BiANE: Bipartite Attributed Network Embedding

Wentao Huang[1], Yuchen Li[2], Yuan Fang[2], Ju Fan[1], Hongxia Yang[3]

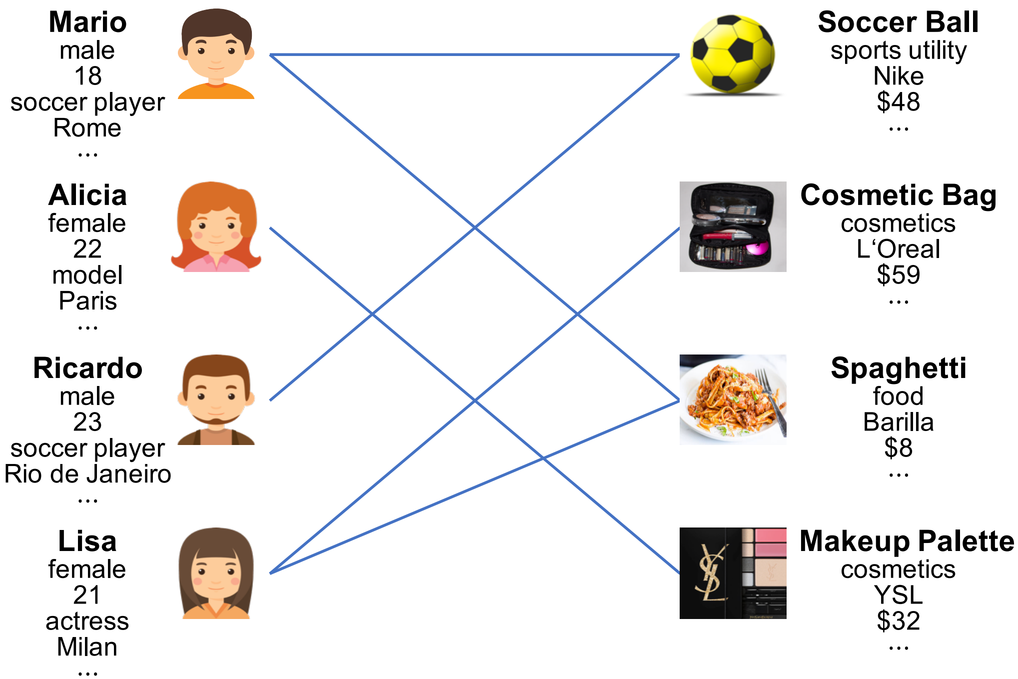
School of Information, Renmin University of China[1]
School of Information System, Singapore Management University[2]
Damo Academy, Alibaba Group[3]

