Deep Fourier Kernel for Self-Attentive Point Processes

Shixiang Zhu, Minghe Zhang, Ruyi Ding, Yao Xie*

Georgia Institute of Technology

Outline

□ Motivation

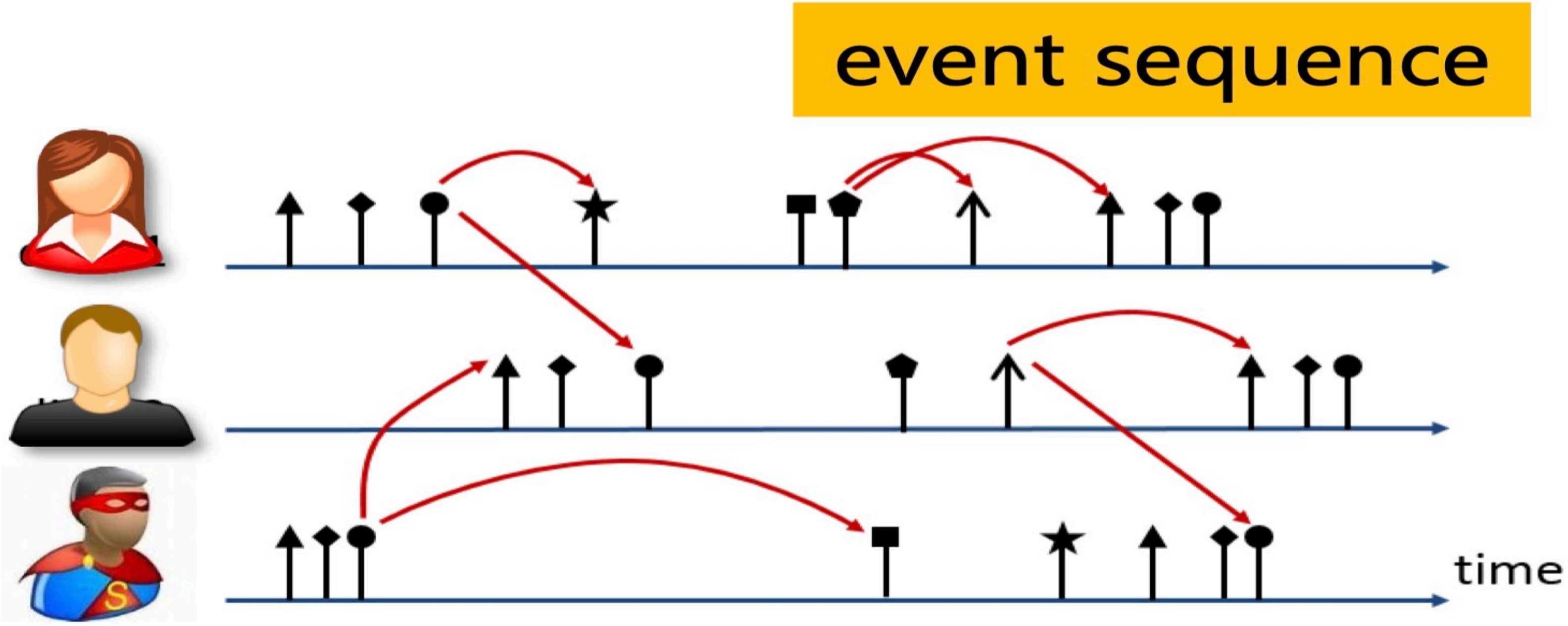
□ Solution & Theoretical Proof

□ Experiments

□ Conclusion

Motivations

- **Problem:**
  Modeling discrete events



Motivation

- **Problem:**

- More Expressive Models are needed:

With the increasing complexity and quantity of modern data, there has been much effort in developing neural network-based point processes, leveraging the rich representation power of neural networks, relying heavily on Recurrent Neural Networks.

- Limitation of Existing Neural Network-Based Models

RNN models are not enough capable of capturing long-range dependencies and RNNs "overemphasize" the recent events and fail to capture the events' influences in the distant past.

