# Heterogeneous Graph Transformer

WWW20


Ziniu Hu*
University of California, Los Angeles
bull@cs.ucla.edu

Yuxiao Dong
Microsoft Research, Redmond
yuxdong@microsoft.com

Kuansan Wang
Microsoft Research, Redmond
kuansanw@microsoft.com

Yizhou Sun
University of California, Los Angeles
yzsun@cs.ucla.edu

# Author

- Second-year CS Ph.D student, advised by Prof. Yizhou Sun

**Associate Professor**

Department of Computer Science
University of California, Los Angeles

**Office:** BH 3531F
**Email:** yzsun at cs dot ucla dot edu

- bachelor degree in Peking University, advised by Prof. Xuanzhe Liu.

Ziniu Hu

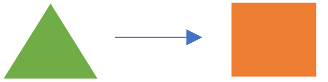
CS Ph.D. Student

University of California
Los Angeles

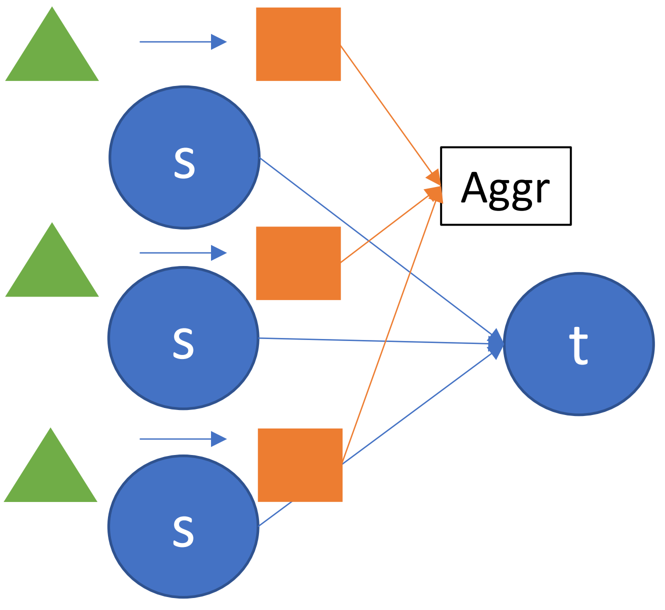
WSDM 2018, WWW 2019, Best Paper Award, ICLR 2019 Workshop, ACL 2019, WWW 2020

# Background

- General GNN Framework



$$H^l[t] \leftarrow \underset{\forall s \in N(t), \forall e \in E(s,t)}{\textbf{Aggregate}} \left( \textbf{Extract}\left( H^{l-1}[s]; H^{l-1}[t], e \right) \right)$$



# Background

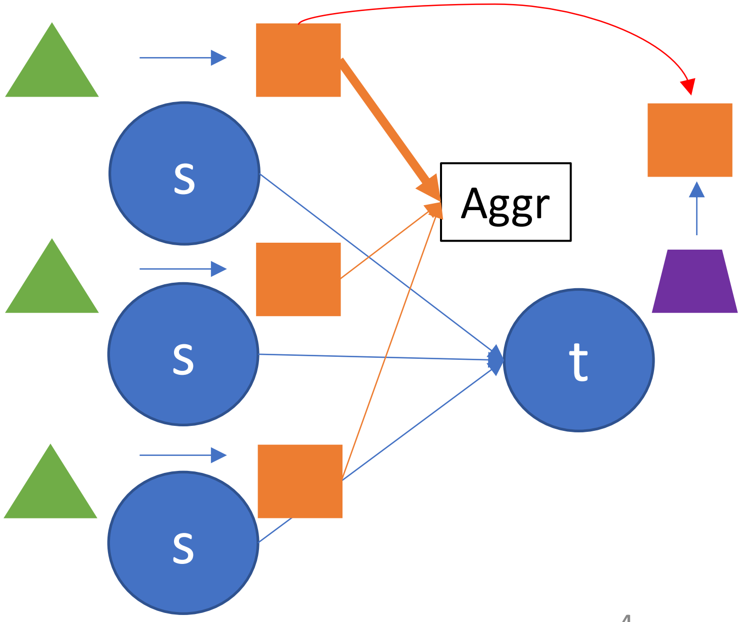
- Graph Attention Network

$$H^l[t] \leftarrow \underset{\forall s \in N(t), \forall e \in E(s,t)}{\textbf{Aggregate}} \left( \textbf{Attention}(s,t) \cdot \textbf{Message}(s) \right)$$

$$\textbf{Attention}_{GAT}(s,t) = \underset{\forall s \in N(t)}{\text{Softmax}} \left( \vec{a} \left( WH^{l-1}[t] \parallel WH^{l-1}[s] \right) \right)$$

$$\textbf{Message}_{GAT}(s) = WH^{l-1}[s]$$

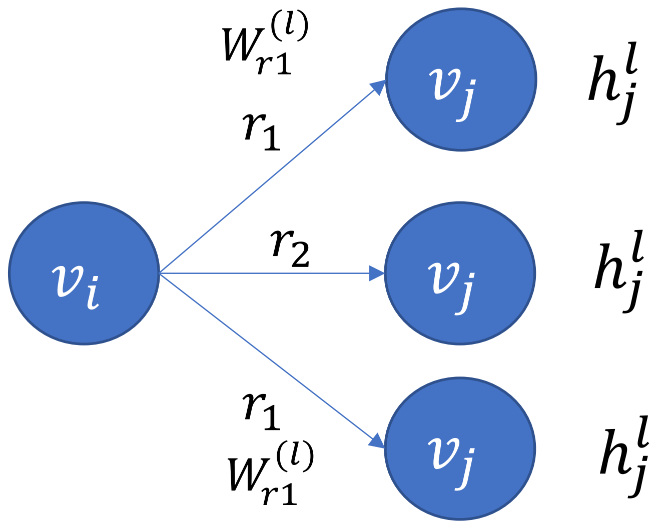
$$\textbf{Aggregate}_{GAT}(\cdot) = \sigma \left( \text{Mean}(\cdot) \right)$$