Outline

- Introduction & Challenge
- Methodology
- Experiment
- Conclusion & Future Work

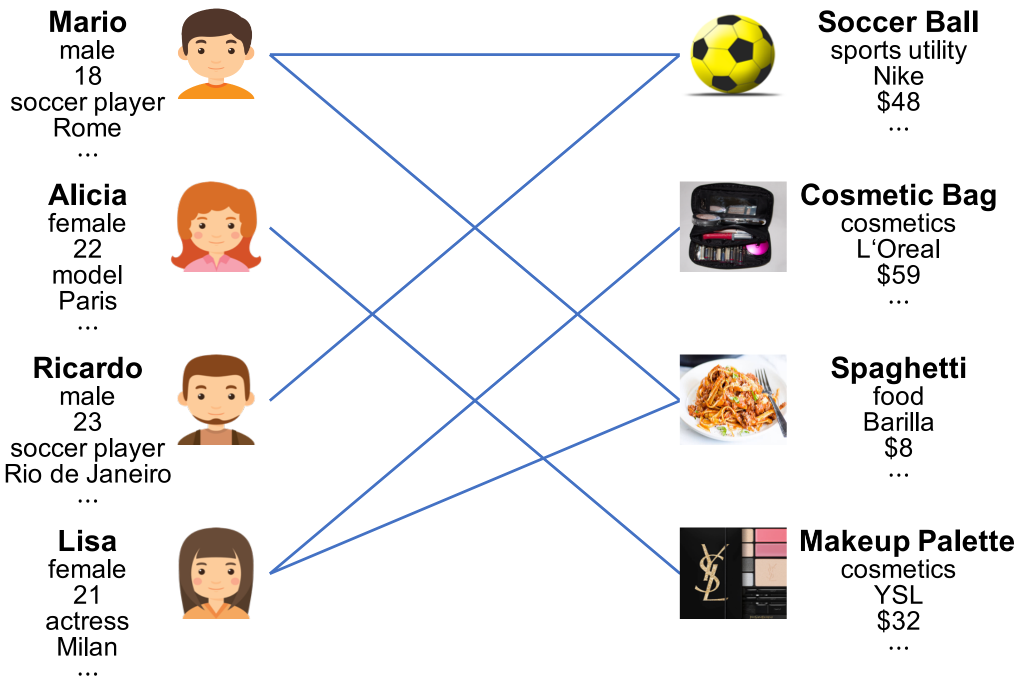
Introduction

- Bipartite Attributed Network
  - ✓ E-Commerce Websites
  - ✓ Recommendation System
  - ✓ Bibliometric Network Analysis
  - ✓ Biological Community Detection
  - ✓ Risk Assessment of Financial Systems



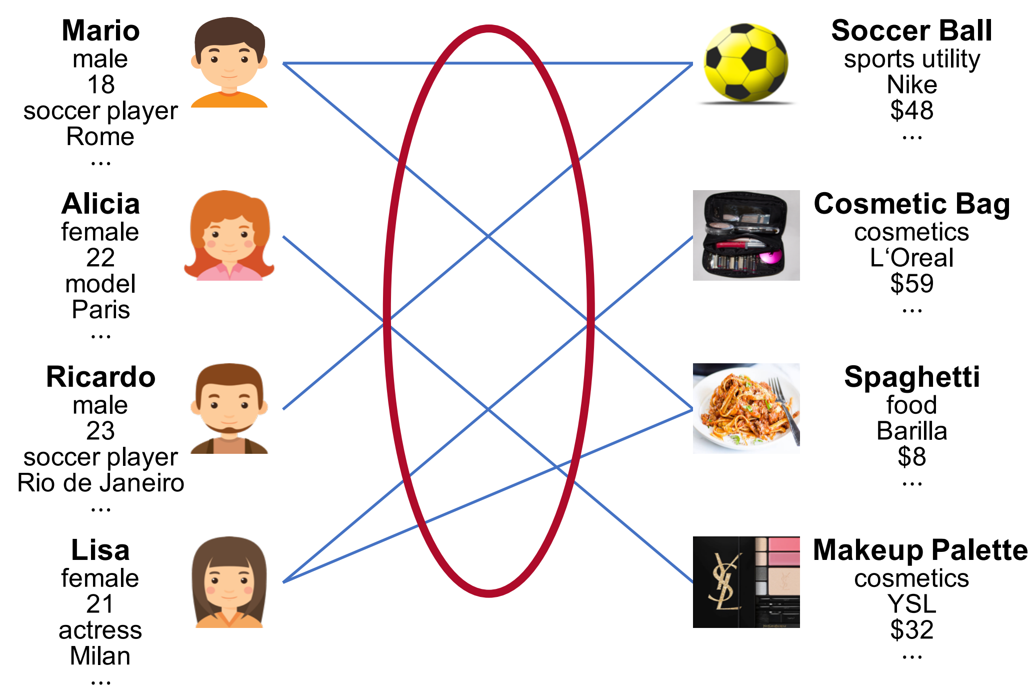
Introduction

- ❑ Bipartite Attributed Network
  - ✓ E-Commerce Websites
  - ✓ Recommendation System
  - ✓ Bibliometric Network Analysis
  - ✓ Biological Community Detection
  - ✓ Risk Assessment of Financial Systems
- ❑ Characteristics



