \tag{16}$$

$$\times \log\left[\frac{\mu_{ij} + \sum_{l=1, l \neq p}^{k-1} z_l^{ij} g_{ij}(t_k - t_l) + g_{ij}(t_k - t_p)}{\mu_{ij} + \sum_{\substack{l=1 \\ l \neq p}}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)} \right]$$

where we have defined $\tilde{\mathcal{Z}}_k^p$ as $\mathcal{Z}_k$ excluding z_p with $\tilde{\mathcal{M}}_{ij}^p(\cdot)$ following the same logic. The numerator term in the logarithm above comes from when pair i and j trigger the k-th event on the p-th event, while the denominator term comes from when they did not.

Finally for the derivative of expectation of Λ_{ij}^T in Equation 13, we use

$$\frac{\partial}{\partial \phi_p^{ij}} E_Q[-\Lambda_{ij}^T] = -\beta_{ij}\{1 - \exp(\omega_{ij}(T - t_p))\} \tag{17}$$

By combining Equation 13 – 17, we obtain an iterative scheme for finding the variational parameters of the form

$$\phi_p^{ij} = f(\{\phi_p^{ij}\}_{k=1; k \neq p}^{n}; \Theta) \tag{18}$$

The above iterations are used until the convergence of all the variational parameters.

Appendix B. **Variational M-step.** The M-step in the EM algorithm computes the parameters by maximizing the expected log-likelihood found in the E-step. The model parameters consists of spatial parameters and temporal parameters. We first look into the update equations of spatial parameters. For some cases, when the spatial pattern is distinct over pairs, we use single Gaussian for each pair, and

the update equations are as below (i.e., the mean and the variance of Gaussian distribution):

$$\mathbf{m}_{ij} \quad \leftarrow \quad \frac{\sum_k \phi_k^{ij} \mathbf{x}_k}{\sum_k \phi_k^{ij}} \tag{19}$$

$$\sigma_{ij,lat}^2 \quad \leftarrow \quad \frac{\sum_k \phi_k^{ij} (\mathbf{x}_{k,lat} - \mathbf{m}_{ij,lat})^2}{\sum_k \phi_k^{ij}} \tag{20}$$

$$\sigma_{ij,long}^2 \quad \leftarrow \quad \frac{\sum_k \phi_k^{ij} (\mathbf{x}_{k,long} - \mathbf{m}_{ij,long})^2}{\sum_k \phi_k^{ij}} \tag{21}$$

When using a Gaussian mixture model, the weight vector of the mixture model for each pair is updated respectively.

$$w_{ij}^c \leftarrow \frac{\sum_k \phi_k^{ij} \frac{\mathcal{N}(\mathbf{x}_k | \mathbf{m}_{ij}^c, \Sigma_{ij}^c)}{\sum_{p=1}^{C} \mathcal{N}(\mathbf{x}_k | \mathbf{m}_{ij}^p, \Sigma_{ij}^p)}}{\sum_k \phi_k^{ij}} \tag{22}$$

The re-estimation of the temporal parameters are more involved. For instance, to estimate μ_{ij}, we nullify the derivative of the likelihood with respect to μ_{ij}, $\frac{\partial \mathcal{L}_{\Phi}}{\partial \mu_{ij}} = 0$, which yields

$$\mu_{ij} \leftarrow \frac{\sum_k \sum_{\mathcal{Z}_{k-1}} \phi_k^{ij} \frac{\mu_{ij} \mathcal{M}_{ij}(\mathcal{Z}_{k-1})}{\mu_{ij} + \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}}{T} \tag{23}$$

Similarly, for re–estimation of β_{ij}, we present the derivative as below:

$$\beta_{ij} \leftarrow \frac{\sum_k \sum_{\mathcal{Z}_{k-1}} \phi_k^{ij} \frac{\mathcal{M}_{ij}(\mathcal{Z}_{k-1}) \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}{\mu_{ij} + \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}}{\sum_k \phi_k^{ij} \int_0^{T-t_k} \omega_{ij} e^{-\omega_{ij} \tau} d\tau} \tag{24}$$

