A Related Works

Hawkes Process Hawkes process has long been used to model event sequences (Hawkes, 1971), such as earthquake aftershock sequences (Ogata, 1999), financial transactions (Bauwens and Hautsch, 2009), and events on social networks (Fox et al., 2016; Farajtabar et al., 2017). Its variant, mixture of Hawkes processes model, has also been proved effective in many area (Yang and Zha, 2013; Li and Zha, 2013; Xu and Zha, 2017). In most cases, the learning methodology is variational inference or maximum likelihood estimation (Rasmussen, 2013; Zhou et al., 2013; Zhao et al., 2015). Other possible methods includes least-squares-based method (Eichler et al., 2017), Wiener-Hopf-based methods (Bacry et al., 2012), and cumulants-based methods (Achab et al., 2017).

Instead of predefine an impact function here, some non-parametric methods use discretization or kernel-estimation when learning models (Reynaud-Bouret et al., 2010; Zhou et al., 2013; Hansen et al., 2015). Those methods usually target small datasets, and do not need a good scalability. Recently, some attempts have been made to further enhance the flexibility of Hawkes processes. The time-dependent Hawkes process (TiDeH) in Kobayashi and Lambiotte (2016) and the neural network-based Hawkes process in Mei and Eisner (2017) learn very flexible Hawkes processes with complicated intensity functions. Those methods usually target very long and multi-dimensional sequences, instead of short sequences.

Existing works targeting short sequences is usually in specific cases (Xu et al., 2017a,b), such as the data is censored. However, there is no work targeting general short sequences as we do here.

There are lines of research that involves both point processes and graphs. One is using point process to find the latent graph (Blundell et al., 2012; Linderman and Adams, 2014; Tran et al., 2015). Another one is considering the interaction of the nodes as point process and use it to construct a dynamic graph, instead of the event happens on nodes as we consider here (Farajtabar et al., 2016; Zarezade et al., 2017; Trivedi et al., 2018). These works have vary different aims from our work.

Meta Learning Meta learning has been studied since last century (Bengio et al., 1990; Chalmers, 1991). Some works focus on learning the hyperparameters, such as learning rates or initial conditions (Maclaurin et al., 2015). Some works aim to learn a metric so that a simple K nearest neighbors can perform well under such a metric (Koch et al., 2015; Vinyals et al., 2016; Sung et al., 2018; Snell et al., 2017). Some works design specific deep neural networks so that the information of different tasks are memorized and thus the model can easily generalize to new tasks (Santoro et al., 2016; Munkhdalai and Yu, 2017; Ravi and Larochelle, 2016).

Model-Agnostic Meta Learning (MAML) method (Finn et al., 2017) opens another line of research, i.e., it designs an optimization scheme so that the model can fast adapt to new tasks. Reptile (Nichol and Schulman, 2018), a variant of MAML, is proposed to simplify the computation of MAML. None of those works, however, considers the relational information between tasks like our method, which is critical in modeling short sequences.

One interesting line of follow-up works of MAML is connecting MAML with Bayesian inference (Finn et al., 2018; Ravi and Beatson, 2018; Grant et al., 2018). Since HARMLESS combines a Bayesian model with MAML, it has the potential to be rewritten into a pure Bayesian model that has better quantification of uncertainty. We left this for future work.

B Definition of Operator $\mathcal{D}$

As we mentioned earlier,

$$\min_{\theta} \sum_{\mathcal{T}_i \in \Gamma} \mathcal{F}_{\mathcal{T}_i}(\widetilde{\theta}_i) = \sum_{\mathcal{T}_i \in \Gamma} \mathcal{F}_{\mathcal{T}_i}(\theta - \eta \mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta))$$

is the loss function for MAML, FOMAML, and Reptile algorithm with different definition of the operator $\mathcal{D}$.

For simplicity, here we define the operator of one gradient step. The cases of few gradient steps can be defined analogously.

For MAML, $\mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$ is defined as $\nabla_\theta(\mathcal{F}_{\mathcal{T}_i}(\theta))$.

For First Order MAML (FOMAML), $\mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$ is also defined as $\nabla_\theta(\mathcal{F}_{\mathcal{T}_i}(\theta))$. The difference is that the output of the operator just a value, not a function of θ, i.e., when we solve the gradient of $\mathcal{F}_{\mathcal{T}_i}(\theta - \eta\mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta))$, the gradient does not back-propagate into $\mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$.

