# Heterogeneous Graph Transformer with Poly-Tokenization

**Zhiyuan Lu, Yuan Fang, Cheng Yang, Chuan Shi**

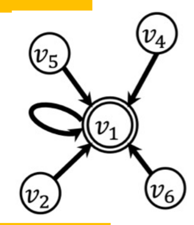
*In the Proceedings of the 33rd International Joint Conference on Artificial Intelligence.*

## Background

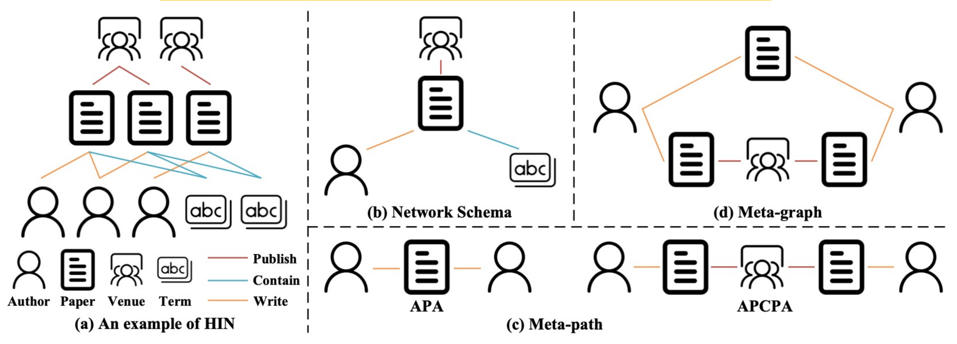
### Graph Neural Networks

$$h_v^l = \mathcal{M}(h_v^{l-1}, \{h_i^{l-1} : i \in \mathcal{N}_v\}; \theta^l)$$