# Background

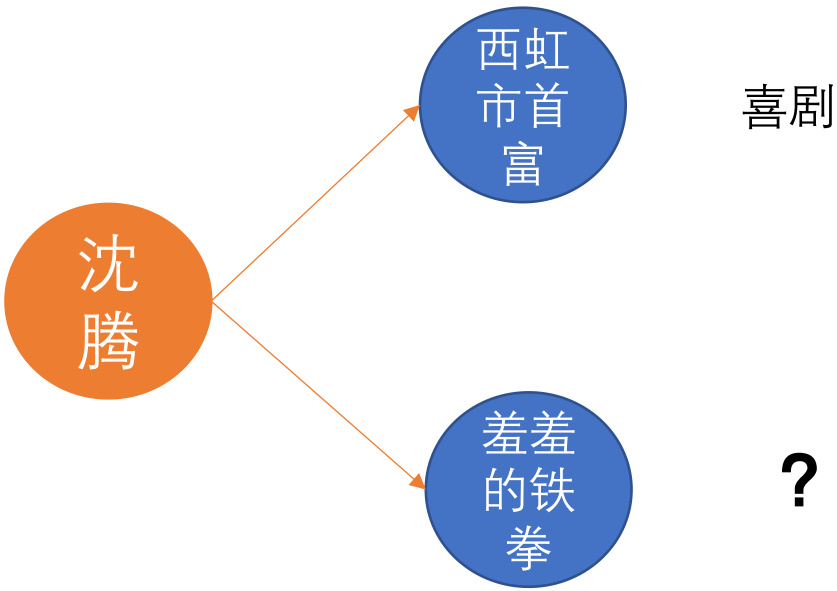
- Relational graph convolutional networks (R-GCN)

$$h_i^{(l+1)} = \mathrm{ReLU}\left(\sum_{r\in R_D}\sum_{v_j\in\mathcal{N}_r(v_i)}\frac{1}{|\mathcal{N}_i^r|}W_r^{(l)}h_j^{(l)}\right)$$

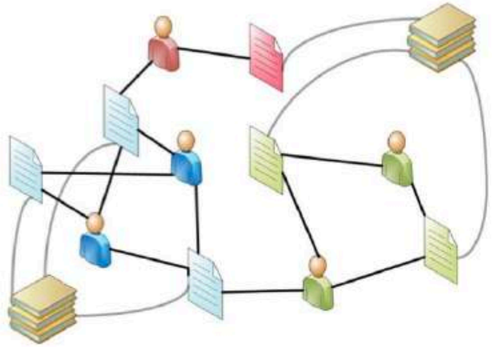


# Background

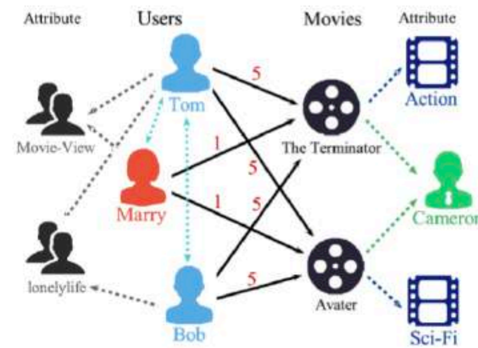
- Node classification



# Heterogeneous Information Networks (HIN)

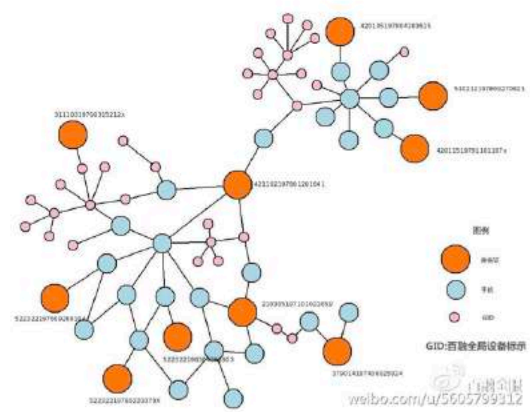


Bibliographic data




Movie data

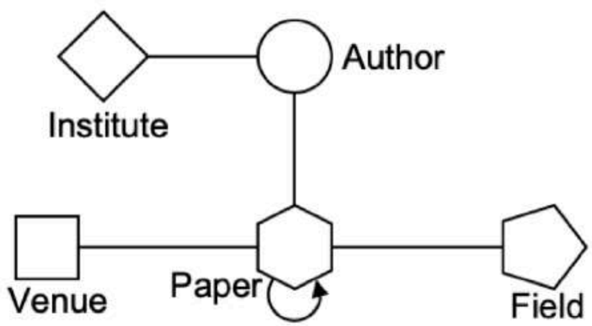
Social network data



Knowledge graph

*From Heterogeneous Graph Neural Network and its Applications in E-Commerce Prof. Chuan Shi*

# OAG Graph




(a) The schema of heterogeneous academic networks

| | | |
|---|---|---|
| Author | *is_(first/last/other)_author_of* | Paper |
| Author | *is_affiliated_with* | Institute |
| Paper | *is_published_(conf/journal)_at* | Venue |
| Paper | *has_*$(L_1 - L_5)$*_field_of* | Field |
| Paper | *has_citation_to* | Paper |

(b) The meta relations of heterogeneous academic networks

**Figure 1: The schema and meta relations of Open Academic Graph (OAG).** Given a Web-scale heterogeneous graph, e.g., an academic network, HGT takes only its one-hop edges as input without manually designing meta paths.

