Learning Parametric Models for Social Infectivity in Multi-Dimensional Hawkes Processes

Liangda Li[1] and **Hongyuan Zha**[2,1]
[1]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[2]Software Engineering Institute, East China Normal University, Shanghai, China

Abstract

Efficient and effective learning of social infectivity presents a critical challenge in modeling diffusion phenomena in social networks and other applications. Existing methods require substantial amount of event cascades to guarantee the learning accuracy and they only consider time-invariant infectivity. Our paper overcomes those two drawbacks by constructing a more compact model and parameterizing the infectivity using time-varying features, thus dramatically reduces the data requirement, and enables the learning of time-varying infectivity which also takes into account the underlying network topology. We replace the pairwise infectivity in the multidimensional Hawkes processes with linear combinations of those time-varying features, and optimize the associated coefficients with lasso-type of regularization. To efficiently solve the resulting optimization problem, we employ the technique of alternating direction method of multipliers which allows independent updating of the individual coefficients by optimizing a surrogate function upper-bounding the original objective function. On both synthetic and real world data, the proposed method performs better than alternatives in terms of both recovering the hidden diffusion network and predicting the occurrence time of social events.

Introduction

How social influence affects people's behaviors in social networks? How to efficiently model time-varying influence in social networks without prior knowledge about the networks' topologies? Such challenging problems attract increasingly interest in social network analysis. Social influences play a major role in determining the path and speed that memes, such as ideas, information, behaviors, or diseases, spread in social networks. For instance, rumors diffuse to public via friendship or kinship, contagious viruses spread among people who interact frequently. For each meme, an event cascade is formed by the sequence of events recording the timestamps that the meme comes to an individual, from which people expect to learn social influence between individuals.

Recent works (Yang and Zha 2013; Zhou, Zha, and Song 2013a; 2013b) frequently employed, one powerful statistical tool, the multi-dimensional Hawkes process (Hawkes 1971), to model timestamped and recurrent events in social networks to learn the degree of pairwise influence between individuals, which our paper calls *infectivity*, by taking each individual as one dimension. Hawkes process is well known for its self-exciting property, a common social phenomenon that the occurrence of one event increases the probability of future events, which is not discussed in other social network analysis models such as Exponential Random Graph Model (ERCM) (Robins et al. 2007) and Cox model (Perry and Wolfe 2013). Hawkes process has been widely used in applications, such as earthquake prediction(Ogata 1988; Zhuang, Ogata, and Jones 2002), market modeling (Errais, Giesecke, and Goldberg 2010; Ait-Sahalia, Cacho-Diaz, and Laeven 2010), crime modeling (Stomakhin, Short, and Bertozzi 2011), and conflict analysis(Z.-Mangion et al. 2012; Li and Zha 2013).

Formally, the multi-dimensional Hawkes process on an event cascade $\{t_l\}$ is defined to be a M-dimensional point process with the intensity of the m-th dimension given by:

$$\lambda_m(t) = \mu_m + \sum_{t_l<t} \alpha_{m_l,m}\kappa(t-t_l)$$

Here μ_m denotes the basic intensity of the m-th dimension, $\kappa(t-t_l)$ is a time-decaying kernel, while $\alpha_{m,m'}$ denotes the infectivity from events in the m-th dimension to events in the m'-th dimension. We call $\mathbf{A} = (\alpha_{m,m'})$ the *infectivity matrix*. The Hawkes parameters need to learn include $O(M)$ μ's and $O(M^2)$ α's.

Unfortunately, although having achieved remarkable performances, existing works suffer from the following drawbacks in learning α:

Problem Complexity. Learning one separate α for each pair of dimensions is daunting. On one hand, learning $O(M^2)$ α's can be both time-consuming and unnecessary under certain scenarios. Modern social networks, such as Fackbook, Twitter, and Youtube, are always participated by numerous individuals, while infectivity only exists in limited individual-pairs. On the other hand, the chances are very high that there are no sufficient historical events for modeling the infectivity within certain individual-pairs. As shown in Figure 1, a multi-dimensional Hawkes model already

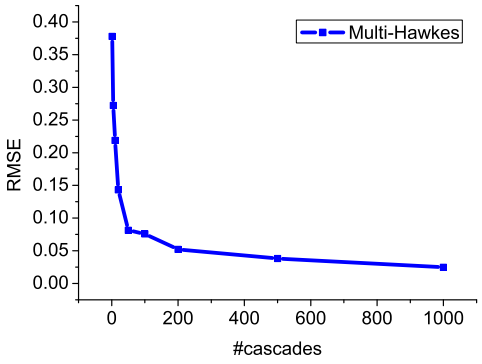


Figure 1: How the increase of cascades affects the performance of a normal multi-dimensional Hawkes model in recovering a 10×10 synthetic network. The average length of event cascades is about 30. The Y-axis uses RMSE to measure the difference between estimated infectivity matrix $\mathbf{A}$ and the ground-truth one.