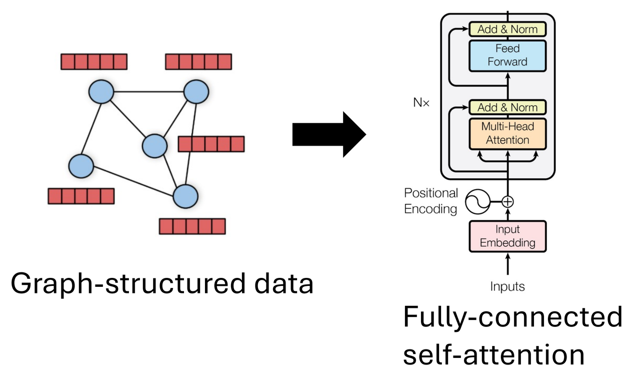
Message passing function




### Heterogeneous Graph




### Graph Transformer



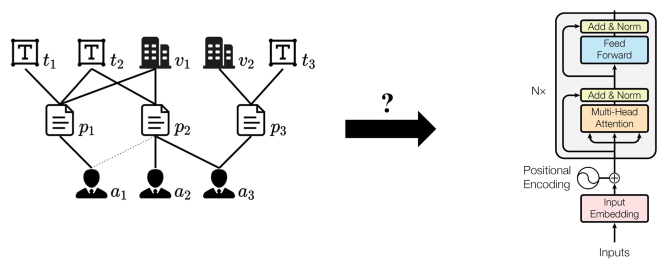

- Stronger expressive power
- Alleviate over-smoothing

## Motivation

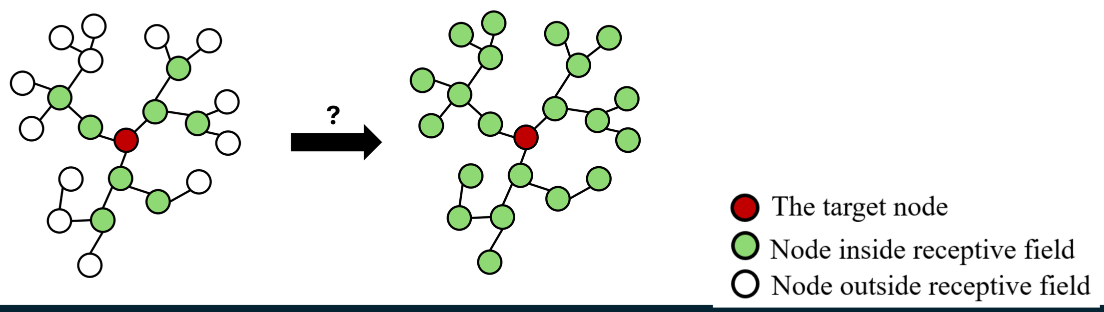
**Problem**: Design transformer models for heterogeneous graph representation learning.

**Challenges:**

**C1:** *How do we integrate the complex semantics on a heterogeneous graph into transformers?*



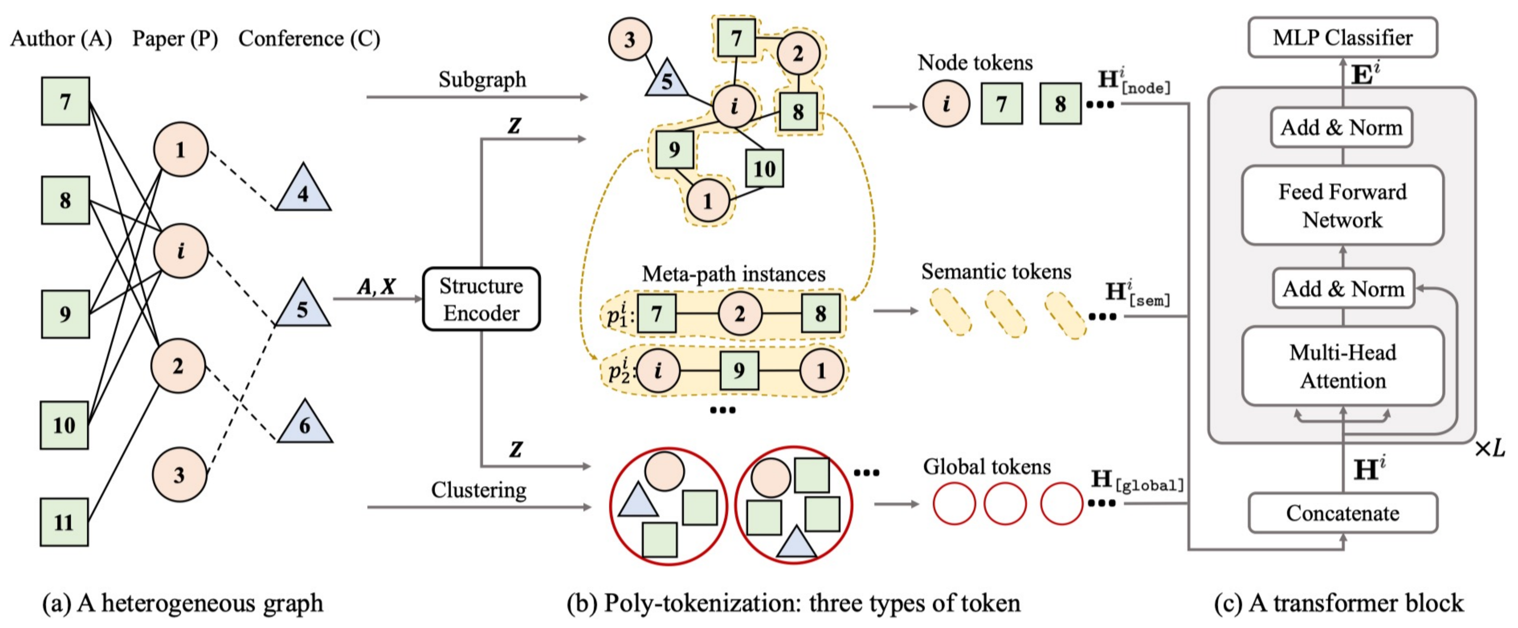
**C2:** *How do we expand the receptive fields for graph transformers to capture long-range interactions on a heterogeneous graph?*