For Reptile, the algorithm of reptile is as follows Nichol and Schulman (2018).

Algorithm 1 Reptile

```
while not converged do
    Sample task T with loss F_T;
    W ← SGD(F_T, θ, k), where k is the number of SGD steps;
    Do the update θ ← θ − η(θ − W);
end while
```

From the algorithm we can see, operator $\mathcal{D}$ is defined as $\mathcal{D}(\mathcal{F}_{\mathcal{T}}, \theta) = \text{SGD}(\mathcal{F}_{\mathcal{T}}, \theta, 1)$. Similar as FOMAML, computing the gradient also does not back-propagate into $\mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$.

C Derivation of Variational EM

Preparation After adding latent variable $\boldsymbol{z}$, the joint distribution is

$$p(\boldsymbol{T}, \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi}) = p(\boldsymbol{T}|\boldsymbol{z})p(\boldsymbol{Y}|\boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow})p(\boldsymbol{z}|\boldsymbol{\pi})p(\boldsymbol{z}_{\leftarrow}|\boldsymbol{\pi})p(\boldsymbol{z}_{\rightarrow}|\boldsymbol{\pi})p(\boldsymbol{\pi}).$$

where

$$\begin{aligned}
p(\boldsymbol{T}|\boldsymbol{z}) &= \prod_{i=1}^{N}\prod_{k=1}^{K}\left(\mathcal{L}_i(\theta_k - \eta\mathcal{D}(\mathcal{L}_i, \theta_k))\right)^{z_{i,k}}, \\
p(\boldsymbol{Y}|\boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}) &= \prod_{i=1}^{N}\prod_{j=1}^{N}(z_{i\rightarrow j}^T\boldsymbol{B}z_{i\leftarrow j})^{Y_{ij}}(1 - z_{i\rightarrow j}^T\boldsymbol{B}z_{i\leftarrow j})^{1-Y_{ij}} \\
p(\boldsymbol{z}|\boldsymbol{\pi}) &= \prod_{i=1}^{N}\prod_{k=1}^{K}\pi_{i,k}^{z_{i,k}}, \\
p(\boldsymbol{z}_{\rightarrow}|\boldsymbol{\pi}) &= \prod_{i=1}^{N}\prod_{j=1}^{N}\prod_{k=1}^{K}\pi_{i,k}^{z_{i\rightarrow j,k}}, \\
p(\boldsymbol{z}_{\leftarrow}|\boldsymbol{\pi}) &= \prod_{i=1}^{N}\prod_{j=1}^{N}\prod_{k=1}^{K}\pi_{j,k}^{z_{i\leftarrow j,k}}, \\
p(\boldsymbol{\pi}) &= \prod_{i=1}^{N}\text{Dirichlet}(\pi_i|\alpha) = \prod_{i=1}^{N}C(\alpha)\prod_{k=1}^{K}\pi_{i,k}^{\alpha-1}.
\end{aligned}$$

Note that in this section we represent $z_i, z_{i\rightarrow j}, z_{i\leftarrow j}$ as one-hot vector, while in the main paper we use scalar $z_i = k$ representing the identities.

The posterior distribution is defined as

$$p(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi}|\boldsymbol{T}, \boldsymbol{Y}, \alpha, \boldsymbol{\theta}, B).$$

We aim to find a distribution $q(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi}) \in Q$, such that the Kullback-Leibler (KL) divergence between the above posterior distribution and $q(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi})$ is minimized. This can be achieved by maximize the Evidence Lower BOund (ELBO),

$$\mathcal{B}(q) = \mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{Y})] - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi})].$$

Variational family We adopt the mean-field variational family, i.e.,

$$q(\boldsymbol{z}, \boldsymbol{z}_{\rightarrow}, \boldsymbol{z}_{\leftarrow}, \boldsymbol{\pi}) = q_1(\boldsymbol{\pi})\prod_i q_2(z_i)\prod_j q_3(z_{i\rightarrow j})q_4(z_{i\leftarrow j}).$$

We pick $q_1(\pi_i)$ as PDF of Dirichlet(β), $q_2(z_i)$ as PDF of Categorical(γ_i), $q_3(z_{i\to j})$ as PDF of Categorical(ϕ_{ij}), $q_4(z_{i\leftarrow j})$ as PDF of Categorical(ψ_{ij}).