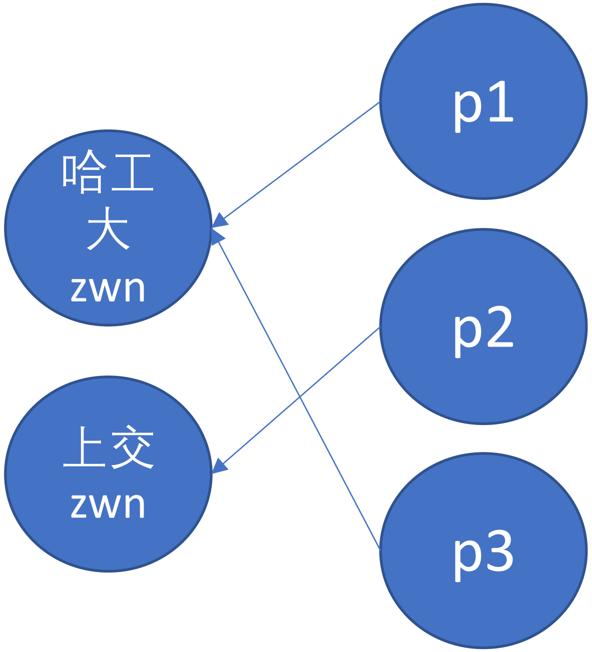
# OAG Graph

| Dataset | #nodes | #edges | #papers | #authors | #fields | #venues | #institutes |
|---|---|---|---|---|---|---|---|
| CS | 11,732,027 | 107,263,811 | 5,597,605 | 5,985,759 | 119,537 | 27,433 | 16,931 |
| Med | 51,044,324 | 451,468,375 | 21,931,587 | 28,779,507 | 289,930 | 25,044 | 18,256 |
| OAG | 178,663,927 | 2,236,196,802 | 89,606,257 | 88,364,081 | 615,228 | 53,073 | 25,288 |

| Dataset | #P-A | #P-F | #P-V | #A-I | #P-P |
|---|---|---|---|---|---|
| CS | 15,571,614 | 47,462,559 | 5,597,606 | 7,190,480 | 31,441,552 |
| Med | 85,620,479 | 149,728,483 | 21,931,588 | 28,779,507 | 165,408,318 |
| OAG | 300,853,688 | 657,049,405 | 89,606,258 | 167,449,933 | 1,021,237,518 |

# Tasks

- Node Classification
  - Paper-Field prediction
    - Paper–Field (L1)
    - Paper–Field (L2)
  - Paper-Venue prediction
- Link prediction
  - Author Disambiguation tasks



# Heterogeneous Graph

each node $v \in \mathcal{V}$ each edge $e \in \mathcal{E}$

$$G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$$

Directed graph

$$\tau(v) : V \rightarrow \mathcal{A} \qquad \phi(e) : E \rightarrow \mathcal{R}$$

Type mapping functions

$v =<$ Heterogeneous Graph Transformer $>$ $e = (HGT, HAN)$

$\tau(v) = paper$ $\emptyset(e) = cited$

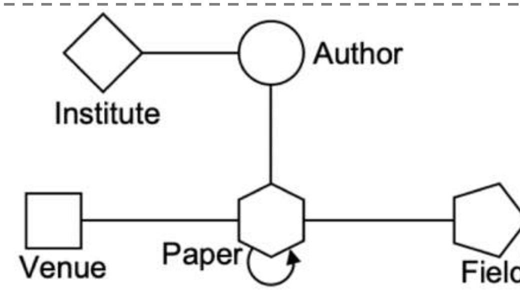
# Meta Relation

$$e = (s, t)$$

$$\langle \tau(s), \phi(e), \tau(t) \rangle$$

$$e = (HGT, HAN)$$

$$< \tau(s), \emptyset(e), \tau(t) >$$
$$=< paper, cited, paper >$$



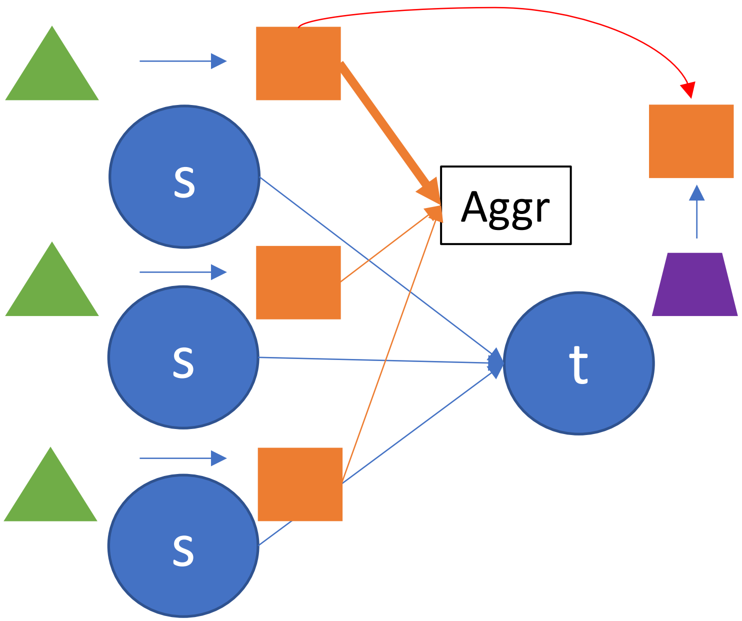

(a) The schema of
heterogeneous academic networks

# Model

- Heterogeneous Mutual Attention
- Heterogeneous Message Passing
- Target-Specific Aggregation

# Heterogeneous Mutual Attention

$$H^l[t] \leftarrow \underset{\forall s \in N(t), \forall e \in E(s,t)}{\textbf{Aggregate}} \left(\textbf{Attention}(s,t) \cdot \textbf{Message}(s)\right)$$