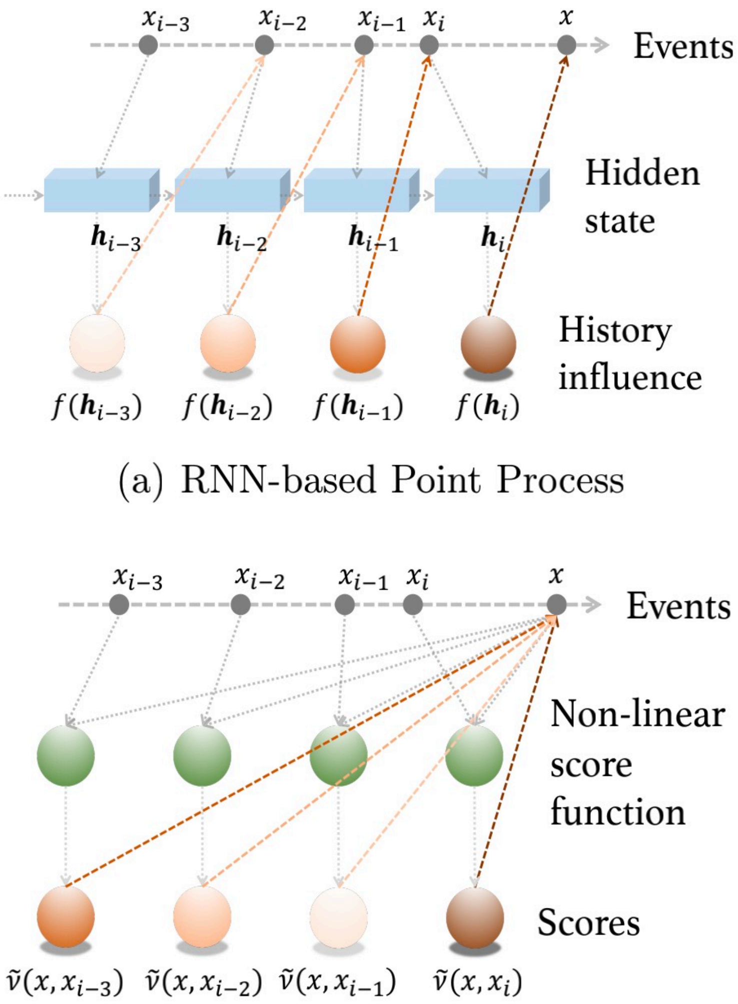
- To tackle the Effect of Non-linear and Long-range Dependence

The similarity between words can be characterized by dot-product score, while discrete events usually exhibit heterogeneous triggering effects regarding their spatio-temporal distances.

Outline

- □ Motivation
- □ Solution & Theoretical Proof
- □ Experiments
- □ Conclusion

Solution



(a) RNN-based Point Process

(b) Deep Attention Point Process

Figure 1: Comparison between RNN-based models and our DAPP.

The color depth of the red balls represent their "importance" in the model. The history influence in (a) are exponentially decaying over the time. The score is a non-linear function with respect to the distance between events and is non-homogeneous over the time.

Solution

- Self-attention in point processes
- Score function via deep Fourier kernel
- Fourier feature generator
- Online attention for streaming data
- Learning and simulation

Solution

Score function via deep Fourier kernel

Score function directly quantifies how likely one event is triggered by the other in a sequence.

This paper proposes a novel deep Fourier kernel as the score function in the attention mechanism.

The optimal spectrum (the distribution of power) is represented by a deep neural network.