Update for q_1 Again, our goal is to maximize

$$\mathcal{B}(q) = \mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{z}_\to, \boldsymbol{z}_\leftarrow, \boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{Y})] - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{z}_\to, \boldsymbol{z}_\leftarrow, \boldsymbol{\pi})].$$

Now we focus on q_1, and treat q_2, q_3 and q_4 as given. We want to maximize

$$\begin{aligned}
\mathcal{F}_{\boldsymbol{\pi}}(q_1) &= \mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{z}_\to, \boldsymbol{z}_\leftarrow, \boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{Y})] - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{z}_\to, \boldsymbol{z}_\leftarrow, \boldsymbol{\pi})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{T}|\boldsymbol{z}) + \log p(\boldsymbol{Y}|\boldsymbol{z}_\leftarrow, \boldsymbol{z}_\to) + \log p(\boldsymbol{z}|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\leftarrow|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\to|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi})] \\
&\quad - \mathbb{E}_{q_1}[\log q_1(\boldsymbol{\pi})] + \text{const} \\
&= \mathbb{E}_q[\log p(\boldsymbol{z}|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\leftarrow|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\to|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi})] - \mathbb{E}_{q_1}[\log q_1(\boldsymbol{\pi})] + \text{const} \\
&= \int q_1(\boldsymbol{\pi}) \left(\mathbb{E}_{q_2}[\log p(\boldsymbol{z}|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\leftarrow|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\to|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi})] - \log q_1(\boldsymbol{\pi})\right) d\boldsymbol{\pi} + \text{const}.
\end{aligned}$$

Take the derivative,

$$\frac{\delta \mathcal{F}_{\boldsymbol{\pi}}(q_1)}{\delta q_1} = \mathbb{E}_{q_2}[\log p(\boldsymbol{z}|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\leftarrow|\boldsymbol{\pi}) + \log p(\boldsymbol{z}_\to|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi})] - \log q_1(\boldsymbol{\pi}) - 1 = 0.$$

Substitute the expressions of the distributions, after some derivation we get the update for $\boldsymbol{\beta}$ as

$$\beta_{i,k} \leftarrow \alpha_k + \gamma_{i,k} + \sum_{j=1}^N \phi_{ij,k} + \sum_{j=1}^N \psi_{ij,k}. \tag{15}$$

Update for q_2 Similarly, we have

$$\begin{aligned}
\mathcal{F}_{\boldsymbol{z}}(q_2) &= \mathbb{E}_q[\log p(\boldsymbol{T}|\boldsymbol{z}) + \log p(\boldsymbol{z}|\boldsymbol{\pi})] - \mathbb{E}_{q_2}[\log q_2(\boldsymbol{z})] + \text{const} \\
&= \int q_2(\boldsymbol{z}) \left(\mathbb{E}_{q_1}[\log p(\boldsymbol{T}|\boldsymbol{\theta}, \boldsymbol{z}) + \log p(\boldsymbol{z}|\boldsymbol{\pi})] - \log q_2(\boldsymbol{z})\right) d\boldsymbol{z} + \text{const}.
\end{aligned}$$

Take the derivative,

$$\frac{\delta \mathcal{F}_{\boldsymbol{z}}(q_2)}{\delta q_2} = \log p(\boldsymbol{T}|\boldsymbol{\theta}, \boldsymbol{z}) + \mathbb{E}_{q_1}[\log p(\boldsymbol{z}|\boldsymbol{\pi})] - \log q_2(\boldsymbol{z}) - 1 = 0.$$

After some derivation, we have

$$\gamma_{i,k} \leftarrow \mathcal{L}_i(\theta_k - \eta \mathcal{D}(\mathcal{L}_i, \theta_k)) \exp\left(f_{\text{dg}}(\beta_{i,k}) - f_{\text{dg}}(\sum_\ell \beta_{i,\ell})\right), \tag{16}$$

$$\gamma_{i,k} \leftarrow \frac{\gamma_{i,k}}{\sum_\ell \gamma_{i,\ell}}, \tag{17}$$

where f_{dg} is the digamma function.