# Heterogeneous Mutual Attention

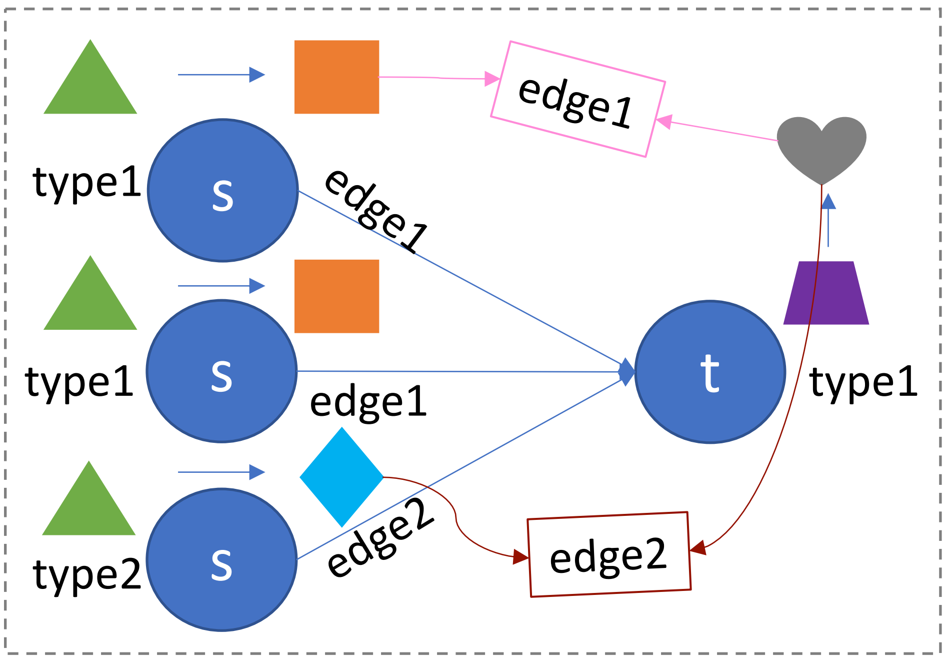
$$\textbf{Attention}_{HGT}(s,e,t) = \underset{\forall s \in N(t)}{\text{Softmax}}\left(\underset{i\in[1,h]}{\Big\|} ATT\text{-}head^i(s,e,t)\right) \quad (3)$$

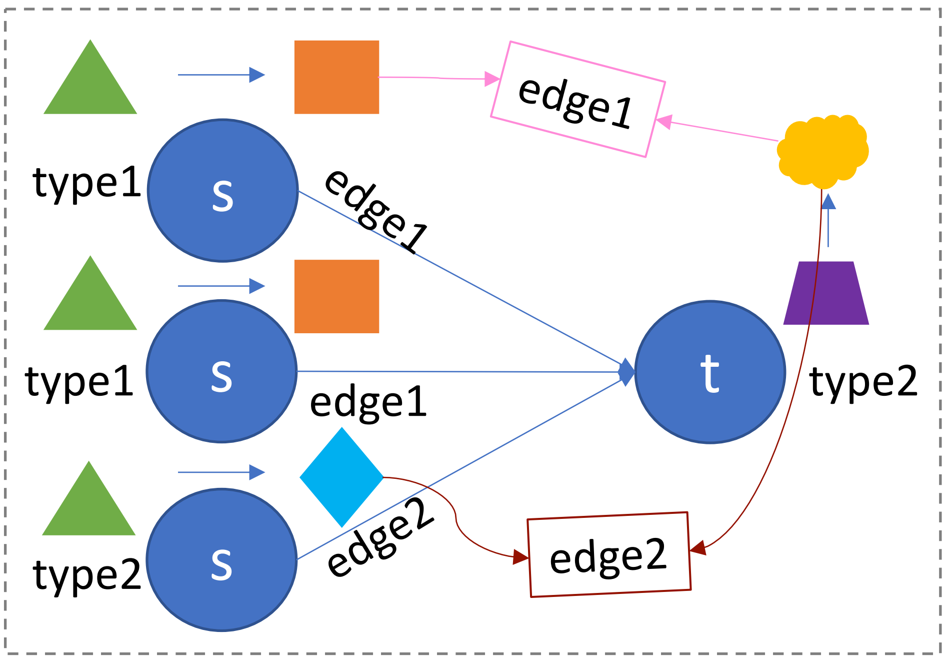
$$ATT\text{-}head^i(s,e,t) = \left(K^i(s)\ W^{ATT}_{\phi(e)}\ Q^i(t)^T\right) \cdot \frac{\mu_{\langle \tau(s),\phi(e),\tau(t)\rangle}}{\sqrt{d}}$$

$$K^i(s) = \text{K-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right)$$

$$Q^i(t) = \text{Q-Linear}^i_{\tau(t)}\left(H^{(l-1)}[t]\right)$$

$$\boxed{\mu \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{R}|\times|\mathcal{A}|}}$$

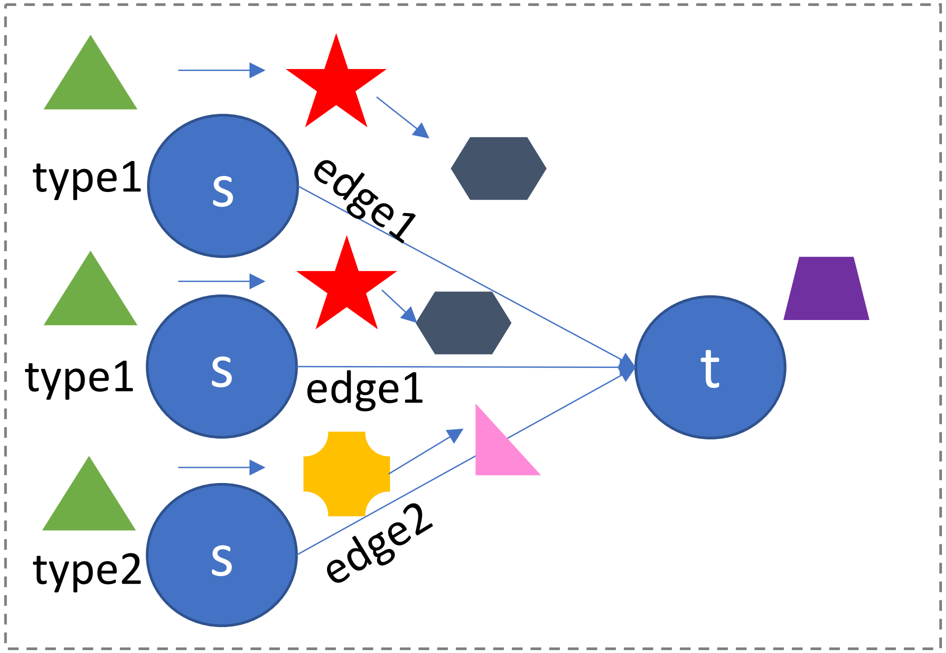


# Heterogeneous Message Passing

$$\mathbf{Message}_{HGT}(s,e,t) = \mathop{\Big\|}_{i\in[1,h]} MSG\text{-}head^i(s,e,t)$$