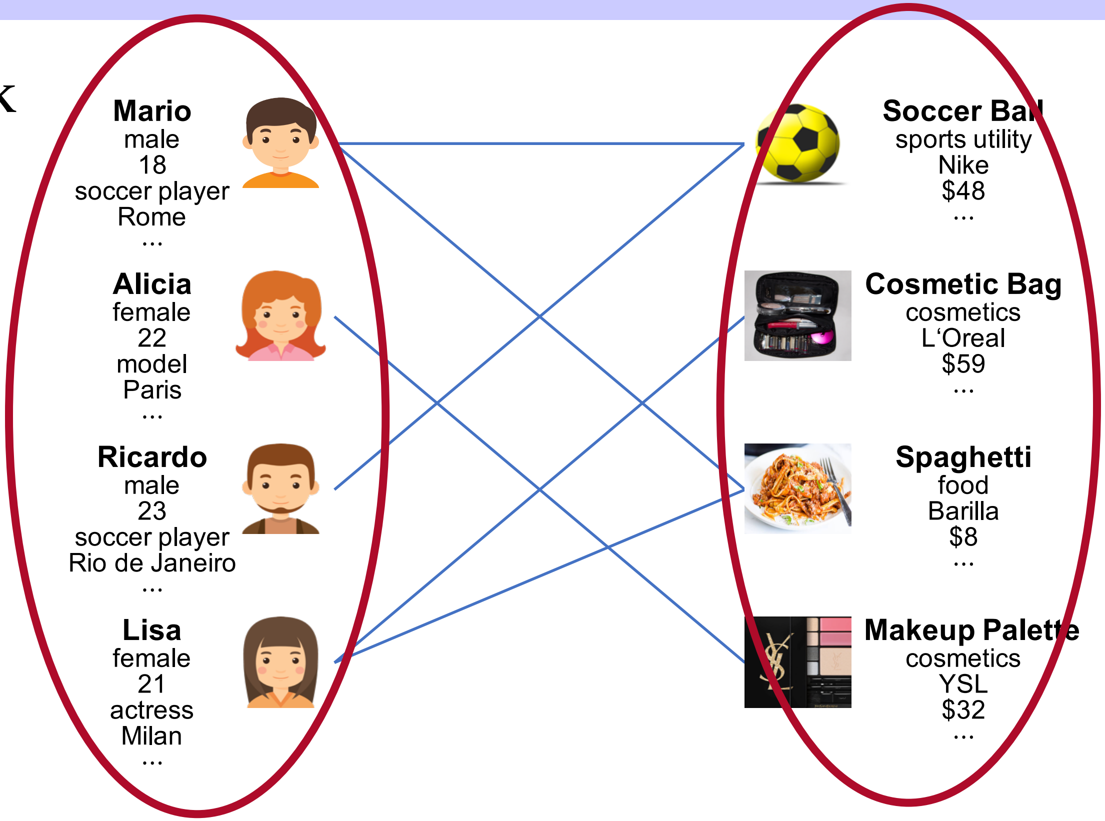
Introduction

- Bipartite Attributed Network
  - E-Commerce Websites
  - Recommendation System
  - Bibliometric Network Analysis
  - Biological Community Detection
  - Risk Assessment of Financial Systems
- Characteristics
  - The Inter-Partition Proximity