Update for q_3 **and** q_4 The derivation of update for q_3 and q_4 is very similar to the update for q_2, so we will not elaborate on that. Readers who are interested might also refer to Airoldi et al. (2008). The updates are

$$\phi_{ij,k} \leftarrow e^{\mathbb{E}_q[\log \pi_{i,k}]} \prod_{\ell=1}^K \left(B_{k\ell}^{Y_{ij}} (1 - B_{k\ell})^{1-Y_{ij}}\right)^{\psi_{ij,\ell}}, \quad \phi_{ij,k} \leftarrow \frac{\phi_{ij,k}}{\sum_\ell \phi_{ij,\ell}}, \tag{18}$$

$$\psi_{ij,\ell} \leftarrow e^{\mathbb{E}_q[\log \pi_{j,\ell}]} \prod_{k=1}^K \left((B_{k\ell})^{Y_{ij}} (1 - B_{k\ell})^{1-Y_{ij}}\right)^{\phi_{ij,k}}, \quad \psi_{ij,k} \leftarrow \frac{\psi_{ij,k}}{\sum_\ell \psi_{ij,\ell}}, \tag{19}$$

Update for $\boldsymbol{\theta}$ We update $\boldsymbol{\theta}$ using gradient ascent. We first pick the terms that is relevant to $\boldsymbol{\theta}$,

$$\begin{aligned}
\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) &= \mathbb{E}_q[\log p(\boldsymbol{T}|\boldsymbol{\theta}, \boldsymbol{z})] + \text{const} \\
&= \int q_2(\boldsymbol{z})[\log p(\boldsymbol{T}|\boldsymbol{\theta}, \boldsymbol{z})] d\boldsymbol{z} + \text{const} \\
&= \sum_{i=1}^N \sum_{k=1}^K \gamma_{i,k} \log \mathcal{L}_i(\theta_k - \eta \mathcal{D}(\mathcal{L}_i, \theta_k)) + \text{const}.
\end{aligned}$$

So the gradient ascent update is,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta_1 \nabla_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{i,k} \log \mathcal{L}_i(\theta_k - \eta \mathcal{D}(\mathcal{L}_i, \theta_k)) \right). \tag{20}$$

Update for α and B From Airoldi et al. (2008), we have the update for α and $\boldsymbol{B}$ as follows

$$\alpha_k \leftarrow \alpha_k + \eta_\alpha \left(N\big(f_{\mathrm{dg}}(\sum_\ell \alpha_\ell) - f_{\mathrm{dg}}(\alpha_k)\big) + \sum_{i=1}^{N} \big(f_{\mathrm{dg}}(\beta_{i,k}) - f_{\mathrm{dg}}(\sum_\ell \beta_{i,\ell})\big) \right), \tag{21}$$

$$B_{k\ell} \leftarrow \frac{\sum_{ij} Y_{ij} \phi_{ij,k} \psi_{ij,\ell}}{\sum_{ij} \phi_{ij,k} \psi_{ij,\ell}}, \tag{22}$$

D Derivation of Evaluation Metric

In this section, we give more details on the evaluate metrics. Specifically, we show how to compute the NLL of the test set. Given a sequence $\boldsymbol{\tau}_i = \{\tau_i^{(1)}, \tau_i^{(2)}, \cdots, \tau_i^{(M_i)}\}$, we would like to predict the timestamp of $\tau_i^{(M_i+1)}$. Here, we use the probability of the arrival at time $\tau_i^{(M_i+1)}$ and no arrival in $[\tau_i^{(M_i)}, \tau_i^{(M_i+1)}]$ given history before $\tau_i^{(M_i)}$ as evaluation metric.

