香港浸會大學 HONG KONG BAPTIST UNIVERSITY | 研發辦公室 Research Office | 北师香港浸会大学 BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY | 研究拓展及知识转移处 RESEARCH DEVELOPMENT AND KNOWLEDGE TRANSFER OFFICE

GUANGDONG AND HONG KONG UNIVERSITIES "1+1+1" Joint Research Collaboration Scheme
粵港高校「1+1+1」聯合資助計劃

Optimizing and Accelerating Graph Neural Networks for Large-Scale Irregular IoT Sensor Data on Chinese NPU Devices

HKBU Amelie Chi Zhou | HKBU Hongning Dai | BNBU Weipeng Zhuo | BNBU Jing Zhao | BNBU Zhiyuan Li

Summary

Efficient and robust graph intelligence for large-scale, irregular, and system-constrained data is important. To achieve this, **SNI-GNN**[1] improves the efficiency of distributed full-graph GNN training by reducing communication overhead through SmartNIC-assisted in-network embedding prediction, while **TAMI**[2] improves temporal graph link prediction by explicitly modeling heterogeneous interaction intervals and preserving pair-specific history. Beyond these two core studies, **MTM**[3] provides a related perspective on learning from irregular temporal observations through multi-scale temporal modeling, and **CFDGraph**[4] highlights the importance of privacy-preserving and communication-efficient graph processing in collaborative environments. Taken together, these works show that advancing graph intelligence in real-world settings requires jointly addressing **system efficiency, temporal irregularity, and deployment constraints**, rather than optimizing model accuracy alone.

SNI-GNN: SmartNIC-Assisted Full-Graph GNN Training with In-Network Embedding Prediction (ICDE 26)

Graph Neural Networks (GNNs) are widely used in recommendation, social networks, fraud detection, and scientific applications. As graphs continue to grow, efficient distributed training becomes a key systems challenge. **Full-graph training** is attractive because it preserves complete neighborhood aggregation and can deliver strong accuracy, but in distributed settings it suffers from **heavy communication overhead** caused by frequent synchronization of **remote embeddings**.

To address this bottleneck, **SNI-GNN** offloads **stale-embedding compensation** to the SmartNIC/DPU instead of relying only on periodic synchronization. Its key insight is that remote node embeddings usually evolve smoothly across epochs, so their near-future values can be estimated from recent history using a **lightweight predictor**. As shown in **Figure 1**, SNI-GNN introduces a **SmartNIC-assisted training pipeline** in which the CPU handles graph partitioning and sampling, the GPU performs **full-graph training** with a local cache, and the SmartNIC/DPU maintains embedding history and generates **predicted remote embeddings** when synchronization is skipped.

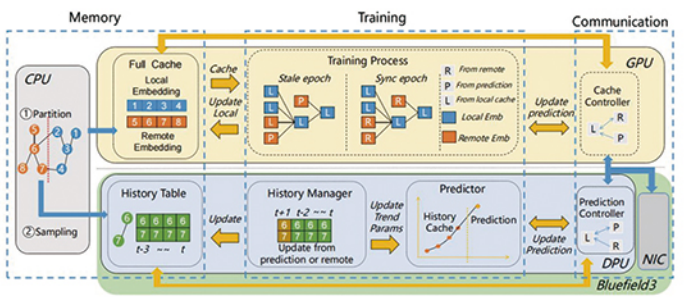


Figure 1. System architecture of SNI-GNN for SmartNIC-assisted distributed full-graph GNN training.

In the system pipeline, the CPU performs graph partitioning and subgraph sampling, while the GPU trains with a **full cache** that stores local embeddings and cached remote embeddings. On the network side, the SmartNIC/DPU maintains **historical records** for remote nodes, updates them through a **history manager**, and predicts near-future remote embeddings for later use by the GPU.

Training alternates between two modes. In a **sync epoch**, workers exchange fresh remote embeddings, which are used to update both the GPU cache and the DPU history. In a **stale epoch**, synchronization is skipped; instead, the GPU queries the SmartNIC/DPU, which predicts remote embeddings from recent history and returns them for **message passing and aggregation**. In this way, SNI-GNN replaces naive stale reuse with **in-network prediction-based compensation**, reducing synchronization overhead while preserving **aggregation quality**.

As shown in **Figure 2**, SNI-GNN consistently reduces **communication overhead** across datasets, leading to substantial **end-to-end speedup** while maintaining **negligible accuracy loss**.

Experiments show that SNI-GNN reduces communication by **21%–45%,** achieves **1.3×–3.6×** end-to-end speedup over BNS-GCN, and delivers up to **1.29×** additional speedup over SANCUS.