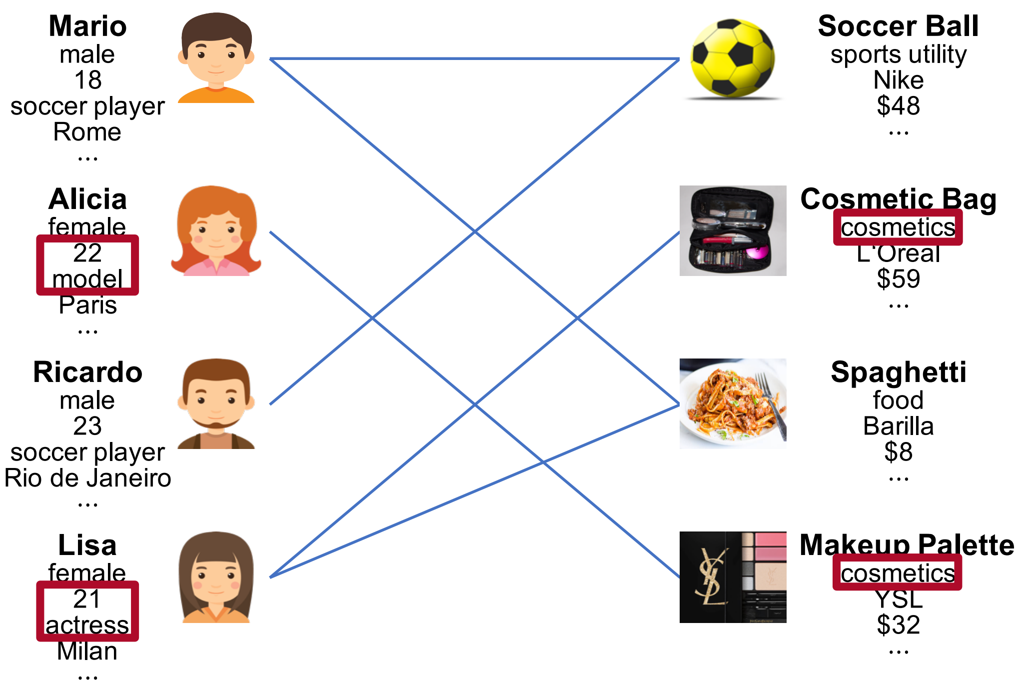
Introduction

- Bipartite Attributed Network
  - E-Commerce Websites
  - Recommendation System
  - Bibliometric Network Analysis
  - Biological Community Detection
  - Risk Assessment of Financial Systems
- Characteristics
  - The Inter-Partition Proximity
  - The Intra-Partition Proximity