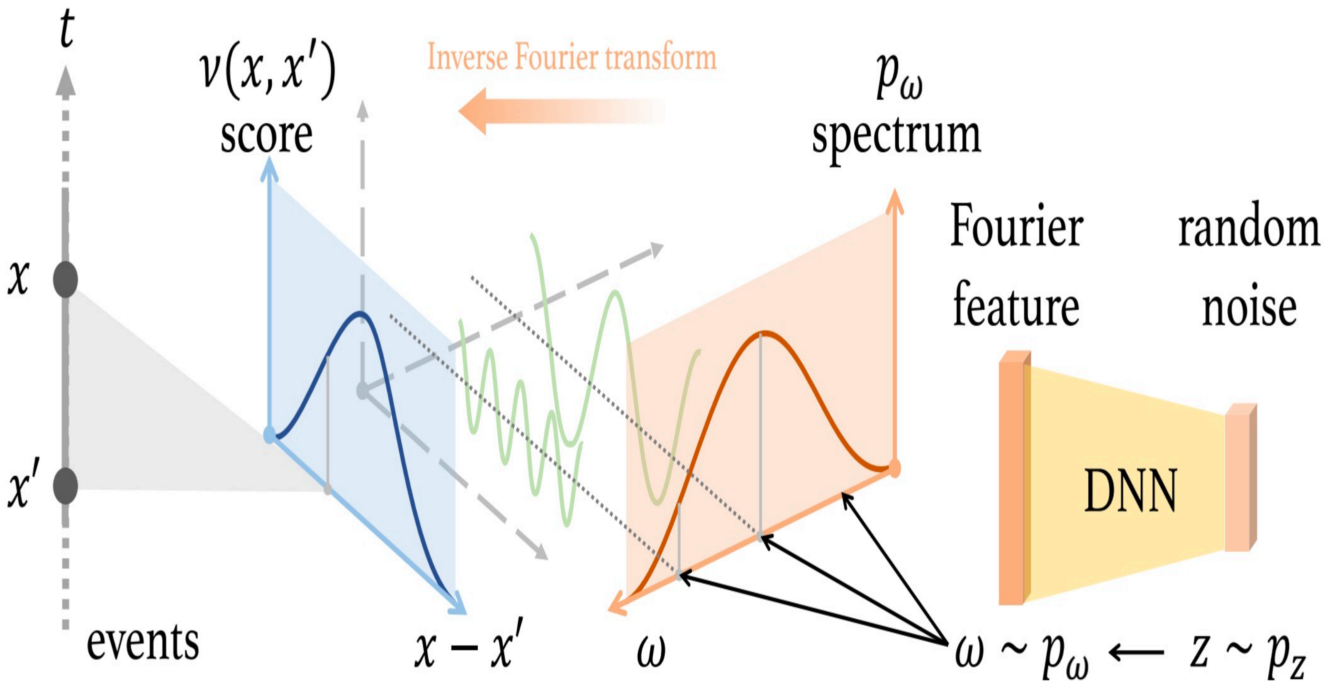


Figure 2: An illustration for the Fourier kernel score

Solution

Fourier feature generator

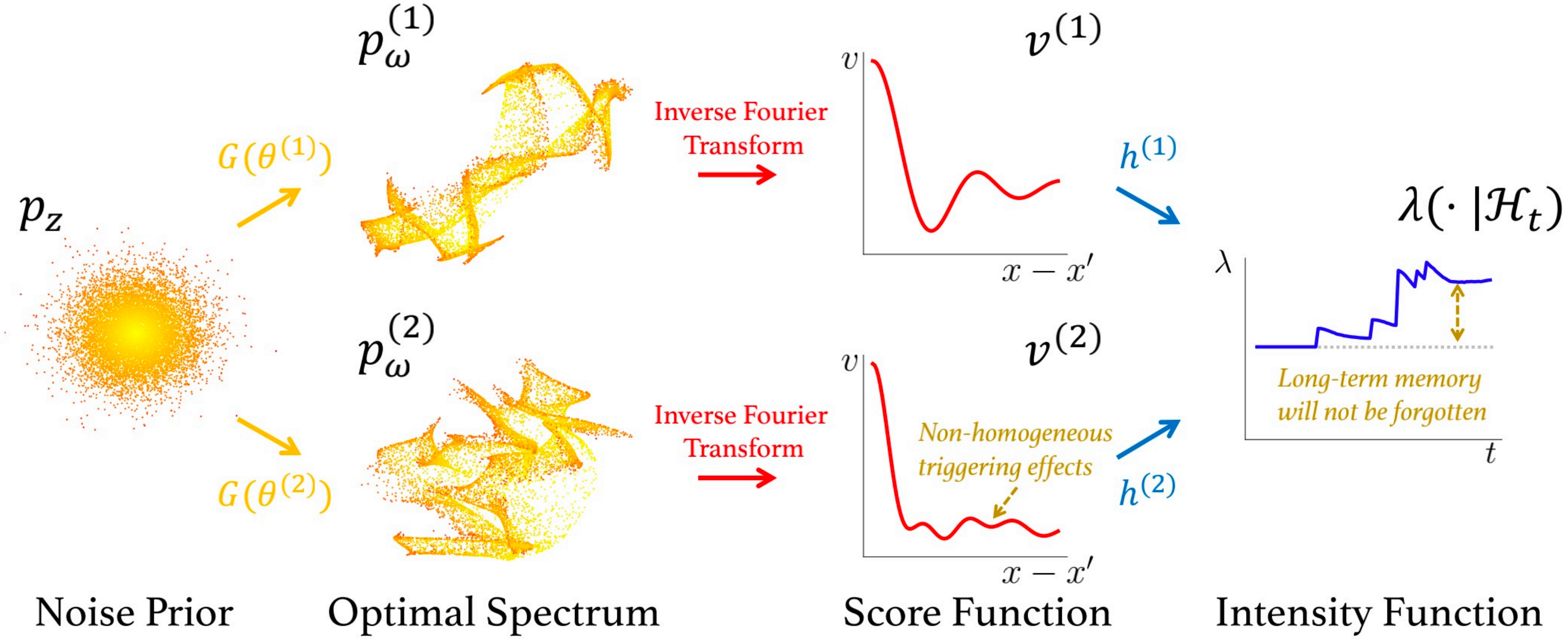


Figure 3: A real example of optimal spectrums, score function, and corresponding intensity function learned from DAPP.

Fourier kernel score is able to capture non-homogeneous triggering effects of events and long-term memory will also not be forgotten in this case.

Outline

- □ Motivation
- □ Solution & Theoretical Proof
- □ Experiments
- □ Conclusion

Experiments

- **Baseline methods:**
    - Recurrent Marked Temporal Point Process (RMTPP)