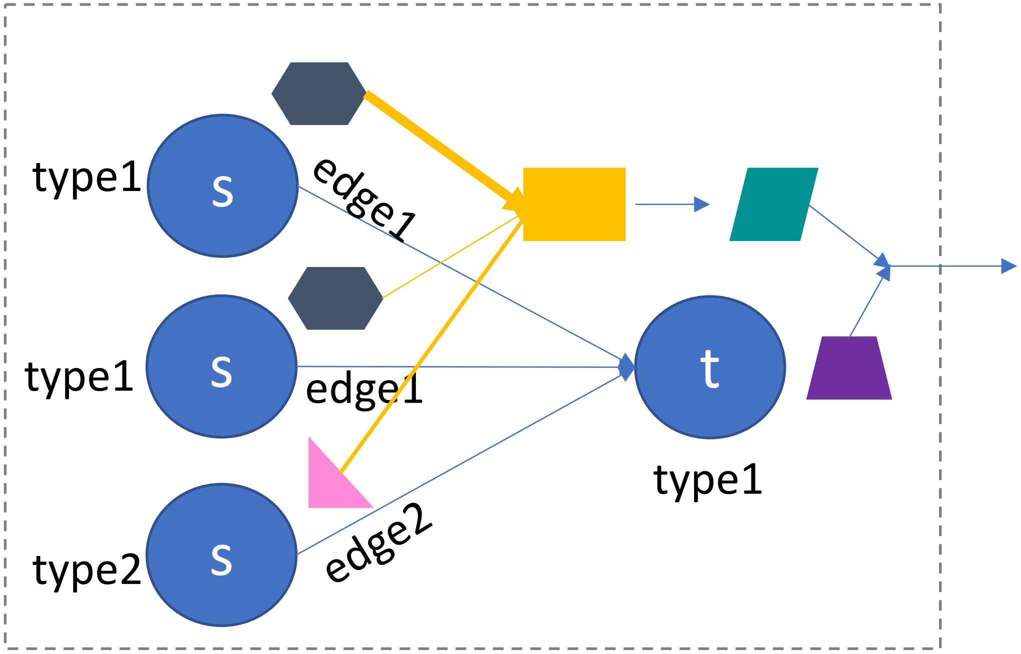
$$MSG\text{-}head^i(s,e,t) = \text{M-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) W^{MSG}_{\phi(e)}$$



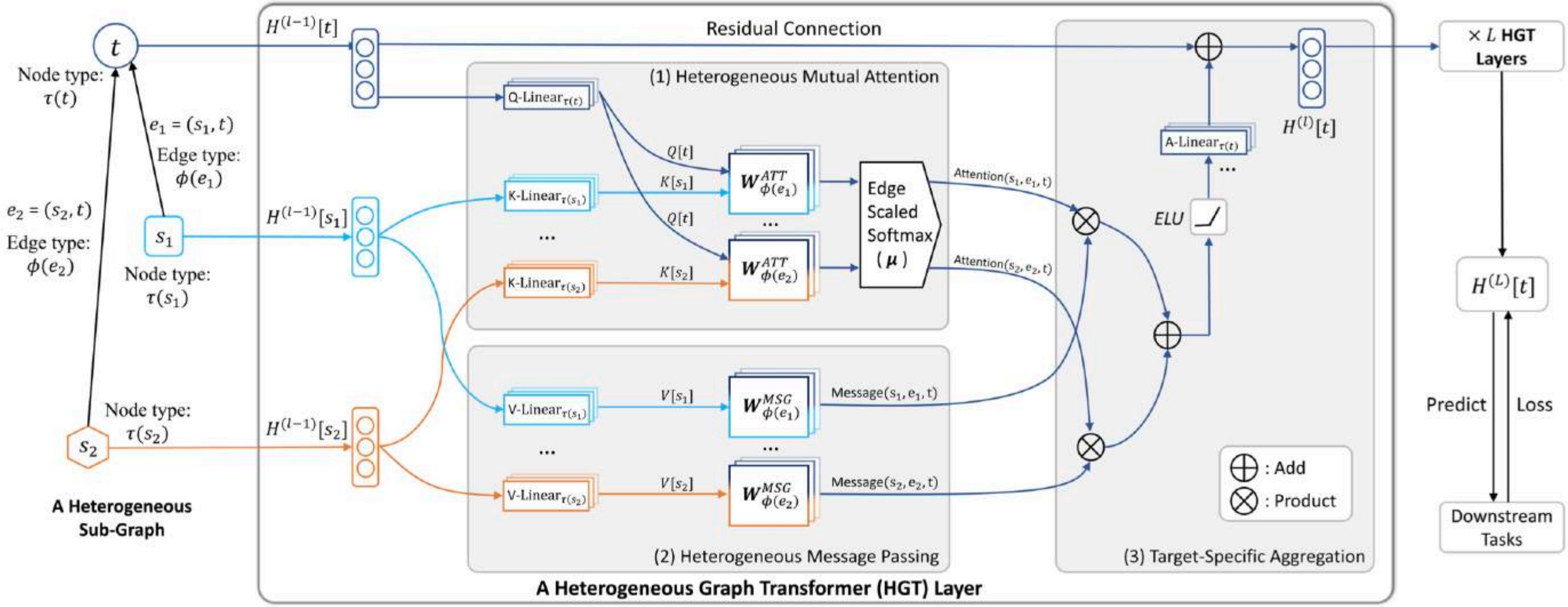
# Target-Specific Aggregation

$$\widetilde{H}^{(l)}[t] = \bigoplus_{\forall s \in N(t)} \Big( \mathbf{Attention}_{HGT}(s,e,t) \cdot \mathbf{Message}_{HGT}(s,e,t) \Big).$$

$$H^{(l)}[t] = \text{A-Linear}_{\tau(t)}\Big(\sigma\big(\widetilde{H}^{(l)}[t]\big)\Big) + H^{(l-1)}[t].$$