Consider a Hawkes process with parameter θ, the probability density is

$$\begin{aligned}
\mathcal{P}(\theta) &= \lambda\big(\tau_i^{(M_i+1)}; \theta, \boldsymbol{\tau}_i)\big) \exp\Big(- \int_{\tau_i^{(M_i)}}^{\tau_i^{(M_i+1)}} \lambda(t; \theta, \boldsymbol{\tau}_i)\, dt \Big) \\
&= \big(\mu + \sum_{m=1}^{M_i} \delta\omega e^{-\omega(\tau_i^{(M_i+1)} - \tau_i^{(m)})}\big) \exp\Big(-\mu(\tau_i^{(M_i+1)} - \tau_i^{(M_i)}) - \delta(1 - e^{-\omega(\tau_i^{(M_i+1)} - \tau_i^{(M_i)})}) \Big).
\end{aligned}$$

In the generative process, for subject i, we first sample z_i, then use parameter $\widetilde{\theta}_{z_i}^{(i)} = \theta_{z_i} - \eta \mathcal{D}(\mathcal{L}_i, \theta_{z_i})$. The posterior distribution of z_i is $q_2(z_i)$, $i.e.$, Categorical(γ_i). Therefore we have

$$\mathbb{P}(z_i = k) = \gamma_{i,k}.$$

So the likelihood of next arrival $\tau_i^{(M_i+1)}$ is

$$\begin{aligned}
\widetilde{\mathcal{L}}_i &= \sum_{k=1}^{K} \mathbb{P}(z_i = k) \mathbb{P}(\text{next arrival is } \tau_i^{(M_i+1)} | \text{ Hawkes model with } \theta_k) \\
&= \sum_{k=1}^{K} \gamma_{i,k} \mathcal{P}(\widetilde{\theta}_k^{(i)}).
\end{aligned}$$

And then we sum $\widetilde{\mathcal{L}}_i$ over every subject.

E Detailed Settings of the Experiments

Note that we can also adopt a non-informative α instead of updating it in every iteration. After some trial experiments, we find setting $\alpha = \mathbf{1}_K$ is numerically more stable than updating it in every iteration. Therefore we adopt $\alpha = \mathbf{1}_K$ in the following experiments.

Besides, we find that ν causes nearly no effect to the result when varying from 10^{-10} to 10^{-1}. We fix it as 10^{-2}.

E.1 Synthetic Dataset

Both the baselines and our proposed methods are fine tuned. We first perform a coarse grid search to find hyper-parameters for all methods. The grid search finds learning rate from 1×10^{-7} to 1 for both inner and outer updates. To perform the multi-split procedure, all hyper-parameters are then selected in the following range listed in Table 4 and Table 5. For each range, we perform experiment on three values: the lower one, the upper one, and the middle one. Method *MTL* adopt $\nu_{\mathrm{mtl}} = 0.1$.

Table 4: Learning rates of experiments.

K_0		1	3	6	10
DMHP	lr.	$1 \pm .1 \times 10^{-3}$	$3 \pm .1 \times 10^{-3}$	$6.5 \pm .1 \times 10^{-3}$	$7 \pm .1 \times 10^{-3}$
Two Step	inner lr.	$1 \pm .1 \times 10^{-5}$	$5 \pm .1 \times 10^{-5}$	$5 \pm .1 \times 10^{-5}$	$1 \pm .1 \times 10^{-4}$
	outer lr.	$1 \pm .1 \times 10^{-3}$	$1 \pm .1 \times 10^{-2}$	$1.5 \pm .1 \times 10^{-2}$	$1 \pm .1 \times 10^{-2}$
HARMLESS	inner lr.	$5 \pm .1 \times 10^{-5}$	$5 \pm .1 \times 10^{-6}$	$2 \pm .1 \times 10^{-4}$	$7 \pm .1 \times 10^{-5}$
(MAML)	outer lr.	$6 \pm .1 \times 10^{-4}$	$2 \pm .1 \times 10^{-4}$	$6 \pm .1 \times 10^{-5}$	$4.5 \pm .1 \times 10^{-6}$
HARMLESS	inner lr.	$5 \pm .1 \times 10^{-4}$	$1 \pm .1 \times 10^{-5}$	$3 \pm .1 \times 10^{-5}$	$1.5 \pm .1 \times 10^{-6}$
(FOMAML)	outer lr.	$6 \pm .1 \times 10^{-4}$	$2 \pm .1 \times 10^{-4}$	$6 \pm .1 \times 10^{-5}$	$4.5 \pm .1 \times 10^{-6}$

Table 5: Learning rates of baseline experiments.