Finally, for ω_{ij}, we obtain

$$\tag{25}$$

$$\sum_k \phi_k^{ij} \Bigg[\sum_{\mathcal{Z}_{k-1}} \Big[\frac{\big(\sum_{l=1}^{k-1} z_l^{ij} (1 - \omega_{ij}(t_k - t_l)) g_{ij}(t_k - t_l) \big)}{\mu_{ij} + \sum_{l=1}^{k-1} z_k^{ij} g_{ij}(t_k - t_l)}$$

$$\times \mathcal{M}_{ij}(\mathcal{Z}_{k-1}) \Big] - \beta_{ij}(t_k - T) e^{-\omega_{ij}(T-t_k)} \Bigg] = 0$$

where

$$g_{ij}(t - t_p) = \beta_{ij} \omega_{ij} \exp\{-\omega_{ij}(t - t_p)\} \tag{26}$$

Unfortunately, the resulting equations do not allow closed form solutions, so they have to be solved using numerical methods, such as the Newton's method employed here. We can also have closed form of update equation of ω_{ij} by approximating the second term to zero in Equation 25. When $\omega_{ij} T$ is fairly large compared to $\omega_{ij} t_k$, we can ignore the second term, and have closed form as below:

$$\omega_{ij} \leftarrow \frac{\sum_k \sum_{\mathcal{Z}_{k-1}} \phi_k^{ij} \frac{\mathcal{M}_{ij}(\mathcal{Z}_{k-1}) \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}{\mu_{ij} + \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}}{\sum_k \sum_{\mathcal{Z}_{k-1}} \phi_k^{ij} \frac{\mathcal{M}_{ij}(\mathcal{Z}_{k-1}) \sum_{l=1}^{k-1} z_l^{ij} (t_k - t_l) g_{ij}(t_k - t_l)}{\mu_{ij} + \sum_{l=1}^{k-1} z_l^{ij} g_{ij}(t_k - t_l)}} \tag{27}$$

The following remark is due: the update equations for both the variational parameters and the model parameters involve summation over the all possible configurations of the latent variables. This sum might become prohibitively extensive for long history windows. However, due to the exponential decay of the self-excitation term, events too far in the past have negligible impact on future events. This observation justifies limiting the summation to a window, i.e., $\lambda_{ij}(t_k|\mathcal{H}_{t_k}) \approx \lambda_{ij}(t_k|\{h_l\}_{l=k-d}^k)$, which discards events that are far in the past. In the results, we use this truncation to speed up the inference process.

REFERENCES

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg and E. P. Xing, Mixed membership stochastic blockmodels, *Journal of Machine Learning Research,* **9** (2008), 1981–2014.

[2] S. Azizpour, K. Giesecke, S. F. Discussions, X. Ding, B. Kim and S. Mudchanatongsuk, *Self-exciting corporate defaults: Contagion vs. frailty*, 2008.

[3] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics, *Nature,* **435** (2005), 207–211.

[4] M. Beal and Z. Ghahramani, The variational bayesian em algorithm for incomplete data: With application to scoring graphical model structures, *Bayesian Statistics,* **7** (2003), 453–464.

[5] P. Bremaud, *Point Processes and Queues : Martingale Dynamics*, Springer series in statistics, Springer-Verlag, New York, 1981.

[6] E. Cho, S. A. Myers and J. Leskovec, Friendship and mobility: User movement in location-based social networks, in *Proc. of the KDD'11,* (2011), 1082–1090.

[7] N. Cressie and C. K. Wikle, *Statistics for Spatio-Temporal Data*, (Wiley Series in Probability and Statistics) Wiley, 2011.

[8] N. Du, L. Song, A. Smola and M. Yuan, *Learning Networks of Heterogeneous Influence*, In Advances Neural Information Processing Systems, 2012.

[9] E. Errais, K. Giesecke and L. Goldberg, Affine point processes and portfolio credit risk, *SIAM Journal on Financial Mathematics,* **1** (2010), 642–665.

[10] R. Eyal, S. Kraus and A. Rosenfeld, *Identifying Missing Node Information in Social Networks*, in AAAI'11, 2011.

[11] Y. Fan and C. R. Shelton, Learning continuous-time social network dynamics, in *Proc. of the 25th Conference on Uncertainty in Artificial Intelligence,* (2009), 161–168.

[12] M. Gomez-Rodriguez, J. Leskovec and A. Krause, Inferring Networks of Diffusion and Influence, in *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (ACM KDD),* **16** (2010), 1019–1028.

[13] M. Gomez-Rodriguez, J. Leskovec and B. Schölkopf, Structure and Dynamics of Information Pathways in Online Media, in *WSDM,* **6** (2013), 23–32.