needs $O(10^3)$ cascades to accurately recover a 10×10 synthetic diffusion network. Thus to recover large-scale real world networks, it will demand far more cascades, which are usually not available. Furthermore, real world networks do not always clarify the cascade membership of new coming events. Thus, existing approaches (Rodriguez, Balduzzi, and Scholkopf 2011; Zhou, Zha, and Song 2013a) generally require the successive event history to be segmented into a number of independent cascades in advance. Such segmentation depends heavily on human annotations which demand massive workloads, or external algorithms which are unable to guarantee the correctness.

Dependency in Infectivity Matrix. Existing works (Stomakhin, Short, and Bertozzi 2011; Ait-Sahalia, Cacho-Diaz, and Laeven 2010) usually ignore the dependency among α's, while under many circumstances α's are closely related. For instance, the friendship between Alice and Bob and that between Alice and Clark probably imply a friendship between Bob and Clark. Recent works (Zhou, Zha, and Song 2013a) attempted to capture such dependency among α's by imposing priors on the network topology, such as sparsity and low-rank structure. However, a priori assumptions on the network topology limit the adaptive social networks of those approaches. The structures of different social networks vary a lot, and even contradict with each other. For instance, the sparsity assumption works in Facebook, where users only influence a small number of acquaintances, however, fails in regional conflicts, where military organizations usually form two alliances and each pair of rivals fight frequently.

Time-varying Infectivity. The infectivity α between each pair of individuals is usually time-variable. A satisfactory purchasing recommendation from Alice to Bob consequently raises Bob's trust on Alice. In a city's gang network, last year's rivals may fight side by side currently due to the variation of conflicts of interest from time to time. Potential solutions for learning time-varying infectivity, such as learning separate α's for each time interval or modeling α with time-dependent functions, greatly increase problem complexity.

In this paper, to address above drawbacks simultaneously, we build a compact model to parameterize the infectivity between individuals. The basic idea is to design a set of K time-varying features, and substitute each α with a linear combination of those features with coefficients to learn. In this way, we 1) only need to estimate K coefficients, which are controlled by the number of features we use, instead of the square of the number of individuals in the given social network. Moreover, the estimation of each coefficient fully utilizes all historical events, thus no longer demands multiple cascades and the a priori cascade assignment of new upcoming events; 2) are free to design features capturing the dependency among infectivies within each individual-pair, based on the pairwise direct or indirect interactions. Compared with methods that impose regularization on $\mathbf{A}$, our idea not only prevents problem complexity from increasing by calculating features ahead, but also avoids making subjective assumptions on social network topology. Our features actually incorporate various kinds of such assumptions, in complementary or in contradictory, and the coefficient estimation process validates assumptions in consistent with the specific social network we observe. For instance, in a sparse network, features recording direct interactions between individuals are more likely to weight higher than features reflecting indirect interactions, since the former ones are more rare than the latter ones; 3) our designed time-varying features are capable of describing the change of infectivity wrt. time. By calculating the values of these features for each individual-pair at each event-timestamp prior to model learning, we avoid increasing problem complexity.

We introduce a set of time-varying features that imply the instant self-properties of each individual, or the instant relationship between each pair of individuals. Replacing α's with linear combinations of time-varying features, we raise the problem of optimizing the corresponding coefficients with lasso regularization on them, and solve the problem efficiently by developing an algorithm that combines the idea of alternating direction method of multipliers (ADMM) (Boyd 2010) and Majorize-Minimization (MM) (Hunter and Lange 2004). Experiments on both synthetic and real world data demonstrate that the proposed method more accurately recovers the hidden network and predicts the occurrence time of events than alternatives.

Problem Formulation

A multi-dimensional Hawkes process estimates basic intensity μ and infectivity α by maximizing the likelihood on each observed event cascade $\{t_n, m_n\}_{n=1}^N$ as:

$$\mathcal{L} = \sum_{n=1}^{N} \log \lambda_{m_n}(t_n) - \sum_{m=1}^{M} \int_0^T \lambda_m(s) ds$$

where M is the number of dimensions, t_n is the timestamp of the n-th event in the cascade, and m_n indicates the dimension/individual where the n-th event occurs.

In real world social networks, M can be very large, dependency exists among α's, and α may varies with respect to time. Thus learning one separate α for each pair of dimensions (m, m') can be both inefficient and ineffective. To

address those issues, instead, we decompose each α into a linear combination of K time-varying features as:

$$\alpha_{m,m'} = \beta^T \mathbf{x}_{m,m'}(t), \tag{1}$$

where β is the vector of coefficients that we are to learn instead of α. $\mathbf{x}_{m,m'}(t)$ is a time-varying dyad-dependent vector of length K, which is supposed to reflect some kind of relationship between dimension m and m'.