Figure 2. SNI-GNN Runtime Breakdown

TAMI: Taming Heterogeneity in Temporal Interactions for Temporal Graph Link Prediction (NeurIPS 25)

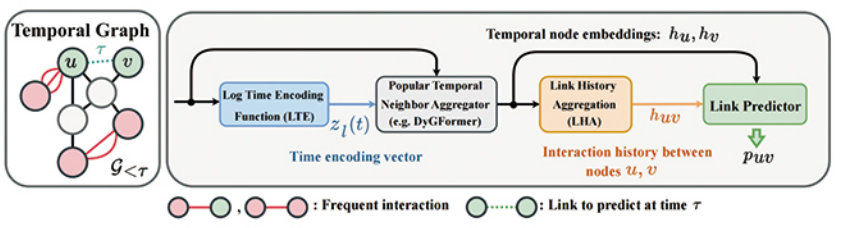


Figure 3. TAMI improves temporal graph link prediction through Log Time Encoding (LTE) and Link History Aggregation (LHA).

Temporal graph link prediction is challenging because **temporal interactions** are highly heterogeneous. In real-world **continuous-time temporal graphs**, a small number of node pairs generate most **interaction events**, while many others interact only occasionally. This leads to highly skewed interaction intervals and makes it difficult for existing models to encode temporal information effectively and preserve useful **pair-specific histories**.

To address this issue, we propose **TAMI**, a general framework for **temporal graph link prediction**. As shown in **Figure 3**, TAMI introduces two lightweight yet effective components: **Log Time Encoding (LTE)** and **Link History Aggregation (LHA)**. LTE applies a logarithmic transformation to temporal differences before time encoding, producing a more balanced temporal input space, while LHA explicitly aggregates recent historical interactions between the target node pair so that infrequent but informative interactions are less likely to be forgotten.

TAMI is designed as a **modular framework** and can be integrated into different **temporal graph neural networks**. This allows the method to improve a broad range of existing models rather than being tied to a single architecture.

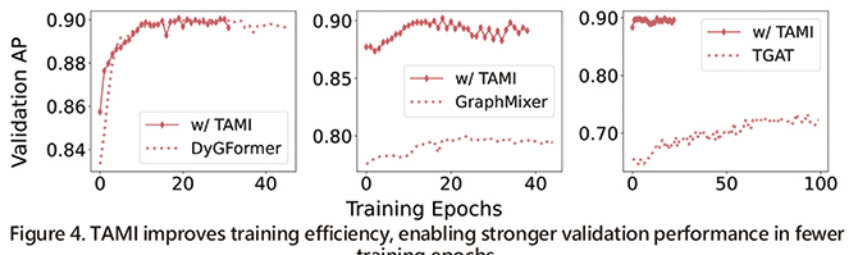
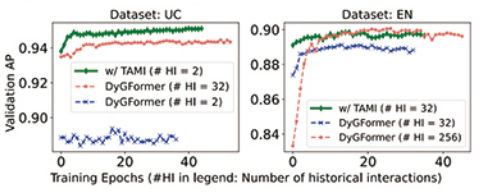


Figure 4. TAMI improves training efficiency, enabling stronger validation performance in fewer training epochs.

In addition to improving **predictive performance**, TAMI also improves **optimization behavior**. As illustrated in **Figure 4**, models equipped with TAMI reach stronger validation performance in fewer training epochs, indicating that the framework makes temporal patterns easier to learn and **accelerates convergence**.

TAMI also improves efficiency from a **systems perspective**. As shown in Figure 5, when integrated into existi -ng methods such as **DyGFormer**, TAMI can achieve equal or better accuracy while using fewer **history--ical interactions**. This reduces **GPU memory** usage and lowers overall training cost, making temporal graph learning more efficient in both comp--utation and **resource consumption**.



Method \ Metric	GPU Memory Usage (MiB) UC	GPU Memory Usage (MiB) EN	Training Time Per Epoch (Second) UC	Training Time Per Epoch (Second) EN
DyGFormer (32, 256)	2429	3303	37	81
DyGFormer (2, 32) w/ TAMI	1359	1545	22	57
Imp. (%)	44.05%	53.22%	40.54%	29.63%

Figure 5. TAMI reduces memory usage and training cost by using fewer historical interactions while maintaining or improving performance.

TAMI improves temporal graph link prediction by explicitly modeling **heterogeneity in temporal interactions** through better **time encoding** and **pair-specific history aggregation.**

[1] G. Yu, S. Chen, Z. Tang, X. Chu, and A. C. Zhou, "SNI-GNN: SmartNIC-Assisted Full-Graph GNN Training with In-Network Embedding Prediction," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2026.

[2] Z. Yu, J. Wu, Z. Wu, S. Zhong, W. Su, C. H. Lee, and W. Zhuo, "TAMI: Taming Heterogeneity in Temporal Interactions for Temporal Graph Link Prediction," in *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[3] S. Zhong, W. Zhuo, S. Song, G. Li, Z. Yu, and S. H. G. Chan, "MTM: A Multi-Scale Token Mixing Transformer for Irregular Multivariate Time Series Classification," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

[4] Q. Wu, A. C. Zhou, T. Allard, S. Ibrahim, Y. Feng, L. Li, and A. Abbadi, "CFDGraph: Privacy-Preserving Graph Processing for Large-Scale Collaborative Fraud Detection," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2026.