# Overall Architecture



# Dynamic Heterogeneous Graph

$v = HGT$

$e = (HGT, WWW)$

$WWW$ 2020

$v = HAN$

$e = (HGT, WWW)$

$WWW$ 2019

$e = (HGT, WWW) \rightarrow$ timestamp 2020

$v = HGT \rightarrow$ timestamp 2020

$v = WWW \rightarrow$ timestamp 2020

$e = (HAN, WWW) \rightarrow$ timestamp 2019

$v = HAN \rightarrow$ timestamp 2019

$v = WWW \rightarrow$ timestamp 2019

# Relative Temporal Encoding

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

$$[\sin(3/10000^{0/128}), \cos(3/10000^{0/128}), \sin(3/10000^{2/128}), \cos(3/10000^{2/128}), \ldots]$$

Transformer

$$\Delta T(t,s) = T(t) - T(s)$$

$$Base\big(\Delta T(t,s), 2i\big) = sin\Big(\Delta T_{t,s}/10000^{\frac{2i}{d}}\Big)$$

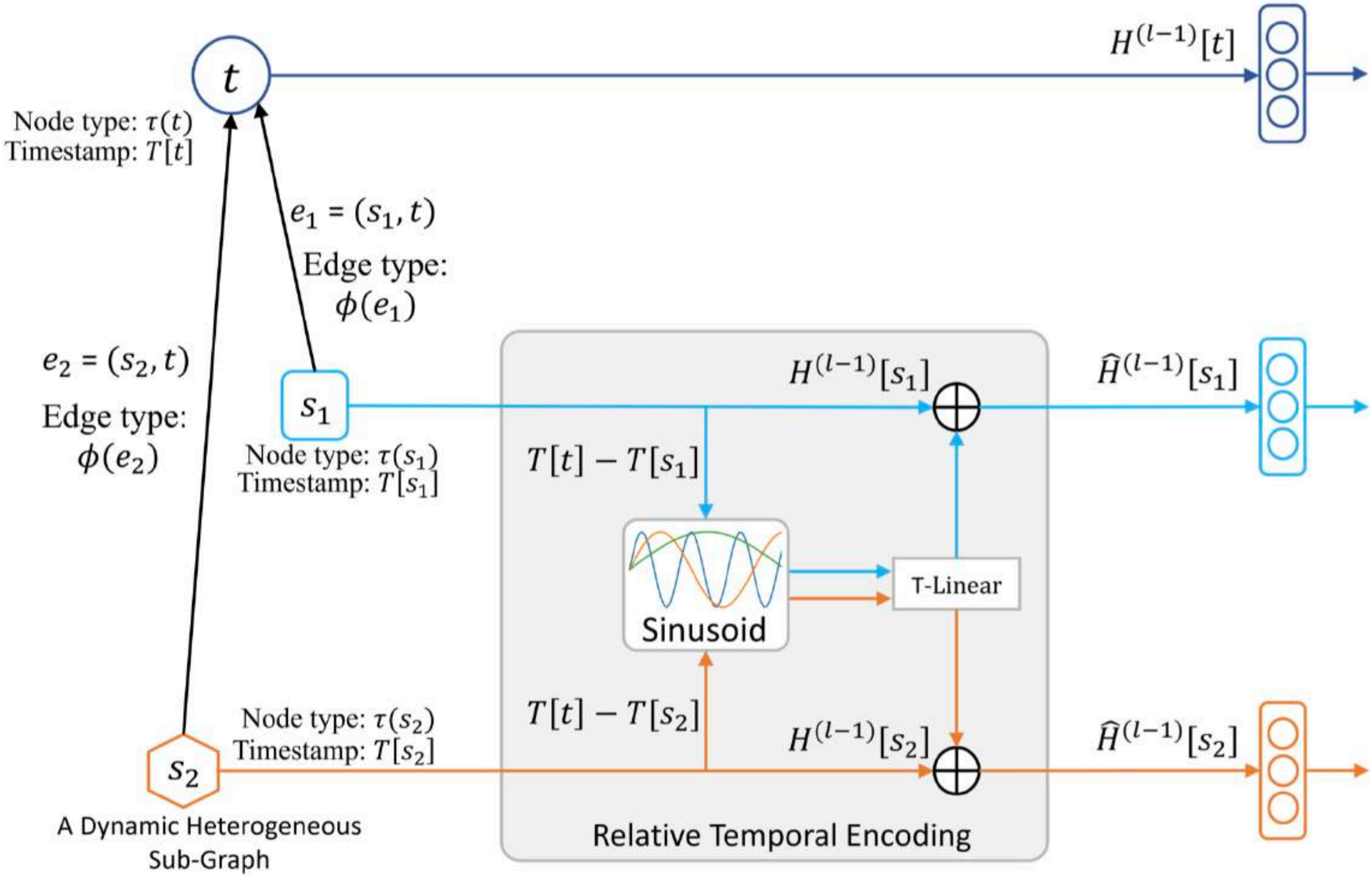
$$Base\big(\Delta T(t,s), 2i+1\big) = cos\Big(\Delta T_{t,s}/10000^{\frac{2i+1}{d}}\Big)$$

