1 Appendix A. Optimal Preserving Thresholds Evaluation

Deep layers analysis. Preserving threshold also represents an explicable approach to overcome over-fitting and over-smoothing problems for deep layers in GNN. To verify our inference, we design three varients of FinEvent with 4/8/32 layers and their corresponding model without preserving thresholds. We still adopt the latest-message strategy and set window size to 1, and report the performance of the models on M_9 to M_{14} in Fig. 1. It is observed that FinEvent with preserving threshold improve the detection effects for all the cases, and the improvement is greater as the layer grows. Different from [1], which drops edges in graph randomly, multi-agents in FinEvent drops edges based on certain rules and their learned experiment. To conclude, FinEvent provides a reasonable edge-drop instruction for the improvement of deep GNNs.

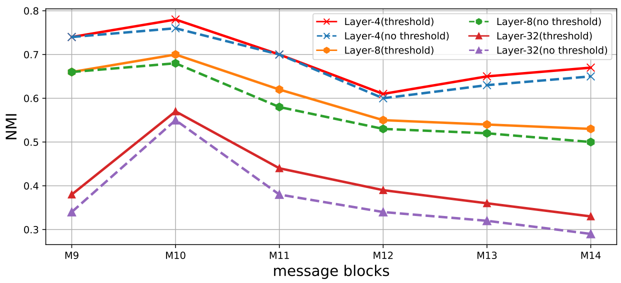

Fig. 1: Inference results for different depths of layers.

Interpretability analysis. In addition, while in training or detection stage, different agents would eventually choose different preserving threshold values. The various thresholds indicate different contributions made by agents, which can be abstracted as macroscopic attention of GNN towards different relations. As depicted in M_3, for instance, relation *M-E-M* gets higher threshold values than others probably for occupying higher importance in this block. We also display the preserving thresholds in a week as radar maps (shown in Fig. 2) for the convenience of observing the flow of relations' contributions. It demonstrates the change of block and event structures with time. The enclosed area can be approximately regarded as the overall contributions one relation makes to events of this week, and hence we can intuitively obtain the global relation importance. All of these give necessary interpretable explanation to GNN for event detection.

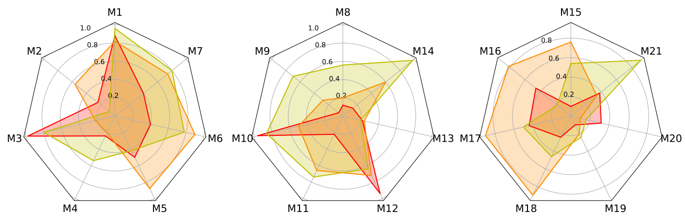

Fig. 2: Comparison on the performance of varients of FinEvent on English dataset in 3 weeks. The colors indicate different relations: *M-U-M* is colored in red, *M-E-M* is colored in orange and *M-L-M* is colored in green.

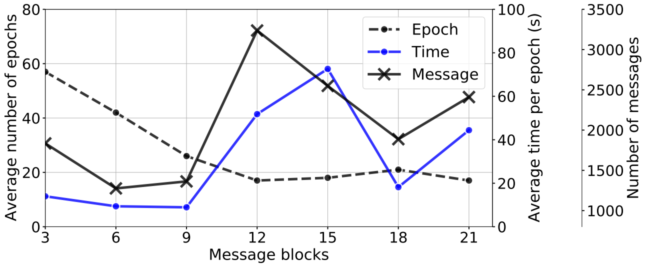

Fig. 3: Time consumption in the online maintenance stage.

2 Appendix B. Hyper-parameter Sensitivity

This subsection studies the effects of hyperparameters in the incremental social event detection experiments. We set the hidden embedding dimension to 128 for Twitter dataset to minimize the graph entropy as much as possible, and only change the output embedding dimension d and window size w. Fig. 4 compares the performance of FinEvent when adopting different output dimension as well as window size for each message blocks. The NMI, AMI and ARI results have average deviations in the range from 0.01-0.03. This suggests that the metrics of FinEvent change with d and w, but rather significantly. The output embedding dimension has little influence on the performance of FinEvent. For example, the block-wise average NMIs of output embedding dimension are 0.701, 0.719, 0.716, respectively. FinEvent reaches the best performance when d is set to 64. Adopting a smaller window size (2 or 3) in general gives a slightly better performance. For example, the block-wise average NMIs of different window sizes are 0.719, 0.729, 0.723, 0.714, respectively and average AMIs are 0.693, 0.702, 0.698, 0.687, respectively. When window size is set to 3, FinEvent achieves the best performance. A possible reason is that for Twitter dataset, $w = 3$ has best adaptability to the continuation of events. In a word, FinEvent is sensitive to the changes in hyperparameters.