[14] M. Gomez-Rodriguez, J. Leskovec and B. Schölkopf, *Modeling Information Propagation with Survival Theory*, in ICML, 2013.

[15] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, *PNAS,* **106** (2009), 22073–22078.

[16] A. Gunawardana, C. Meek and P. Xu, A model for temporal dependencies in event streams, in *Advances in Neural Information Processing Systems,* **24** (2011), 1962–1970.

[17] S. Hanneke, W. Fu and E. Xing, Discrete temporal models of social networks, *Electronic Journal of Statistics,* **4** (2010), 585–605.

[18] R. Hegemann, E. Lewis and A. Bertozzi, An Estimate & Score Algorithm for simultaneous parameter estimation and reconstruction of incomplete data on social network, *Security Informatics,* **2** (2013).

[19] M. Kim and J. Leskovec, The network completion problem: Inferring missing nodes and edges in networks, in *SDM, SIAM / Omnipress,* (2011), 47–58.

[20] G. Kossinets, Effects of missing data in social networks, *Social Networks,* **28** (2006), 247–268.

[21] D. Liben-Nowell and J. Kleinberg, The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology,* **58** (2007), 1019–1031.

[22] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg and G. E. Tita, Self-exciting point process modeling of crime, *Journal of the American Statistical Association,* **106** (2011), 100–108.

[23] P. Netrapalli and S. Sanghavi, Learning the graph of epidemic cascades, In *Procs. of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems,* (2012), 211–222.

[24] Y. Ogata, Space-time point process models for earthquake occurrences, *Ann. Inst. Statist. Math.,* **50** (1988), 379–402.

[25] D. Pelleg and A. W. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *Proceedings of the Seventeenth International Conference on Machine Learning,* ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, (2000), 727–734.

[26] P. O. Perry and P. J. Wolfe, Point process modeling for directed interaction networks, *Journal of the Royal Statistical Society, Series B,* **75** (2013), 821–849.

[27] P.Lewis and G.Shedler, Simulation of nonhomogenous Poisson processes by thinning, *Naval Research Logistics Quarterly,* **26** (1979), 403–413.

[28] M. D. Porter and G. White, Self-exciting hurdle models for terrorist activity, *The Annals of Applied Statistics,* **6** (2012), 106–124.

[29] S. Rajarm, T. Graepel and R. Herbrich, *Poisson-networks: A Model for Structured Point Processes*, in Proceedings of the International Workshop on Artificial Intelligence and Statistics, 2005.

[30] A. Simma and M. I. Jordan, *Modeling Events with Cascades of Poisson Processes*, in Proceedings of the Twenty sixth International Conference on Uncertainty in Artificial Intelligence, 2010.

[31] C. Steglich, T. A. B. Snijders and M. Pearson, *Dynamic Networks and Behavior: Separating Selection from Influence*, Sociological Methodology, 2010.

[32] J. Stehlé, A. Barrat and G. Bianconi, Dynamical and bursty interactions in social networks, *Phys. Rev. E,* **81** (2010), 035101.

[33] A. Stomakhin, M. B. Short and A. L. Bertozzi, Reconstruction of missing data in social networks based on temporal patterns of interactions, *Inverse Problems,* **27** (2011), 115013.

[34] G. Tita, J. K. Riley, G. Ridgeway, C. Grammich, A. F. Abrahamse and P. Greenwood, *Reducing Gun Violence: Results from an Intervention in East Los Angeles*, RAND Press, 2003.

[35] G. Ver Steeg and A. Galstyan, *Information Transfer in Social Media*, in Proc. of WWW, 2012.

[36] G. Ver Steeg and A. Galstyan, *Information-Theoretic Measures of Influence Based on Content Dynamics*, in Proc. of WSDM, 2013.

[37] D. Vu, A. U. Asuncion, D. Hunter and P. Smyth, Continuous-time regression models for longitudinal networks, in *NIPS,* (2011), 2492–2500.

[38] L. Wang, S. Ermon and J. Hopcroft, Feature-enhanced probabilistic models for diffusion network inference, *Lecture Notes in Computer Science*, **7524** (2012), 499–514.

E-mail address: yoonsik@isi.edu

E-mail address: galstyan@isi.edu

E-mail address: branting@ucla.edu

E-mail address: gtita@uci.edu