$$RTE\big(\Delta T(t,s)\big) = \text{T-Linear}\Big(Base(\Delta T_{t,s})\Big)$$

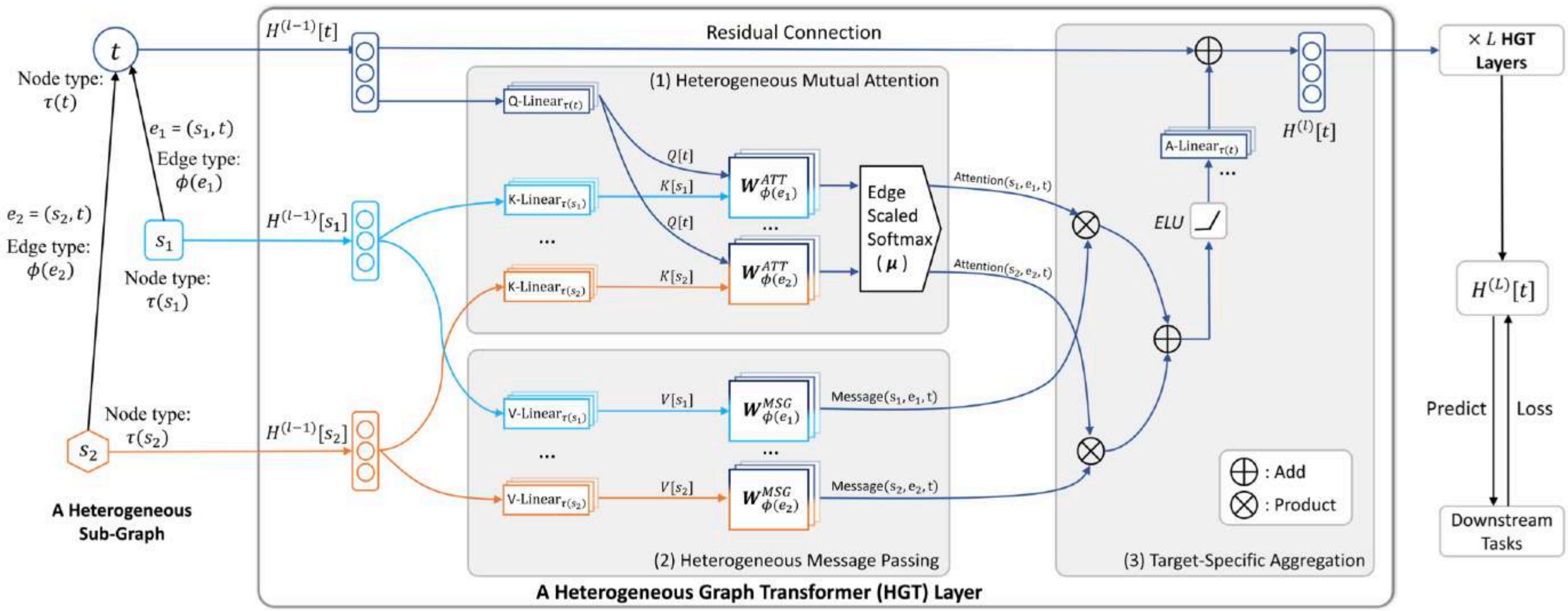
$$\widehat{H}^{(l-1)}[s] = H^{(l-1)}[s] + RTE\big(\Delta T(t,s)\big)$$

Relative Temporal Encoding

# Relative Temporal Encoding



# Overall Architecture



# HGSampling

- keep a similar number of nodes and edges for each type
- keep the sampled sub-graph dense to minimize the information loss and reduce the sample variance.

# Baselines

- GCN
- GAT
- R-GCN
- HetGNN (KDD19 Heterogeneous Graph Neural Network)
- HAN (WWW19 Heterogeneous Graph Attention Network)

$\text{HGT}_{-Heter}^{-RTE}$ $\text{HGT}_{-Heter}^{+RTE}$ $\text{HGT}_{+Heter}^{-RTE}$ $\text{HGT}_{+Heter}^{+RTE}$

# Input Features

- Paper
  - pre-trained XLNet to get the representation of each word in its title.
  - Average them weighted by each word’s attention to get the title representation for each paper.
- Author
  - average of his/her published papers’ representations
- Field, venue, and institute
  - metapath2vec

# Results

| GNN Models | | | GCN [9] | RGCN [14] | GAT [22] | HetGNN [27] | HAN [23] | HGT$^{-RTE}_{-Heter}$ | HGT$^{+RTE}_{-Heter}$ | HGT$^{-RTE}_{+Heter}$ | HGT$^{+RTE}_{+Heter}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of Parameters | | | 1.69M | 8.80M | 1.69M | 8.41M | 9.45M | 3.12M | 3.88M | 7.44M | 8.20M |
| Batch Time | | | 0.46s | 1.24s | 0.97s | 1.35s | 2.27s | 1.11s | 1.14s | 1.48s | 1.50s |
| CS | Paper–Field ($L_1$) | NDCG | .608±.062 | .603±.065 | .622±.071 | .612±.063 | .618±.058 | .662±.051 | .689±.042 | .705±.036 | **.718±.014** |
| | | MRR | .679±.069 | .683±.056 | .694±.065 | .689±.060 | .691±.051 | .751±.036 | .779±.027 | .799±.023 | **.823±.019** |
| | Paper–Field ($L_2$) | NDCG | .344±.021 | .322±.053 | .357±.058 | .346±.071 | .352±.051 | .362±.048 | .371±.043 | .379±.047 | **.403±.041** |
| | | MRR | .353±.053 | .340±.061 | .382±.057 | .373±.051 | .388±.065 | .394±.072 | .397±.064 | .414±.076 | **.439±.078** |
| | Paper–Venue | NDCG | .406±.081 | .412±.076 | .437±.082 | .431±.074 | .449±.072 | .456±.069 | .461±.066 | .468±.074 | **.473±.054** |
| | | MRR | .215±.066 | .216±.105 | .239±.089 | .245±.069 | .254±.074 | .258±.085 | .265±.090 | .275±.089 | **.288±.088** |
| | Author Disambiguation | NDCG | .826±.039 | .835±.042 | .864±.051 | .850±.056 | .859±.053 | .867±.048 | .875±.046 | .886±.048 | **.894±.034** |
| | | MRR | .661±.045 | .665±.054 | .694±.052 | .668±.061 | .688±.049 | .703±.036 | .712±.032 | .727±.038 | **.732±.038** |
| Med | Paper–Field ($L_1$) | NDCG | .560±.056 | .571±.061 | .584±.076 | .598±.068 | .607±.054 | .654±.048 | .667±.045 | .683±.037 | **.709±.029** |
| | | MRR | .465±.055 | .470±.082 | .493±.069 | .509±.054 | .575±.057 | .620±.066 | .642±.062 | .659±.055 | **.688±.048** |
| | Paper–Field ($L_2$) | NDCG | .334±.035 | .337±.051 | .344±.063 | .342±.048 | .350±.059 | .359±.053 | .365±.047 | .374±.050 | **.384±.046** |
| | | MRR | .337±.061 | .343±.063 | .370±.058 | .373±.061 | .379±.052 | .385±.071 | .397±.069 | .408±.071 | **.417±.074** |
| | Paper–Venue | NDCG | .377±.059 | .383±.062 | .388±.065 | .412±.057 | .416±.068 | .421±.083 | .432±.078 | **.446±.083** | .445±.085 |
| | | MRR | .211±.045 | .217±.058 | .244±.091 | .259±.072 | .271±.056 | .277±.081 | .282±.085 | .288±.074 | **.291±.062** |
| | Author Disambiguation | MRR | .776±.042 | .779±.048 | .828±.044 | .824±.058 | .834±.056 | .838±.047 | .844±.041 | .864±.043 | **.871±.040** |
| | | NDCG | .614±.051 | .625±.049 | .663±.046 | .659±.061 | .667±.053 | .683±.055 | .691±.046 | .708±.041 | **.718±.043** |
| OAG | Paper–Field ($L_1$) | NDCG | .508±.141 | .511±.128 | .534±.103 | .543±.084 | .544±.096 | .571±.089 | .578±.086 | .595±.089 | **.615±.084** |
| | | MRR | .556±.136 | .565±.105 | .610±.096 | .616±.076 | .622±.092 | .649±.081 | .657±.078 | .675±.082 | **.702±.081** |
| | Paper–Field ($L_2$) | NDCG | .318±.074 | .328±.046 | .339±.049 | .336±.062 | .342±.051 | .350±.045 | .354±.046 | .358±.052 | **.367±.048** |
| | | MRR | .322±.067 | .332±.052 | .348±.045 | .350±.053 | .358±.049 | .362±.057 | .369±.058 | .371±.064 | **.378±.071** |
| | Paper–Venue | NDCG | .302±.066 | .313±.051 | .317±.057 | .309±.071 | .327±.062 | .334±.058 | .341±.059 | .353±.064 | **.355±.062** |
| | | MRR | .194±.070 | .193±.047 | .196±.052 | .192±.059 | .214±.067 | .229±.061 | .233±.060 | .243±.048 | **.247±.061** |
| | Author Disambiguation | NDCG | .738±.042 | .755±.048 | .797±.044 | .803±.058 | .821±.056 | .835±.043 | .841±.041 | .847±.043 | **.852±.048** |
| | | MRR | .612±.064 | .619±.057 | .645±.063 | .649±.052 | .660±.049 | .668±.059 | .674±.058 | .683±.066 | **.688±.054** |