Method	Learning Rate
MLE-Sep	$5 \pm .1 \times \times 10^{-5}$
MLE-Com	$1 \pm .1 \times 10^{-3}$
MTL	$1 \pm .1 \times 10^{-3}$

E.2 Real Datasets

In this section, we introduce the experimental detail of the real datasets. We run our experiment with same inner and outer learning rate, denoted by η. For simplicity, we also set $\eta = \eta_\alpha = \eta_{\boldsymbol{\theta}}$, and search over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\} \otimes \{1, 2, 3, 4, 5\}$, where the element-wise product of two sets is defined as $A \otimes B = \{ab | a \in A, b \in B\}$. We search $K \in \{2, 3, 5\}$ and ν_{mtl} in range $\{0.1, 0.01, 0.001\}$. We perform grid search over the hyper-parameters, and obtain the candidate models. Then we perform multi-split procedure.

Because StackOverflow dataset is very large, it is too expensive to perform grid search. To accommodate this, we first split a validation set and a test set, then performing hyper-parameter search by flipping. Each experiment of StackOverflow dataset is run under 5 different settings.

In Table 6 we report one of the models that is picked by multi-split procedure. We remark that in most cases, the procedure picks only one model repeatedly.

Table 6: Settings of experiments.

data type	911-Calls	Linkedin	MathOverflow	StackOverflow
Baseline 1	$\eta = 4 \times 10^{-4}$	$\eta = 1 \times 10^{-3}$	$\eta = 5 \times 10^{-4}$	$\eta = 5 \times 10^{-4}$
Baseline 2	$\eta = 3 \times 10^{-}$	$\eta = 5 \times 10^{-3}$	$\eta = 1 \times 10^{-3}$	$\eta = 1 \times 10^{-3}$
MTL	$\eta = 3 \times 10^{-5}, \nu_{\text{mtl}} = 0.1$	$\eta = 1 \times 10^{-2}, \nu_{\text{mtl}} = 0.1$	$\eta = 4 \times 10^{-4}, \nu_{\text{mtl}} = 0.1$	$\eta = 5 \times 10^{-4}, \nu_{\text{mtl}} = 0.1$
DMHP	$\eta = 3 \times 10^{-5}, K = 2$	$\eta = 1 \times 10^{-3}, K = 3$	$\eta = 4 \times 10^{-3}, K = 3$	$N \backslash A$
MAML	$\eta = 3 \times 10^{-4}, K = 3$	$\eta = 5 \times 10^{-1}, K = 3$	$\eta = 3 \times 10^{-4}, K = 3$	$\eta = 1 \times 10^{-3}, K = 2$
FOMAML	$\eta = 3 \times 10^{-5}, K = 2$	$\eta = 1 \times 10^{-2}, K = 5$	$\eta = 2 \times 10^{-4}, K = 2$	$\eta = 4 \times 10^{-4}, K = 3$
Reptile	$\eta = 5 \times 10^{-3}, K = 2$	$\eta = 2 \times 10^{-1}, K = 3$	$\eta = 4 \times 10^{-2}, K = 2$	$\eta = 4 \times 10^{-2}, K = 2$

E.3 Ablation study

In this section we introduce the experimental detail of the ablation study. Specifically, the tuning process of the ablation study is as follows: We start from the same setting as the corresponding real experiment in previous section. For example, experiment *Remove graph (FOMAML)* corresponds to HARMLESS (FOMAML). We first use the same learning rate and K as HARMLESS (FOMAML) to perform experiment. If the experiment runs well, we adopt the experiment result. If the training does not converge, we decrease the learning rate and run again.

References

ACHAB, M., BACRY, E., GAÏFFAS, S., MASTROMATTEO, I. and MUZY, J.-F. (2017). Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research*, **18** 6998–7025.

Table 7: Learning rates of experiments of ablation study.

data type	LR
Remove inner heterogeneity ($K = 3$)	0.1
Remove inner heterogeneity ($K = 5$)	0.1
Remove grouping (MAML)	0.1
Remove grouping (FOMAML)	0.01
Remove grouping (Reptile)	0.2
Remove graph (MAML)	0.2
Remove graph (FOMAML)	0.005
Remove graph (Reptile)	0.2

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, **9** 1981–2014.

BACRY, E., DAYRI, K. and MUZY, J.-F. (2012). Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, **85** 157.