    - Neural Hawkes Process (NHP)
    - Self-Attentive Hawkes Process (SAHP)
    - Hawkes Process (HP)

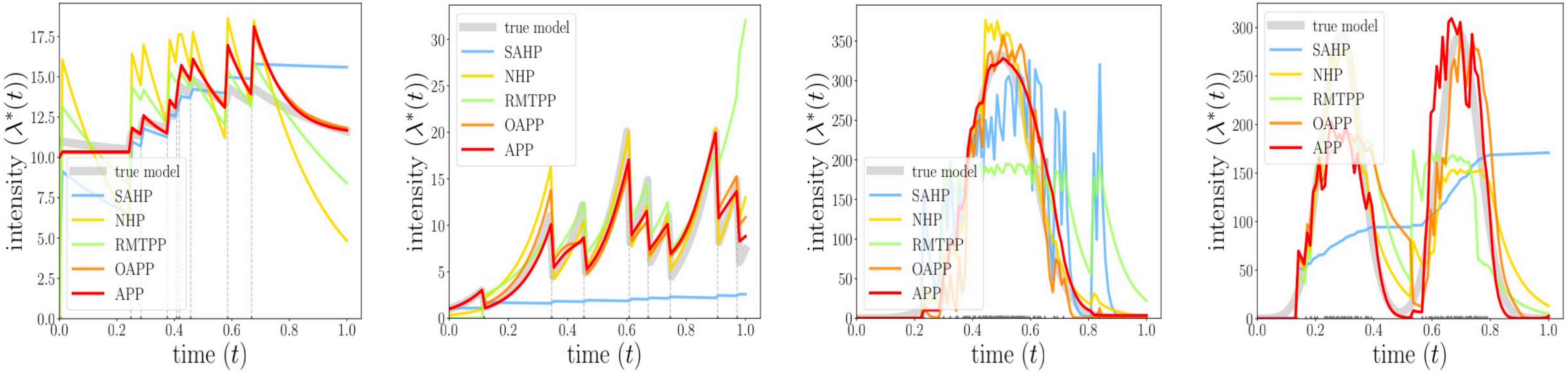
Experiments

- **Data**

Synthetic data sets: obtained by the following four generative processes:

(1) Hawkes process (2) self-correction point process

(3) nonhomogeneous Poisson 1 (4) non-homogeneous Poisson 2



(a) Hawkes (b) self-correction (c) non-homogeneous 1 (d) non-homogeneous 2

Figure 5: Conditional intensity function estimated from synthetic data sets. Triangles at the bottom of each panel represent events. The ground truth of conditional intensities is indicated by the grayline.