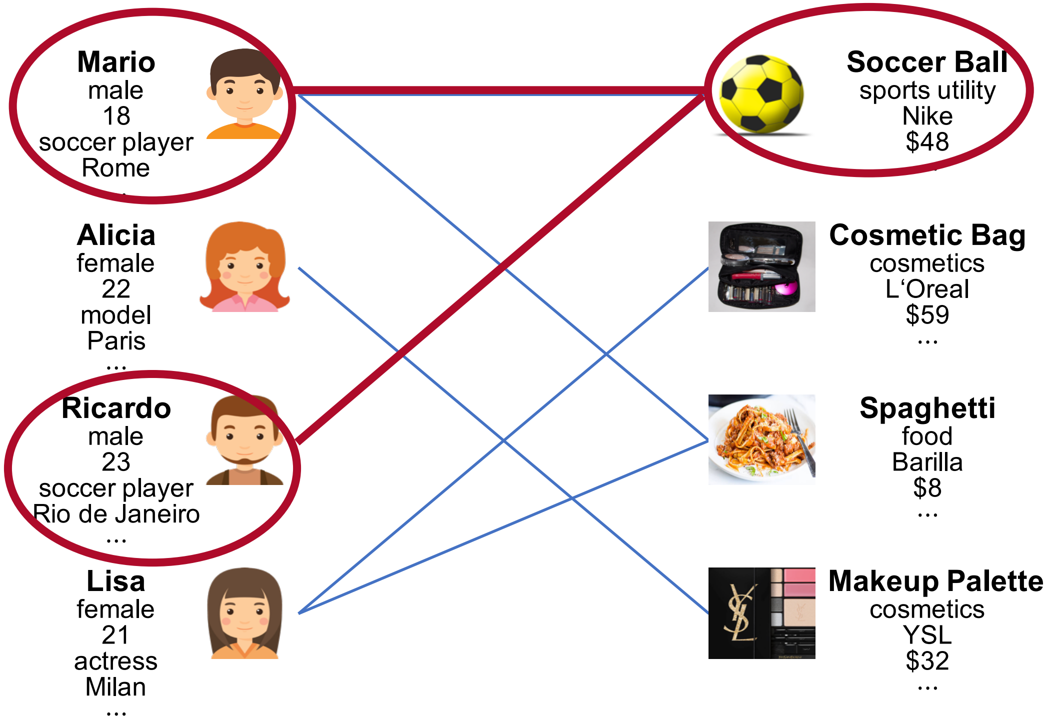
Introduction

- Bipartite Attributed Network
  - E-Commerce Websites
  - Recommendation System
  - Bibliometric Network Analysis
  - Biological Community Detection
  - Risk Assessment of Financial Systems
- Characteristics
  - The Inter-Partition Proximity
  - The Intra-Partition Proximity
    1) The Attribute Proximity



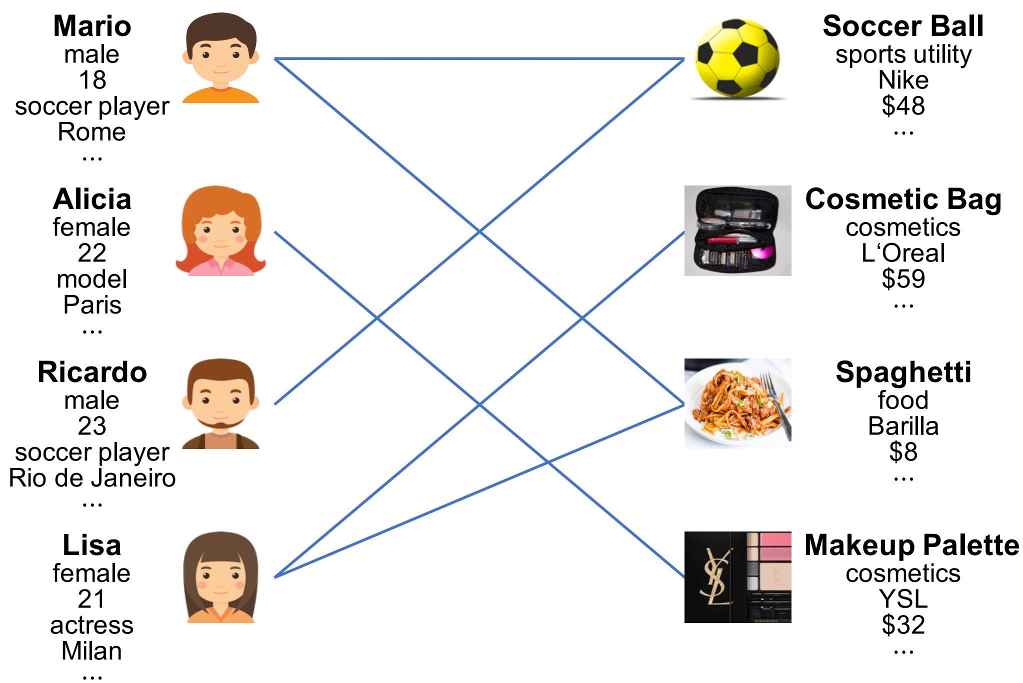
Introduction

- Bipartite Attributed Network
  - E-Commerce Websites
  - Recommendation System
  - Bibliometric Network Analysis
  - Biological Community Detection
  - Risk Assessment of Financial Systems
- Characteristics
  - The Inter-Partition Proximity
  - The Intra-Partition Proximity
    1) The Attribute Proximity
    2) The Structure Proximity



Introduction

- Bipartite Attributed Network
  - ✓ E-Commerce Websites
  - ✓ Recommendation System
  - ✓ Bibliometric Network Analysis
  - ✓ Biological Community Detection
  - ✓ Risk Assessment of Financial Systems
- Characteristics
  - The Inter-Partition Proximity
  - The Intra-Partition Proximity
    1) The Attribute Proximity
    2) The Structure Proximity