Plugging Eqn (1) into the intensity function of multi-dimensional Hawkes processes, we can write the log-likelihood of model parameters μ, β as:

$$\mathcal{L}(\mu,\beta) = \sum_{n=1}^{N} \log(\mu_{m_n} + \beta \sum_{l=1}^{n-1} \kappa(t_n - t_l)\mathbf{x}_{m_l,m_n}(t_n)) - T\sum_{m=1}^{M}\mu_m$$
$$-\beta^T \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{n-1} \mathbf{x}_{m_l,m_n}(t_n)(K(t_n - t_l) - K(t_{n-1} - t_l))$$

where $K(t) = \int_0^t \kappa(s)ds$.

To select effective features and avoid overfitting, we enforce the sparsity of coefficients β by imposing lasso type of regularization as $\|\beta\|_1$. Under this lasso regularization, β_k will be non-zero only when its corresponding feature is highly correlated with the infectivity between two dimensions; otherwise, β_k will be enforced to be zero. In summary, we are to optimize model parameters μ, β as:

$$\min_{\mu\geq 0,\beta\geq 0} -\mathcal{L}(\mu,\beta) + \lambda\|\beta\|_1 \tag{2}$$

where λ is the regularization parameter that trades off the sparsity of the coefficients and the data likelihood.

Optimization

Optimizing β against $\mathcal{L}$ is relatively difficult, since the non-smooth regularizer on β makes the objective function non-differentiable. To optimize such an objective, we employ alternating direction method of multipliers (ADMM) (Boyd 2010) to reduce this ℓ_1 regularized loss minimization problem to a sequence of ℓ_2 regularized loss minimization problems, which are much easier to solve. ADMM is known as a special case of the more general Douglas-Rachford splitting method, which has good convergence properties under some fairly mild conditions (Eckstein and Bertsekas 1992).

Derivation of ADMM

In ADMM, the optimization problem in Eqn (2) can be re-written to the following equivalent form by introducing an auxiliary variable $\mathbf{z}$:

$$\min_{\mu\geq 0,\beta\geq 0,\mathbf{z}} -\mathcal{L}(\mu,\beta) + \lambda\|\mathbf{z}\|_1,$$
$$\text{subject to} \quad \beta = \mathbf{z}.$$

The corresponding augmented Lagrangian of the problem is:

$$\mathcal{L}_\rho = -\mathcal{L}(\mu,\beta) + \lambda\|\mathbf{z}\|_1 + \rho\mathbf{u}(\beta - \mathbf{z}) + \frac{\rho}{2}\|\beta - \mathbf{z}\|_2^2,$$

where $\mathbf{u}$ is the scaled dual variables corresponding to the constraint $\beta = \mathbf{z}$, and ρ is the penalty parameter, which is usually used as the step size in updating the dual variable.

Then we solve the above augmented Lagrangian using the ADMM algorithm consisting of the following iterative steps:

$$\mu^{i+1}, \beta^{i+1} = \text{argmin}_{\mu\geq 0,\beta\geq 0} -\mathcal{L}_\rho(\mu,\beta,\mathbf{z}^i,\mathbf{u}^i),$$
$$\mathbf{z}^{i+1} = S_{\lambda/\rho}(\beta^{i+1} + \mathbf{u}^i),$$
$$\mathbf{u}^{i+1} = \mathbf{u}^i + \beta^{i+1} - \mathbf{z}^{i+1}.$$

where S_κ is the soft thresholding operator (Donoho and Johnstone 1995). We will derive the algorithm for optimizing μ and β in the following, which is a proximal operator evaluation.

Estimation of μ and β

In order to update each μ and β independently, we choose to optimize a surrogate function which breaks down the log-sum of $\log\lambda_{m_n}(t_n)$ based on Jensen's inequality, and upper-bounds of $-\mathcal{L}_\rho(\mu,\beta,\mathbf{z}^i,\mathbf{u}^i)$. By optimizing this surrogate function in the Majorize-Minimization (MM) algorithm (Hunter and Lange 2004), we can reach the gloabl optimum of $-\mathcal{L}_\rho$. We define the surrogate function as:

$$g(\mu,\beta|\mu^{(j)},\beta^{(j)}) = \rho\mathbf{u}^i(\beta - \mathbf{z}^i) + \frac{\rho}{2}\|\beta - \mathbf{z}^i\|_2^2 - \sum_{n=1}^{N}\eta_{n0}\log\left(\frac{\mu_{m_n}}{\eta_{n0}}\right)$$
$$+\sum_{n=1}^{N}\sum_{k=1}^{K}\eta_{nk}\log\left(\frac{\beta_k\sum_{l=1}^{n-1}x_{m_l,m,k}(t_n)\kappa(t_n - t_l)}{\eta_{nk}}\right)$$
$$-\beta^T\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{n-1}\mathbf{x}_{m_l,m_n}(t_n)(K(t_n - t_l) - K(t_{n-1} - t_l)),$$

