Supplementary Material
Distill2Vec: Dynamic Graph Representation Learning with Knowledge Distillation

Stefanos Antaris
KTH Royal Institute of Technology
Hive Streaming AB
Sweden
antaris@kth.se

Dimitrios Rafailidis
Maastricht University
Netherlands
dimitrios.rafailidis@maastrichtuniversity.nl

Parameter Settings We tuned the hyper-parameters of each examined model following a cross-validation strategy. In Table I, we present the concluded values of the hyper-parameters of each model. In our experiments, we optimized the weight parameters of each model, employing the mini-batch gradient descent with the Adam optimizer [1]. We initialized the learning rate to $1e-03$ and train each model for 200 epochs. All experiments were performed on an Intel(R) Xeon(R) Bronze 3106 CPU 1.70GHz machine and GPU accelerated with the GEFORCE RTX 2080 Ti graph card.

TABLE I
CONCLUDED HYPER-PARAMETERS OF EACH EXAMINED MODEL

	Yelp			**ML-10M**		
Model	**Embedding** d	**Window** l	**Heads** h/c	**Embedding** d	**Window** l	**Heads** h/g
Distill2Vec-$\mathcal{T}$	256	5	16	512	5	8/8
Distill2Vec-$\mathcal{S}$	64	2	2	128	2	4/4
Distill2Vec-$\mathcal{L}$	64	2	2	128	2	4/4
DynVGAE	256	3	N/A	128	2	N/A
DynamicTriad	256	3	N/A	512	2	N/A
TDGNN	512	2	N/A	256	3	N/A
DyREP	128	3	N/A	256	3	N/A
DMTKG-$\mathcal{T}$	512	N/A	N/A	256	N/A	N/A
DMTKG-$\mathcal{S}$	256	N/A	N/A	64	N/A	N/A

REFERENCES

[1] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.