

# Meta Learning with Relational Information for Short Sequences



Yujia Xie<sup>1</sup>, Haoming Jiang<sup>1</sup>, Feng Liu<sup>2</sup>, Tuo Zhao<sup>1</sup>, Hongyuan Zha<sup>1</sup> <sup>2</sup>Florida Atlantic University <sup>1</sup>Georgia Institute of Technology

#### HARMLESS – HAwkes Relational Meta LEarning for Short Sequence Introduction **Event** sequences Identity Proportion Vector • Key idea: identify and incorporate the relational information between tasks - The timestamps of tweets of a twitter user Sport 10% - Social graphs often exhibit community patterns Music 30% - The job hopping history of a person - Each subject may belong to multiple communities and thus Marvel 20% have multiple identities Short sequences $\rightarrow$ Assign each subject *i* a sum-to-one **identity proportion vector** – Sparse event pattern Alice $\pi_i \in [0,1]^K$ , where K is the number of communities - Job hopping histories **Mixture of Hawkes Process Model** Identity - Narrow observation window Mixture of Hawkes process model Alice - The criminal incidents after a regulation is published For the k-th identity of subject i, we adopt Hawkes process

• Challenging inference for short sequences

- - MLE for each sequence
    - Their lengths are insufficient for reliable inference.
  - Treat the collection of short sequences as i.i.d.
    - Highly biased against certain individuals.

# **Problem Setting**

- Given:
  - A collection of sequences  $T = \{ \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \cdots, \boldsymbol{\tau}_N \}$

- Grpah relational information among sequences, described by an  $N \times N$  adjacency matrix as  $\boldsymbol{Y}$ .



• **Goal**: Relational information helps predicting the occurrence of future events.

 $\mathcal{H}(\theta_{k}^{(i)})$  to model the timestamps of the associated events. The O MMM N likelihood for the *i*-th sequence  $\boldsymbol{\tau}_i$  is **Observed sequences**  $p(\boldsymbol{\tau}_i) = \sum_{k=1}^{\infty} \pi_{i,k} \mathcal{L}_i(\widetilde{\theta}_k^{(i)}).$ Sport **Mixed Membership Blockmodel** Music Mixed Membership stochastic Blockmodel (MMB) Marvel  $-z_{i \rightarrow j}$ : the identity of subject *i* when subject *i* approaches subject  $\bigcirc$  $-z_{i\leftarrow j}$ : the identity of subject j when j is approached by i  $\bigcirc$  $-z_{i \rightarrow j}^T B z_{i \leftarrow j}$ : the probability of whether subject i and j **Observed** connections have a connection • Generative process: • For each subject *i*, Draw a K dimensional identity proportion vector  $\pi_i \sim \text{Dirichlet}(\alpha)$ . — - Sample the *i*-th sequence  $\tau_i$  from the mixture of Hawkes processes in (1). • For each pair of subject *i* and *j*, - Draw identity indicator for the initiator  $z_{i \rightarrow j} \sim \text{Categorical}(\pi_i)$ - Draw identity indicator for the receiver  $z_{i \leftarrow j} \sim \text{Categorical}(\pi_j)$ - Sample whether there is an edge between *i* and *j*,  $Y_{ij} \sim \text{Bernoulli}(z_{i \rightarrow j}^T B z_{i \leftarrow j})$ .

Here, the observed variables are  $\tau_i$  and  $Y_{ij}$ . The parameters are  $\alpha$ ,  $\theta_k^{(i)}$ , and B. The latent variables are  $\pi_i$ ,  $z_i$ ,  $z_{i \rightarrow j}$  and  $z_{i \leftarrow j}$ .