BAUWENS, L. and HAUTSCH, N. (2009). Modelling financial high frequency data using point processes. In *Handbook of financial time series*. Springer, 953–979.

BENGIO, Y., BENGIO, S. and CLOUTIER, J. (1990). *Learning a synaptic learning rule*. Université de Montréal, Département d'informatique et de recherche ?

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112** 859–877.

BLUNDELL, C., BECK, J. and HELLER, K. A. (2012). Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*.

BOX, G. E. and TIAO, G. C. (2011). *Bayesian inference in statistical analysis*, vol. 40. John Wiley & Sons.

CHALMERS, D. J. (1991). The evolution of learning: An experiment in genetic connectionism. In *Connectionist Models*. Elsevier, 81–90.

CLEEREMANS, A. and MCCLELLAND, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, **120** 235.

EICHLER, M., DAHLHAUS, R. and DUECK, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, **38** 225–242.

FARAJTABAR, M., YANG, J., YE, X., XU, H., TRIVEDI, R., KHALIL, E., LI, S., SONG, L. and ZHA, H. (2017). Fake news mitigation via point process based intervention. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

FARAJTABAR, M., YE, X., HARATI, S., SONG, L. and ZHA, H. (2016). Multistage campaigning in social networks. In *Advances in Neural Information Processing Systems*.

FINN, C., ABBEEL, P. and LEVINE, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

FINN, C., XU, K. and LEVINE, S. (2018). Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*.

FOX, E. W., SHORT, M. B., SCHOENBERG, F. P., CORONGES, K. D. and BERTOZZI, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, **111** 564–584.

GIRVAN, M. and NEWMAN, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99** 7821–7826.

Grant, E., Finn, C., Levine, S., Darrell, T. and Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*.

Hansen, N. R., Reynaud-Bouret, P., Rivoirard, V. et al. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, **21** 83–143.

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

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, **14** 1303–1347.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, **83** 016107.

Kobayashi, R. and Lambiotte, R. (2016). Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*.

Koch, G., Zemel, R. and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, vol. 2.

Laub, P. J., Taimre, T. and Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

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 Information & Knowledge Management*. ACM.

Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. In *International Conference on Machine Learning*.

Luo, D., Xu, H., Zhen, Y., Ning, X., Zha, H., Yang, X. and Zhang, W. (2015). Multi-task multi-dimensional hawkes processes for modeling event sequences. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Maclaurin, D., Duvenaud, D. and Adams, R. (2015). Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*.

Munkhdalai, T. and Yu, H. (2017). Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Nichol, A., Achiam, J. and Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Nichol, A. and Schulman, J. (2018). Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. In *Seismicity patterns, their statistical significance and physical meaning*. Springer, 471–507.

Paranjape, A., Benson, A. R. and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, **15** 623–642.

Ravi, S. and Beatson, A. (2018). Amortized bayesian meta-learning.

Ravi, S. and Larochelle, H. (2016). Optimization as a model for few-shot learning.

Reynaud-Bouret, P., Schbath, S. et al. (2010). Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, **38** 2781–2822.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B. and Bristow, V. L. (1996). *Stochastic processes*, vol. 2. Wiley New York.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. and Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning*.

Snell, J., Swersky, K. and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tran, L., Farajtabar, M., Song, L. and Zha, H. (2015). Netcodec: Community detection from individual activities. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM.

Trivedi, R., Farajtabar, M., Biswal, P. and Zha, H. (2018). Dyrep: Learning representations over dynamic graphs.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems*.

Xie, J., Kelley, S. and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, **45** 43.

Xu, H., Luo, D., Chen, X. and Carin, L. (2017a). Benefits from superposed hawkes processes. *arXiv preprint arXiv:1710.05115*.

Xu, H., Luo, D. and Zha, H. (2017b). Learning hawkes processes from short doubly-censored event sequences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Xu, H. and Zha, H. (2017). A dirichlet mixture model of hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems*.

Yang, S.-H. and Zha, H. (2013). Mixture of mutually exciting processes for viral diffusion. In *International Conference on Machine Learning*.

Zarezade, A., Khodadadi, A., Farajtabar, M., Rabiee, H. R. and Zha, H. (2017). Correlated cascades: Compete or cooperate. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A. and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Zhou, K., Zha, H. and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*.