## The Proposed Model: PHGT

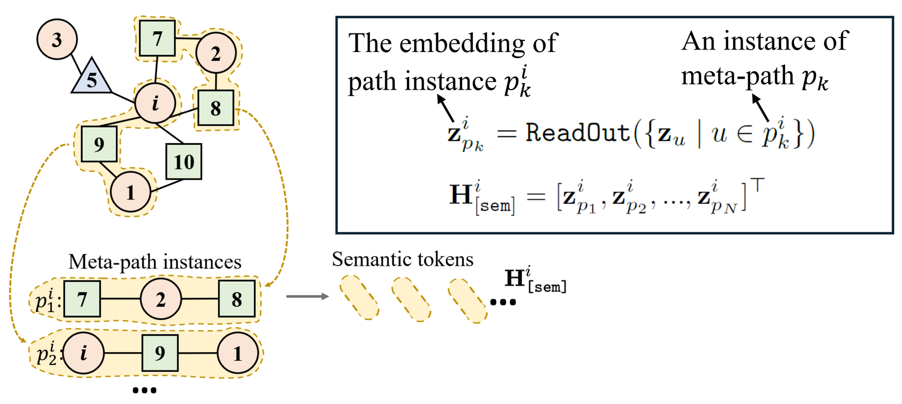
### Overall Framework

The proposed poly-tokenization mechanism




(a) A heterogeneous graph (b) Poly-tokenization: three types of token (c) A transformer block

### Semantic Token



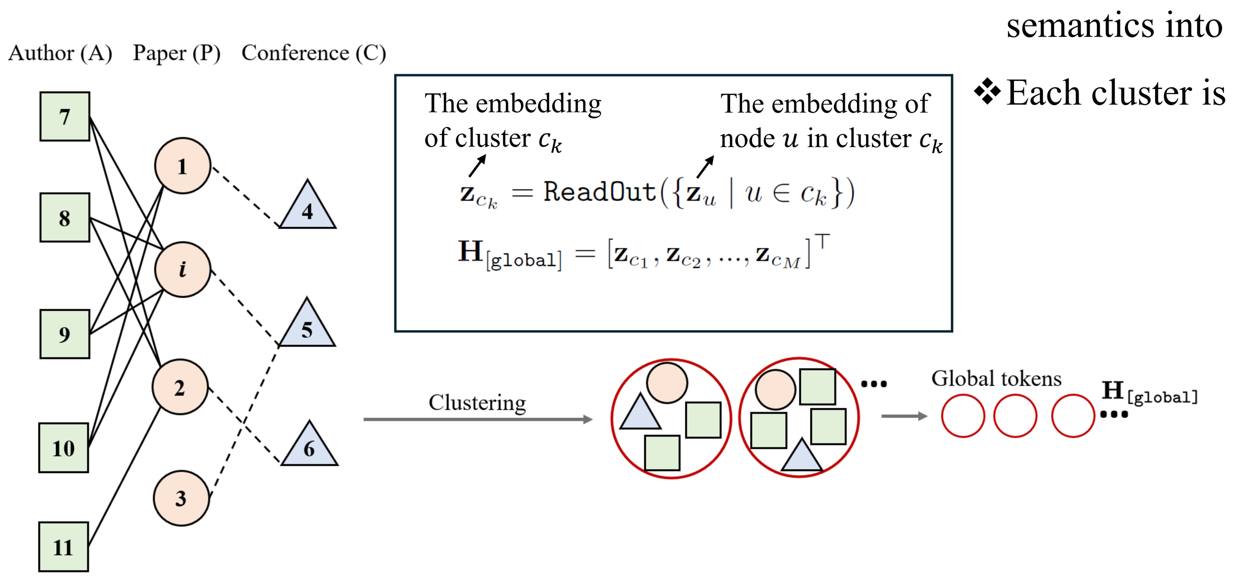

Solve Challenge 1:

- ❖Sample meta-path instances according to a pre-defined meta-path set.
- ❖Each meta-path instance is converted to a semantic token

### Global Token

Solve Challenge 2:

- ❖Summarize nodes with similar structure and semantics into a cluster.
- ❖Each cluster is converted to a global token.