where $\{\eta\}$ is a set of branching variables formulated by:

$$\eta_{n0} = \frac{\mu_{m_n}^{(j)}}{\mu_{m_n}^{(j)} + \sum_{k=1}^{K}\beta_k^{(j)}\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)\kappa(t_n - t_l)},$$
$$\eta_{nk} = \frac{\beta_k^{(j)}\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)\kappa(t_n - t_l)}{\mu_{m_n}^{(j)} + \sum_{k=1}^{K}\beta_k^{(j)}\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)\kappa(t_n - t_l)}.$$

Notice here we can interpret η_{n0} as the infectivity of all historical events on the n-th event with regard to the k-th feature, while η_{n0} is the probability that the n-th event is sampled from the base intensity.

As proved in (Zhou, Zha, and Song 2013a), optimizing the surrogate function g ensures that $\mathcal{L}_\rho$ decreases monotonically, thus guarantees that $\mathcal{L}_\rho$ will converge to a global optimum. Then by optimizing g, we are able to update μ and β independently with closed-form solutions, and automatically take care of the non-negativity constraints as follows:

$$\mu_m = \frac{1}{T}\sum_{n:m_n=m}\eta_{n0}, \quad \beta_k = \frac{1}{2\rho}\left(-b_k + \sqrt{b_k^2 + 4\rho\sum_{n=1}^{N}\eta_{nk}}\right),$$

where

$$b_k = \sum_{n=1}^{N}\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)(K(t_n - t_l) - K(t_{n-1} - t_l)) + \rho(u_k^i - z_k^i).$$

Complexity Analysis. The majority of our computation lies in the estimation of μ and β, where we need to calculate a vector of η for each event n. Since feature-related computations such as $\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)\kappa(t_n - t_l)$ and $\sum_{n=1}^{N}\sum_{l=1}^{n-1}x_{m_l,m_n,k}(t_n)(K(t_n - t_l) - K(t_{n-1} - t_l))$ can be done ahead, this estimation procedure has a computational cost of $O(N * K + M)$ only. The updates of $\mathbf{z}$ and $\mathbf{u}$

in each iteration only cost $O(K)$. Thus, our algorithm costs $O(N * K + M)$ in total, where $K \ll N$ can be ensured by controlling by the number of features we use. Note that M is the number of dimensions, N is the number of events, thus we can view the computational cost as linear in the number of events and the number of individuals, such cost is much smaller compared with multi-dimensional Hawkes models that estimate pairwise infectivity directly, which cost at least $O(N^2 + M^2)$.

Time-varying Features

Time-varying features (Swan and Allan 1999) attract ever increasing attentions in analyzing temporal data, such as email communication (Perry and Wolfe 2013), seismic events (Cardenas-Pena, Orozco-Alzate, and Castellanos-Dominguez 2013), and Heart Rate Variability (HRV) signals (Mendez et al. 2010). In a given social network where memes diffuse, our paper collect both *individual features*, which imply the instant self-properties of each individual, and *dyadic features*, which imply the instant relationship between each pair of individuals. These features count the number of appearances of a certain pattern involving one individual or one individual-pair in a certain time range formulated as:

$$x(p)(t, \Delta t) = \#\{p \in [t - \Delta t, t)\},$$

where p represents a certain defined pattern, $[t - \Delta t, t)$ is the time interval from some ancient timestamp to the current timestamp. Table 1 shows several patterns we adopt in this paper. Our feature design is inspired by the features proposed in (Perry and Wolfe 2013). The novelty of our design is that we propose features in more general forms, and also explore brand-new patterns in networks, thus produce far more features.

As shown in Table 1, our features generally originate from individuals' involvements in the diffusion paths of memes in networks, and reflect implicit individual property or pairwise relationship. These features can be either categorized by the number of individuals involved, or by the path length. If provided with explicit self-properties of individuals or the relationship between individuals, we can propose new features accordingly. Based on above collected features, we are able to form a feature vector $\mathbf{x}_{m,m'}(t)$ for each individual-pair (m, m') at any given timestamp t through:

$$\mathbf{x}_{m,m'}(t) = \{x(p)(t, \Delta t) | p \in \mathcal{P}_{m,m'}, \Delta t > 0\},$$

where $\mathcal{P}_{m,m'}$ refers to the set of patterns involving at least one individual among $\{m, m'\}$. Thus for each timestamp t, we have a unique set of feature vectors $\{\mathbf{x}_{m,m'}(t)\}$ utilized in intensity function $\lambda(t)$.