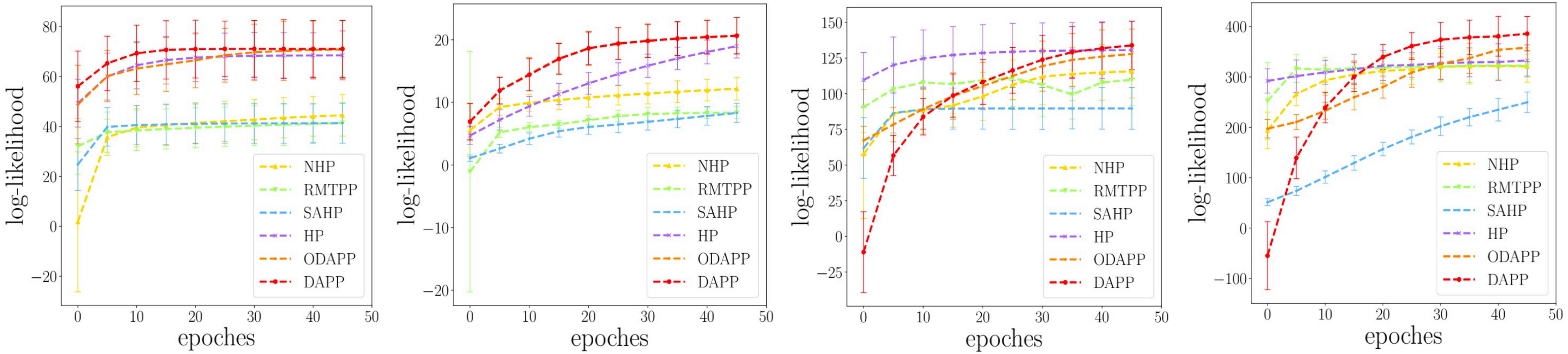
Experiments

- **Data**

Real data sets:

(1) Traffic Congestions (traffic) (2) Electrical Medical Records (MIMIC-III

(3) Financial Transactions (stock) (4) Memes (meme)



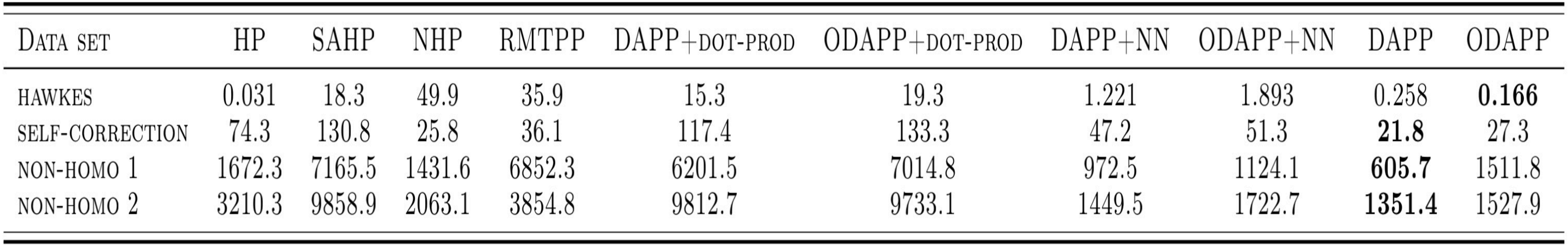
(a) stock (b) MIMIC-III (c) meme (d) traffic

Figure 6: The average log-likelihood of real data sets versus training epochs. For each real data set, we evaluate performance of the five methods according to the final log-likelihood averaged per event calculated for the test data.

Experiments

- **Synthetic Data Experiment Results**

Table 1: the mean square error of recovering the intensity.



Data set	HP	SAHP	NHP	RMTPP	DAPP+dot-prod	ODAPP+dot-prod	DAPP+NN	ODAPP+NN	DAPP	ODAPP
HAWKES	0.031	18.3	49.9	35.9	15.3	19.3	1.221	1.893	0.258	**0.166**
SELF-CORRECTION	74.3	130.8	25.8	36.1	117.4	133.3	47.2	51.3	**21.8**	27.3
NON-HOMO 1	1672.3	7165.5	1431.6	6852.3	6201.5	7014.8	972.5	1124.1	**605.7**	1511.8
NON-HOMO 2	3210.3	9858.9	2063.1	3854.8	9812.7	9733.1	1449.5	1722.7	**1351.4**	1527.9

The proposed methods achieve the minimal error in recovering intensities.