## Experiments

### Experimental Setup

| Dataset | Nodes | # Node types | #Edges | # Edge types | Target | #Classes |
|---|---|---|---|---|---|---|
| DBLP | 26,128 | 4 | 239,566 | 6 | author | 4 |
| IMDB | 21,420 | 4 | 86,642 | 6 | movie | 5 |
| ACM | 10,942 | 4 | 547,872 | 8 | paper | 3 |
| Freebase | 43,854 | 4 | 151,034 | 6 | movie | 3 |

**Heterogeneous GNNs**
- ❖ RGCN [1]
- ❖ MAGNN [5]
- ❖ HAN [2]
- ❖ HGT [6]
- ❖ GTN [3]
- ❖ Simple-HGN [7]
- ❖ HetGNN [4]

**Homogeneous Graph Transformers**
- ❖ ANS-GT [8]
- ❖ NodeFormer [9]

**Heterogeneous Graph Transformers**
- ❖ HINormer [10]
- ❖ PHGT

### Node Classification

| Methods | DBLP | | IMDB | | ACM | | Freebase | |
|---|---|---|---|---|---|---|---|---|
| | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 |
| RGCN | 92.07±0.50 | 91.52±0.50 | 62.05±0.15 | 58.85±0.26 | 91.41±0.75 | 91.55±0.74 | 60.82±1.23 | 59.08±1.44 |
| HAN | 92.05±0.62 | 91.67±0.49 | 64.63±0.58 | 57.74±0.96 | 90.79±0.43 | 90.89±0.43 | 61.42±3.56 | 57.05±2.06 |
| GTN | 93.97±0.54 | 93.52±0.55 | 65.14±0.45 | 60.47±0.98 | 91.20±0.71 | 91.31±0.70 | - | - |
| HetGNN | 92.33±0.41 | 91.76±0.43 | 51.16±0.65 | 48.25±0.67 | 86.05±0.25 | 85.91±0.25 | 62.99±2.31 | 58.44±1.99 |
| MAGNN | 93.76±0.45 | 93.28±0.51 | 64.67±1.67 | 56.49±3.20 | 90.77±0.65 | 90.88±0.64 | 64.43±0.73 | 58.18±3.87 |
| HGT | 93.49±0.25 | 93.01±0.23 | 67.20±0.57 | 63.00±1.19 | 91.00±0.76 | 91.12±0.76 | 66.43±1.88 | 60.03±2.21 |
| Simple-HGN | 94.46±0.22 | 94.01±0.24 | 67.36±0.57 | 63.53±1.36 | 93.35±0.45 | 93.42±0.44 | 67.49±0.97 | 62.49±1.69 |
| ANS-GT | 93.15±0.51 | 92.75±0.43 | 66.65±0.35 | 62.52±0.61 | 92.55±0.54 | 93.67±0.62 | 67.33±0.61 | 61.24±0.57 |
| NodeFormer | 93.68±0.42 | 93.05±0.38 | 65.86±0.42 | 62.15±0.77 | 91.89±0.31 | 92.72±0.84 | 67.01±0.52 | 60.83±1.41 |
| HINormer | 94.94±0.21 | 94.57±0.23 | 67.83±0.34 | 64.65±0.53 | 93.15±0.36 | 93.28±0.43 | 67.78±0.39 | **62.76±1.10** |
| PHGT | **95.33±0.18** | **94.96±0.17** | **68.81±0.08** | **65.91±0.30** | **93.72±0.40** | **93.79±0.39** | **68.74±1.42** | 61.73±1.86 |

- ❖ PHGT demonstrates superior performance in most scenarios, outperforming other baselines.
- ❖ PHGT consistently outperforms homogeneous graph transformers (ANS-GT and NodeFormer).
- ❖ PHGT surpasses all the message passing-based HGNN baselines in almost all cases.