Experiments

We conducted experiments on both synthetic and real-world data sets , and compared the performance of our model with alternatives to demonstrate the effectiveness of our model.

Table 1: Patterns in Constructing Time-varying Features

Pattern p	Description
i	node i appears on one diffusion path.
$\mathrm{dist}(i)$	node i appears on one diffusion path of a certain meme(the appearance on the path of the same meme will not be counted twice).
$in(i)$	node i gets infected by another node (the appearance of the same node will not be counted twice).
$out(i)$	node i infects another node (the appearance of the same node will not be counted twice).
$i \circlearrowleft^{(v)}$	there exists a length-v diffusion path from node i to itself.
$\mathrm{pure}(i \circlearrowleft^{(v)})$	there exists a length-v diffusion path from node i to itself, and there exists one meme that diffuses on the entire path (Similar patterns are designed for all dyad-dependent patterns below).
$i \xrightarrow{(v)} j$	$v-1$ intermediate nodes exist on the diffusion path from node i to j.
$j \xleftarrow{(v)} i$	$v-1$ intermediate nodes exist on the diffusion path from node j to i.
$i \xleftrightarrow{(v,v')} j$	there exists a node h that is the ancestor of both node i and j, and the corresponding path length is v and v', respectively.
$i \overset{(v,v')}{\rightarrow\leftarrow} j$	there exists a node h that is the descendant of both node i and j, and the corresponding path length is v and v', respectively.

To facilitate the description of each pattern, we take a social network as a graph, and each individual as a node, and the paths that memes diffuse as directed edges.

Synthetic Data

Data set. We sample the synthetic data according to the proposed model in the following manner: Given model dimensions (M, N, K), we start by drawing the basic intensity vector μ of size M, and the coefficient vector β of size K. Each element μ_m and β_k is randomly generated in $[0.5\hat{\mu}, 1.5\hat{\mu}]$ and $[0, 2\hat{\beta}]$ respectively before simulation. Then we randomly draw a fixed feature vector $\mathbf{x}_{m,m'}$ for each pair of dimensions m and m', and finally sample event cascades from the proposed model specified by μ, β, and $\mathbf{x}$. We also generate the ground-truth infectivity matrix $\hat{\mathbf{A}}$ based on the ground-truth β and $\mathbf{x}$. Our synthetic data are simulated with two different settings:

- `Small`: M=100, N=1,200, K=10, $\hat{\mu}$=0.01, $\hat{\beta}$=0.05. Simulations were run 100 times.
- `Large`: M=1,000, N=50,000, K=100, $\hat{\mu}$=0.01, $\hat{\beta}$=0.005. Simulations were run 5 times.

We sample 100 cascades to ensure that normal multi-dimensional Hawkes models can obtain promising results, which our model doesn't necessarily need as shown in experiments. To test the how the lasso regularization works, we generate `Sparse Synthetic` data with a sparse β by randomly selecting 80% elements in the vector β to be 0.We also generate `Time-varying Synthetic` data with time-varying feature vectors. For each timestamp in a event cascade, we calculate a separate $\mathbf{x}_{m,m'}$ based on the

generative process of the proposed time-varying features in the network, thus ensure $\mathbf{x}_{m,m'}$ to be time-varying.

Evaluation metrics. We consider the following evaluation metrics: 1) first, we compare the average log probability on the training data, and the average log predictive likelihood on events falling in the final 10% of the total time of each event cascade; 2) next we compare the average relative distance between the estimated parameters and ground-truth ones by $\frac{1}{K}\sum_k |\frac{\beta_k - \hat{\beta}_k}{\beta_k}|$ and $\frac{1}{M}\sum_m |\frac{\mu_m - \hat{\mu}_m}{\mu_m}|$, and evaluate the learned infectivity α by $\frac{1}{M(M-1)}\sum_{ij:i\neq j} |\frac{\alpha_{ij} - \hat{\alpha}_{ij}}{\alpha_{ij}}|$. We classify these three metrics for parameter estimation into the class of Mean Absolute Error (MAE). 3) we also employ the metric RankCorr (Zhou, Zha, and Song 2013a), which is defined as the averaged Kendall's rank correlation coefficient between each row of $\mathbf{A}$ and $\hat{\mathbf{A}}$. It measures whether the relative order of the estimated social infectivities is correctly recovered or not.

Baselines. To demonstrate the effectiveness of the proposed model, we compare it with the following alternatives:

`Multi-Hawkes:` This is a normal multi-dimensional Hawkes model with no regularizer on $\mathbf{A}$.