• Meta inference for  $\theta$  and  $\theta$ . Instead of specifying that  $\theta_k^{(i)}$  is sampled from a prior distribution, we adapt the k-th common model  $\mathcal{H}(\theta_k)$  to sequence *i* using MAML-type updates,  $\tilde{\theta}_k^{(i)} = \theta_k - \eta \mathcal{D}(\log \mathcal{L}_i, \theta_k)$ . The gradient descent step on the log-likelihood of  $\theta$  can then be written as

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

B

(1)

## Background - Hawkes Process

 A Hawkes processes is a doubly stochastic temporal point process  $\mathcal{H}(\theta)$  with conditional intensity function  $\lambda = \lambda(t; \theta, \boldsymbol{\tau})$ defined as

$$\lambda(t; \theta, \boldsymbol{\tau}) = \mu + \sum_{\tau^{(m)} < t} \delta \omega e^{-\omega(t - \tau^{(m)})},$$
$$-\theta = \{\mu, \delta, \omega\},$$

 $-\mu$  is the base intensity,  $- \boldsymbol{\tau} = \{ \tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(M)} \}$  are the timestamps of the events occurring in a time interval  $[0, t_{end}]$ .



Self-exciting

- The past events always increase the chance of arrivals of new events

# Background - Meta Learning

• Meta Learning

 $\mathcal{T}_i \in \Gamma$ 

- Given a set of tasks  $\Gamma = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N\}$ – Each task contains a very small amount of data

#### **Experiment** – Synthetic Graphs

• Data generation: 50 Nodes, 6 Communities, S: Sparsity of the Graph,  $K_0$ : Number of Specified Communities

• Experiment: Community Assignment





# Experiment – Real Graphs

HARMLESS (Reptile)

Dataset	911-Calls	LinkedIn	MathOverflow	StackOverflow
MLE-Sep	$4.0030 \pm 0.3763$	$0.8419 \pm 0.0251$	$0.5043 \pm 0.0657$	$0.2862 \pm 0.0177$
MLE-Com	$4.5111 \pm 0.3192$	$0.8768 \pm 0.0028$	$1.7805 \pm 0.0345$	$1.5594 \pm 0.0134$
DMHP	$4.4812 \pm 0.3434$	$0.8348 \pm 0.0030$	$1.5394 \pm 0.0347$	N ackslash A
MTL	$4.4621 \pm 0.3173$	$0.9270 \pm 0.0027$	$1.7225 \pm 0.0336$	$1.4910 \pm 0.0089$
HARMLESS (MAML)	$4.5208 \pm 0.3256$	$1.4070 \pm 0.0105$	$1.8563 \pm 0.0345$	$1.3886 \pm 0.0082$
HARMLESS (FOMAML)	$4.6362 \pm 0.3241$	$1.0129 \pm 0.004$	$1.8344 \pm 0.0348$	$1.5988 \pm 0.0083$

 $0.9540 \pm 0.0082$ 

 $4.4929 \pm 0.3503$ 

- Model-Agnostic Meta Learning (MAML)
- Train a common model for all tasks,

 $\min_{\theta} \sum \mathcal{F}_{\mathcal{T}_i}(\theta - \eta \nabla_{\theta} \mathcal{F}_{\mathcal{T}_i}(\theta))$ 



where  $\mathcal{F}_{\mathcal{T}_i}$  is the loss function of task  $\mathcal{T}_i$ ,  $\theta$  is the parameter of the common model,  $\eta$  is the step size.

- Find the common model that is expected to produce maximally effective behavior on that task after performing update  $\theta - \eta \mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$ .
- Variants to alleviate the computational burden:
  - First Order MAML (FOMAML)
  - Reptile

Experiment – Ablation Study				
IVIETNOD	Log-Likelinood			
HARMLESS (MAML)	$1.4070 \pm 0.0105$			
HARMLESS (FOMAML)	$1.0129 \pm 0.0042$			
HARMLESS (Reptile)	$0.9540 \pm 0.0082$			
Remove inner heterogeneity $(K = 3)$	$0.9405 \pm 0.0032$			
Remove inner heterogeneity $(K = 5)$	$0.9392 \pm 0.0032$			
Remove grouping (MAML)	$0.9432 \pm 0.0031$			
Remove grouping (FOMAML)	$0.9376 \pm 0.0031$			
Remove grouping (Reptile)	$0.9455 \pm 0.0041$			
Remove graph (MAML)	$0.9507 \pm 0.0032$			
Remove graph (FOMAML)	$0.9446 \pm 0.0032$			
Remove graph (Reptile)	$0.9489 \pm 0.0072$			

### Reference

• HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*fg, 58 83?90.

 $1.6017 \pm 0.0097$ 

 $1.8663 \pm 0.0342$ 

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

• 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.