3 Appendix C. Statistics of Social Streams

This section depicts the number of messages and the number of events composed in each block from English and French dataset, respectively. The details are shown in Table 1 and Table 2. In addition, the time consumption of the social stream is given in Figure 3

References

[1] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," in *Proceedings of the ICLR*, 2020.

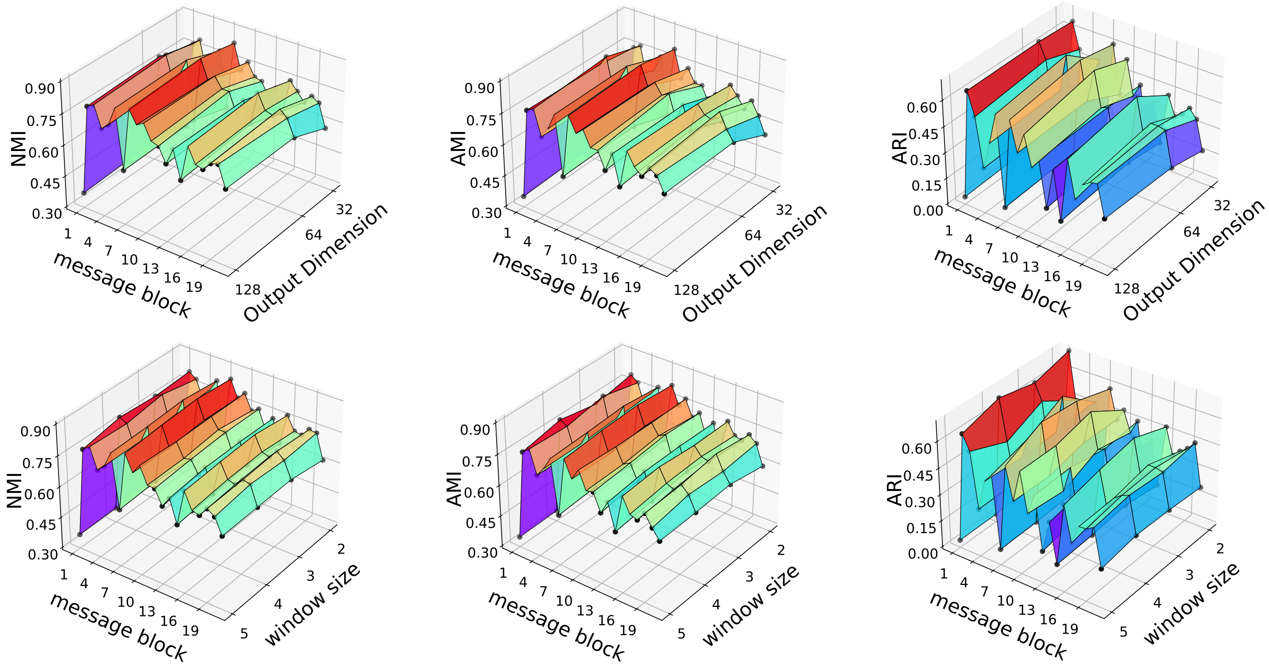

Fig. 4: FinEvent with different hyperparameters.

TABLE 1: The statistics of the social stream from English Twitter Dataset.

Blocks	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
# of messages # of events	$20,254$ 155	$8,722$ 41	$1,491$ 30	$1,835$ 33	$2,010$ 38	$1,834$ 30	$1,276$ 44	$5,278$ 57	$1,560$ 53	$1,363$ 38	$1,096$ 33
Blocks	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}
# of messages # of events	$1,232$ 30	$3,237$ 42	$1,972$ 40	$2,956$ 43	$2,549$ 42	910 27	$2,676$ 35	$1,887$ 32	$1,399$ 28	893 34	$2,410$ 32

TABLE 2: The statistics of the social stream from French Twitter Dataset.

Blocks	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7
# of messages # of events	$14,328$ 79	$5,356$ 22	$3,186$ 19	$2,644$ 15	$3,179$ 19	$2,662$ 27	$4,200$ 26	$3,454$ 23
Blocks	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}
# of messages # of events	$2,257$ 25	$3,669$ 31	$2,385$ 32	$2,802$ 31	$2,927$ 29	$4,884$ 28	$3,065$ 26	$2,411$ 25