`Cox:` This is a multiplicative Cox model that parameterizes intensity. Our experiments learn this model using the same feature set as our proposed model. Note that Cox has no parameter μ (Perry and Wolfe 2013).

`LowRankSparse:` This is a multi-dimensional Hawkes model with the infectivity matrix $\mathbf{A}$ regularized by both nuclear norm and ℓ_1 norm (Zhou, Zha, and Song 2013a).

`NetRate:` This is a continuous time model for diffusion networks (Rodriguez, Balduzzi, and Scholkopf 2011). It cannot model the recurrent events, thus we only keep the first event occurrence at each individual.

`Para-Hawkes:` This is our proposed model, besides estimating μ and β, we also infer the infectivity matrix $\mathbf{A}$ accordingly for the comparison with Hawkes models which directly estimate infectivity,

Model Fitness on Synthetic Data. Table 2 compares the performance of the proposed model with several alternative point process models measured by both likelihood and the accuracy of parameter estimation. On synthetic data simulated with non-sparse β, Para-Hawkes fits the data better than Cox, while Cox performs better than Multi-Hawkes. On synthetic data simulated with sparse β, Para-Hawkes performs better than the non-sparse case, which demonstrates that the lasso regularization on coefficients β does work. The performance of Multi-Hawkes is rarely affected since the sparsity of β only influences the relationship within $\mathbf{A}$, which is ignored by Multi-Hawkes. Cox performs worse, as it imposes no regularization on coefficients. On larger synthetic data, the advantage of Para-Hawkes over others become greater. This illustrates that Para-Hawkes is adept in modeling more complexity diffusion networks .

Fitness on Synthetic Data with Time-varying Infectivity. Table 2 also shows that using time-varying features instead of invariant features slightly harms the performances of Para-Hawkes and Cox, while Multi-Hawkes performs poor, which illustrates the advantage of estimating coefficients β rather than the infectivity α directly . The degree of degeneration of the performance of Para-Hawkes is smaller than that of Cox, which proves that Para-Hawkes is more suitable for networks with time-varying infectivity.

Table 2: Model Fitness on Synthetic Data

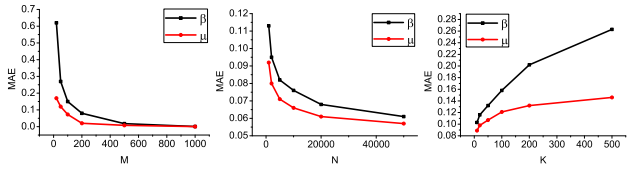
Data set	Metric	P-Hawkes	M-Hawkes	Cox
`S-Synthetic`	Training	-73.91	-89.13	-79.82
	Predictive	-136.23	-151.21	-143.21
	MAE(β or α)	0.103	0.257	0.148
	MAE(μ)	0.089	0.116	
`L-Synthetic`	Training	-107.89	-172.21	-135.95
	Predictive	-190.26	-310.85	-233.90
	MAE(β or α)	0.120	0.342	0.161
	MAE(μ)	0.113	0.148	
`S-Sparse`	Training	-70.73	-89.72	-80.27
	Predictive	-133.91	-151.30	-144.84
	MAE(β or α)	0.094	0.258	0.157
	MAE(μ)	0.086	0.117	
`L-Sparse`	Training	-102.46	-172.26	-137.27
	Predictive	-182.91	-310.81	-239.64
	MAE(β or α)	0.116	0.344	0.168
	MAE(μ)	0.102	0.149	
`S-T-varying`	Training	-81.62	-176.32	-97.28
	Predictive	-140.83	-418.20	-172.74
	MAE(β or α)	0.115	0.923	0.165
	MAE(μ)	0.104	0.363	
`L-T-varying`	Training	-122.43	-218.38	-160.92
	Predictive	-207.22	-693.67	-269.30
	MAE(β or α)	0.131	1.327	0.184
	MAE(μ)	0.128	0.616	

In the column of "Metric", "Training" stands for training likelihood, while "Predictive" stands for predictive likelihood. "P-Hawkes" stands for Para-Hawkes, "M-Hakwes" stands for Multi-Hawkes, "S-" stands for data setting `Small`, "L-" stands for `Large`. "T-varying" stands for `Time-varying`.

Model Dimension Variation. Figure 2 shows how the variation in the setting of model dimensions influences the fitness of the proposed model on the synthetic data. *When increasing the number of dimensions M and fixing all other model dimensions*, the error in both the learning of coefficients β and the estimation of Hawkes parameter μ will be significantly reduced. *Secondly, along with the increase of events N*, the proposed model fits the synthetic data better. *When the number of features K increases*, Para-Hawkes finds it more and more difficult to fit the synthetic data.