Experiments

- **Real Data Experiment Results**

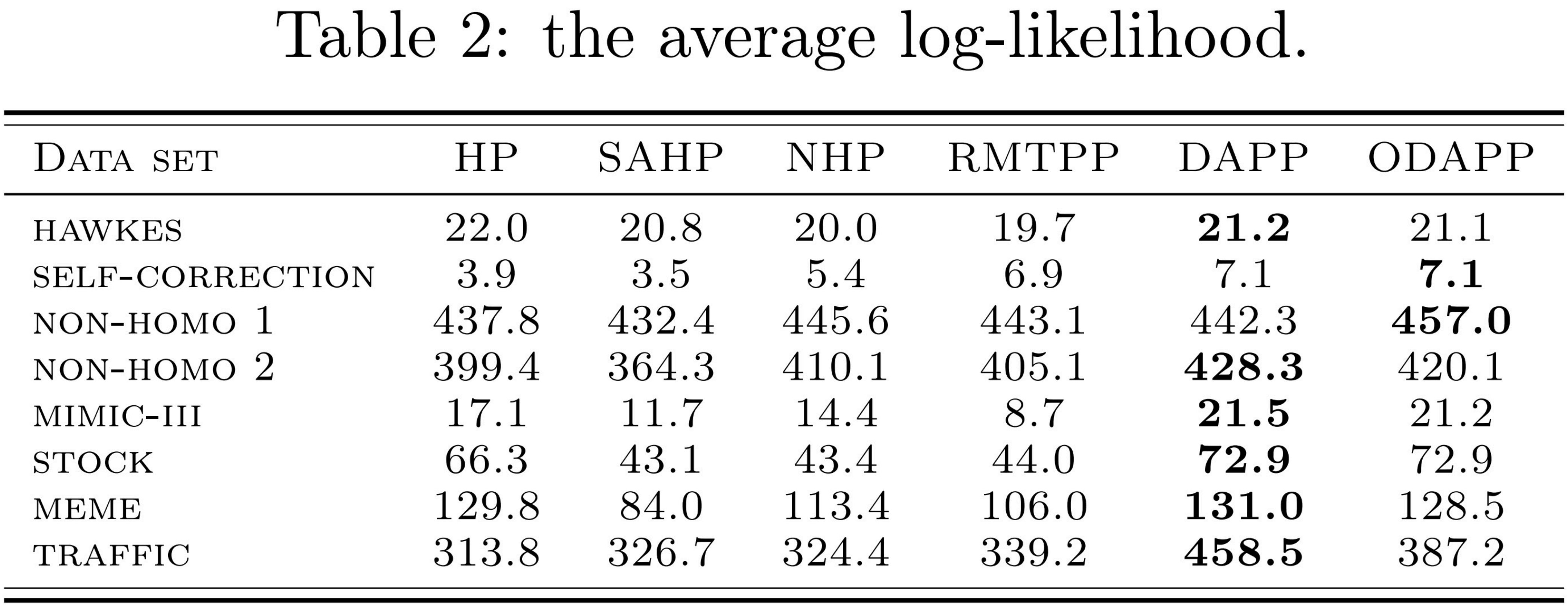


Table 2: the average log-likelihood.

Data set	HP	SAHP	NHP	RMTPP	DAPP	ODAPP
HAWKES	22.0	20.8	20.0	19.7	**21.2**	21.1
SELF-CORRECTION	3.9	3.5	5.4	6.9	7.1	**7.1**
NON-HOMO 1	437.8	432.4	445.6	443.1	442.3	**457.0**
NON-HOMO 2	399.4	364.3	410.1	405.1	**428.3**	420.1
MIMIC-III	17.1	11.7	14.4	8.7	**21.5**	21.2
STOCK	66.3	43.1	43.4	44.0	**72.9**	72.9
MEME	129.8	84.0	113.4	106.0	**131.0**	128.5
TRAFFIC	313.8	326.7	324.4	339.2	**458.5**	387.2

Outline

- □ Motivation
- □ Solution & Theoretical Proof
- □ Experiments
- □ Conclusion

Conclusion

Contributions

- Introducing a general probabilistic attention-based point process model for discrete event data;

- Introducing a novel similarity kernel based on Fourier kernel embedding and neural network represented spectrum (in contrast to the standard dot-product kernel).

Thank you!