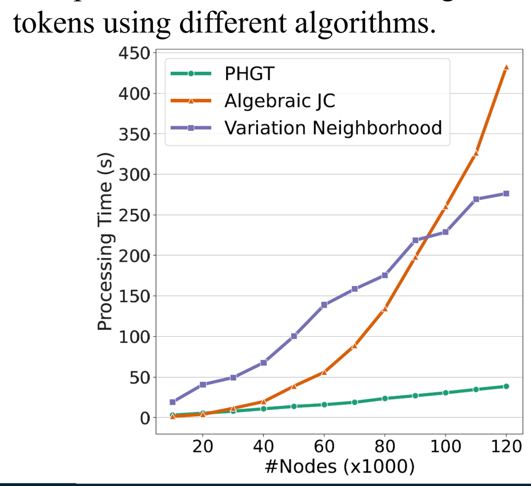
### Ablation Study

| | DBLP | IMDB | ACM | Freebase |
|---|---|---|---|---|
| w/o both | 94.80 | 68.35 | 93.34 | 67.58 |
| w/o semantic token | 94.94 | 68.58 | 93.41 | 67.73 |
| w/o global token | 94.91 | 68.54 | 93.55 | 68.06 |
| PHGT | **95.33** | **68.81** | **93.72** | **68.89** |

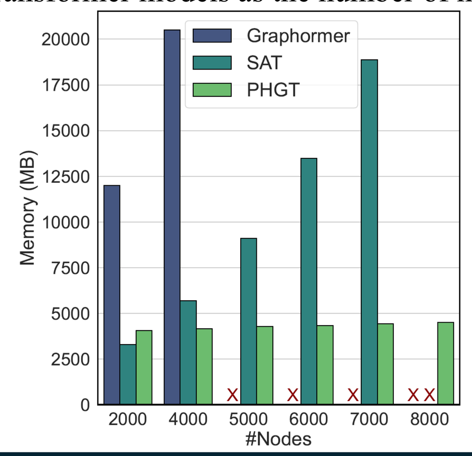
- ❖ w/o semantic token: The semantic token is removed to gauge its impact on performance;
- ❖ w/o global token: the global token is removed to assess its contribution;
- ❖ w/o both: In this variant, both the semantic tokens and the global tokens are removed, retaining only the node tokens.

### Efficiency Studies

Comparison of time overhead for generating global tokens using different algorithms.




Comparison of memory usage among different Transformer models as the number of nodes increases.




## Conclusions

➢ PHGT addressed the two limitations of existing graph transformer models:

(1) the inability to capture heterogeneous semantics;

(2) the incapacity to model intricate long-range dependencies.

➢ Through comprehensive experiments on four benchmark datasets, we demonstrate the efficacy of our PHGT approach.

## References


[1] Schlichtkrull et al. Modeling relational data with graph convolutional networks. ISWC 2018.
[2] Wang et al. Heterogeneous graph attention network. WWW 2019.
[3] Yun et al. Graph transformer networks. NeurIPS 2019.
[4] Zhang et al. Heterogeneous graph neural network. SIGKDD 2019.
[5] Fu et al. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. WWW 2020.
[6] Hu et al. Heterogeneous graph transformer. WWW2020.
[7] Lv et al. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. SIGKDD 2021.
[8] Zhang et al. Hierarchical graph transformer with adaptive node sampling. NeurIPS 2021.
[9] Wu et al. Nodeformer: A scalable graph structure learning transformer for node classification. NeurIPS 2022.
[10] Mao et al. Hinormer: Representation learning on heterogeneous information networks with graph transformer. WWW 2023.


OpenHGNN

shichuan.org