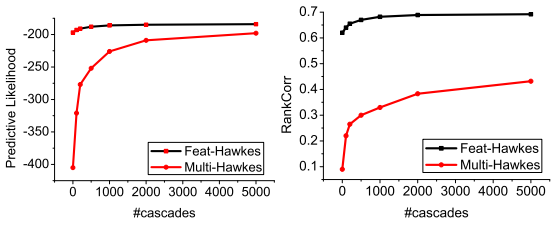
How the Number of Cascades Affects Performance. Figure 3 shows that when the number of cascades increases, both the data fitness and the accuracy of the social infectivity estimation of Para-Hawkes are rarely affected, while Multi-Hawkes performs significantly better. However, even trained with a large number of event cascades, Para-Hawkes still performs much better than Multi-Hawkes. Such phenomenon demonstrates that the proposed model works well without multiple cascades, while a normal multi-dimensional Hawkes model requires a large number of cascades to gain a satisfactory performance.

Coefficient Learning on Synthetic Networks with Various Topologies. This series of experiments sample event



(a) M (b) N (c) K

Figure 2: How the variation in model dimension influences the fitness of the proposed model on the synthetic data.



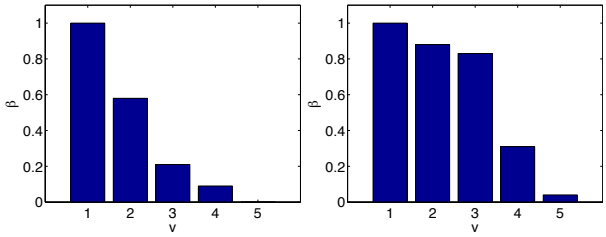
(a) Predictive Likelihood (b) RankCorr

Figure 3: Performance Comparison wrt. Cascade Number.

cascades from the normal multi-dimensional Hawkes process specified by sparsity and low-rank $\mathbf{A}$'s, respectively, and estimate the coefficients in our proposed model on both data sets to explore the appropriate set of features for modeling different network topologies. Figure 4 shows that, when characterized by different path lengths v, our features weight different in modeling various network topologies. In a sparse social network, the weights of features characterized by a short path are larger, while in a low-rank network, the weights of features characterized by a medium-length path are relatively more significant. One explanation can be that, in a very sparse network, individuals are more unlikely to influence each other via middlemen than in a low-rank network where people form groups, and influence every other group members. Moreover, in both networks, the weights of features characterized by paths of length 1 are the highest, which emphasizes the importance of individuals' direct interaction in determining social influence; the weights of features characterized by paths of length>3 are negligible, which illustrates that too much middlemen can greatly weaken the influence between individuals, e.g. two people connected by over three acquaintances are rarely acquainted each other.

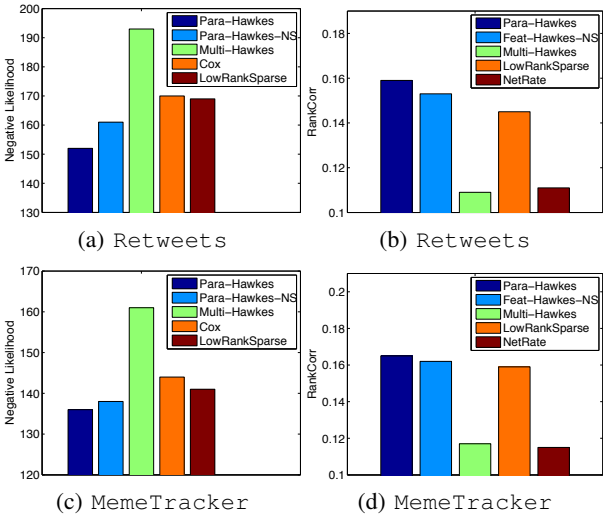
Real World Dataset

To further study how our model works in real world social networks, we apply the proposed model on `Retweets` and `MemeTracker` data sets. The `Retweets` data set contains the timestamped information flowing among tweet users. When a new post is issued by some user, other users will retweet this post or those retweets. In this way, the content of the original post diffuses in the network, and all the timestamped retweets concerning that post form an event cascade. From the `Retweets` data set, we extract 5000 most popular posts diffusing among around 5000 users. The `MemeTracker` data set contains the timestamped information flows captured by hyper-links among different sites. These timestamped hyperlinks form an event cascade for the particular piece of information flowing among numerous web sites. In particular, we extract a network consisting of top 500 sites with all hyperlinks among them.



(a) Sparse (b) Low-rank

Figure 4: Coefficients Learned on Synthetic Networks with Different Topologies. The Y axis denotes the average value of the learned coefficients of features characterized by path length v. These average values are scaled to the range of $[0, 1]$ to clarify the comparison of relative importance of different features.



(a) `Retweets` (b) `Retweets`

(c) `MemeTracker` (d) `MemeTracker`

Figure 5: Performance Comparison on Real World Data Sets.