**Table 2: Experimental results of different methods over the three datasets.**

# Visualize Meta Relation Attention

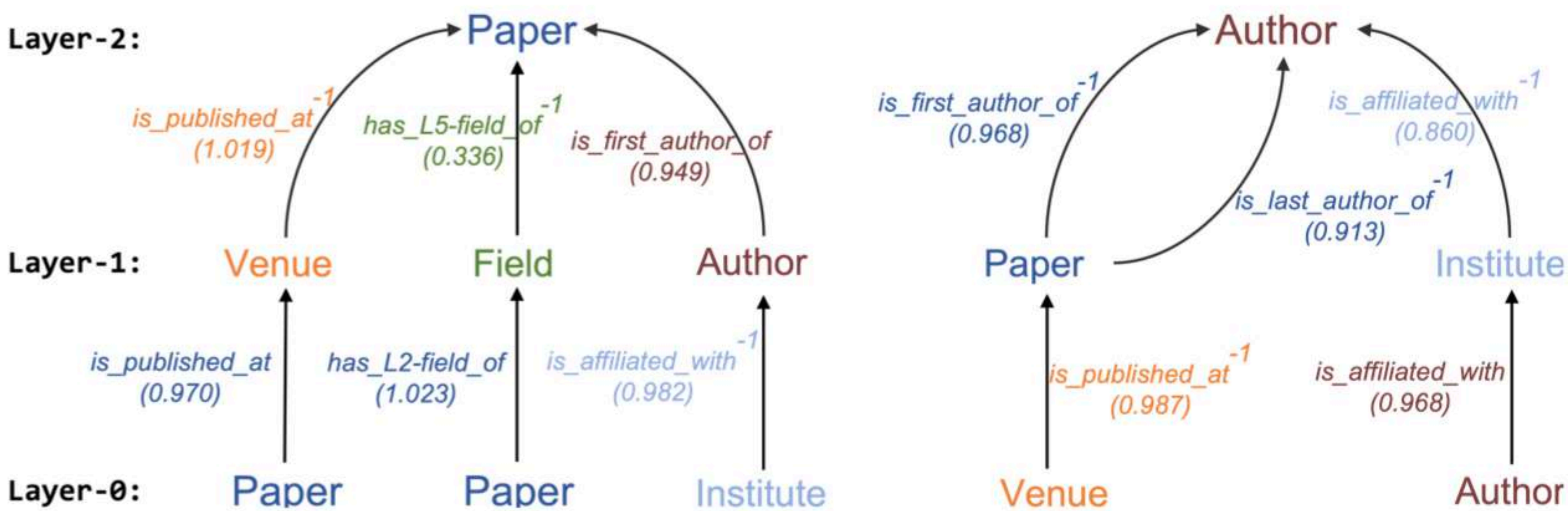



**Figure 5: Hierarchy of the learned meta relation attention.**

# Papers

| Paper | Conference |
|---|---|
| ☆ Heterogeneous Graph Transformer | WWW20 |
| Author Name Disambiguation on Heterogeneous Information Network with Adversarial Representation Learning | AAAI20 |
| Graph-Driven Generative Models for Heterogeneous Multi-Task Learning | AAAI20 |
| ☆ An Attention-based Graph Neural Network for Heterogeneous Structural Learning | AAAI20 |
| ☆ Spam Review Detection with Graph Convolutional Networks | CIKM19 |
| Heterogeneous Graph Learning for Visual Commonsense Reasoning | NIPS19 |
| Metapath-guided Heterogeneous Graph Neural Network for Intent Recommendation | KDD19 |
| ☆ Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification | EMNLP19 |
| ☆ Heterogeneous Graph Attention Network | WWW19 |

# Thanks!