Figure 5 compares the performance of Para-Hawkes with baselines measured by both predictive likelihood and RankCorr. Notice that both real datasets have ground-truth, from which we can derive the relative order of influence between individual and accordingly calculate the RankCorr score. In this series of experiments, we add a new model named Para-Hawkes-NS, which is our proposed model with no lasso regularizer. From Figure 5, we can see that our proposed models perform better the all compared baselines, which demonstrates the effectiveness of using time-varying features. The advantage over LowRankSparse illustrates that appropriate weighting of generic features can capture specific network topologies, such as sparsity and low-rank structure. Our advantage over LowRankSparse on `Retweets` is much larger than that on `MemeTracker`.

One explanation may be that the proposed model suits various networks while using no prior knowledge of the network topology. Methods using topological priors only work in networks with some specific structure. Moreover, our performance advantage measured by likelihood is much greater than that measured by RankCorr, which implies that the proposed model is capable of precisely modeling observed diffusion, rather than just predicting the relative significance of pairwise infectivities. We also find that a thresholding of the inferred $\mathbf{A}$ with a small constant will result in an infectivity matrix with sparsity degree similar as that learned by LowRankSparse. Meanwhile, Para-Hawkes performs better than Para-Hawkes-NS, which illustrates the importance of selecting effective features among all designed features.

Conclusion and Future Work

We propose a novel multi-dimensional Hawkes model that parameterizes pairwise infectivities using linear combinations of time-varying features. Alternating direction method of multipliers (ADMM) is employed to estimate the proposed features' coefficients, which are regularized by a ℓ_1 norm to select effective features. In future work, it would be interesting to consider additional time-varying features, and investigate the performance of the proposed model in other kinds of social networks. Moreover, we'll study the coefficient learning for the messy structure in real networks.

Acknowledgments

This work is supported in part by NIH 1R01GM108341, NSF IIS 1116886 and NSFC 61129001.

References

Ait-Sahalia, Y.; Cacho-Diaz, J.; and Laeven, R. 2010. Modeling financial contagion using mutually exciting jump processes. *Tech. rep.*

Boyd, S. 2010. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1):1–122.

Cardenas-Pena, D.; Orozco-Alzate, M.; and Castellanos-Dominguez, G. 2013. Selection of time-variant features for earthquake classification at the nevado-del-ruiz volcano. *Comput. Geosci.* 51:293–304.

Donoho, D. L., and Johnstone, I. M. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90(432):1200–1224.

Eckstein, J., and Bertsekas, D. P. 1992. On the douglas rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming* 55(1):293–318.

Errais, E.; Giesecke, K.; and Goldberg, L. R. 2010. Affine point processes and portfolio credit risk. *SIAM J. Fin. Math.* 1(1):642–665.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90.

Hunter, D. R., and Lange, K. 2004. A tutorial on mm algorithms. *Amer. Statist* 30–37.

Li, L., and Zha, H. 2013. Dyadic event attribution in social networks with mixtures of hawkes processes. In *Proceedings of the 22nd ACM International Conference on Conference on Information; Knowledge Management*, CIKM '13, 1667–1672. New York, NY, USA: ACM.

Mendez, M. O.; Matteucci, M.; Castronovo, V.; Strambi, L. F.; Cerutti, S.; and Bianchi, A. M. 2010. Sleep staging from heart rate variability: time-varying spectral features and hidden markov models. *International Journal of Biomedical Engineering and Technology* 3:246–263.

Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association.* 83(401):9–27.

Perry, P. O., and Wolfe, P. J. 2013. Point process modeling for directed interaction networks. In *Journal of the Royal Statistical Society*, volume 8, 821–849.

Robins, G.; Pattison, P.; Kalish, Y.; and Lusher, D. 2007. An introduction to exponential random graph models for social networks. *Social Networks* 29(2):173–191.

Rodriguez, M. G.; Balduzzi, D.; and Scholkopf, B. 2011. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 561–568. New York, NY, USA: ACM.

Stomakhin, A.; Short, M. B.; and Bertozzi, A. L. 2011. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems.* 27(11).

Swan, R., and Allan, J. 1999. Extracting significant time varying features from text. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, 38–45. New York, NY, USA: ACM.

Yang, S., and Zha, H. 2013. Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, 1–9.

Z.-Mangion, A.; Dewarc, M.; Kadirkamanathand, V.; and Sanguinetti, G. 2012. Point process modelling of the afghan war diary. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, volume 109, 12414–12419.

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of 16th International Conference on Artificial Intelligence and Statistics (AISTATS-13)*, volume 31, 641–649.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, 1301–1309.

Zhuang, J.; Ogata, Y.; and Jones, D. V. 2002. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association.* 97(458):369–380.