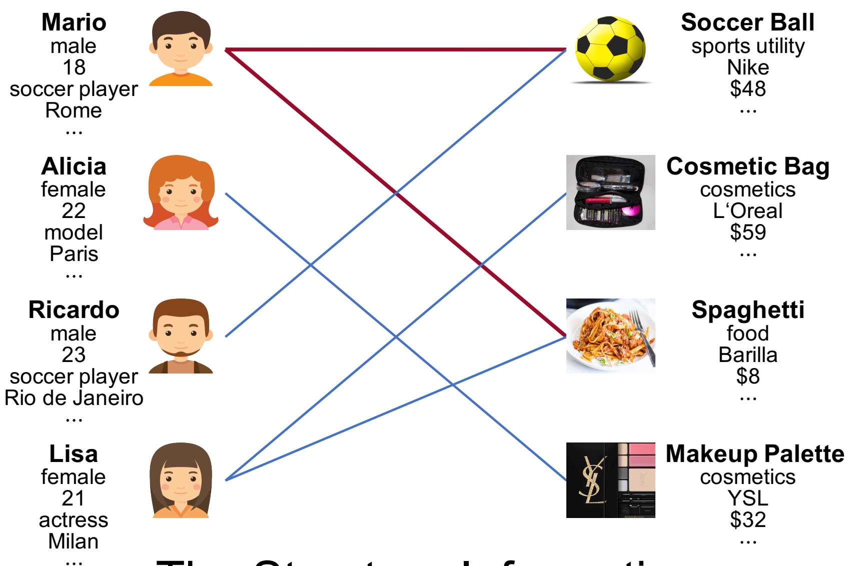


- Goal:

Given a bipartite attributed network $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathsf{E}, \mathbf{X}_{\mathcal{U}}, \mathbf{X}_{\mathcal{V}})$, we want to learn a mapping function to transform each node to a vector in a low-dimension space.

Technical Challenges

- The Attribute-Structure Correlation
  - Complementarity & Coherence

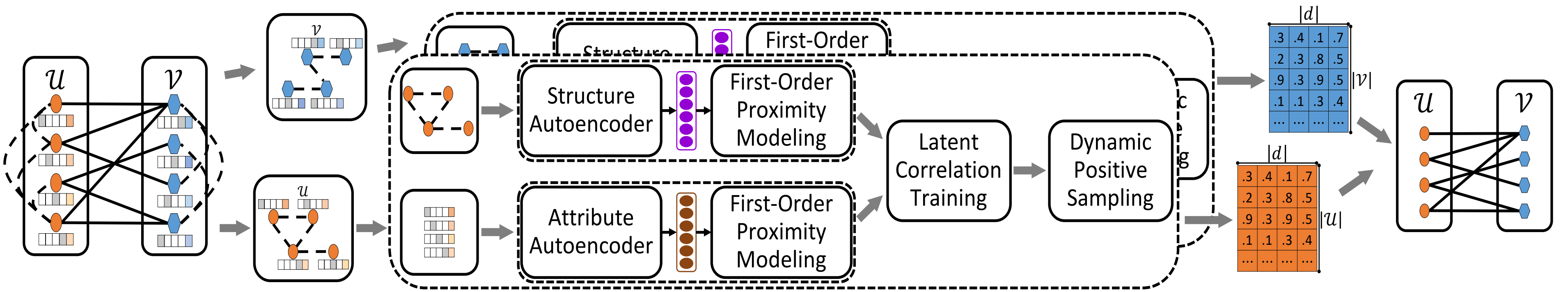


The Structure Information

The Attribute Information

- Negative Sampling Strategy
  - Static sampling strategies can not reflect the variation of embedding space.
  - Dynamic sampling strategies will result in the scalability issue.

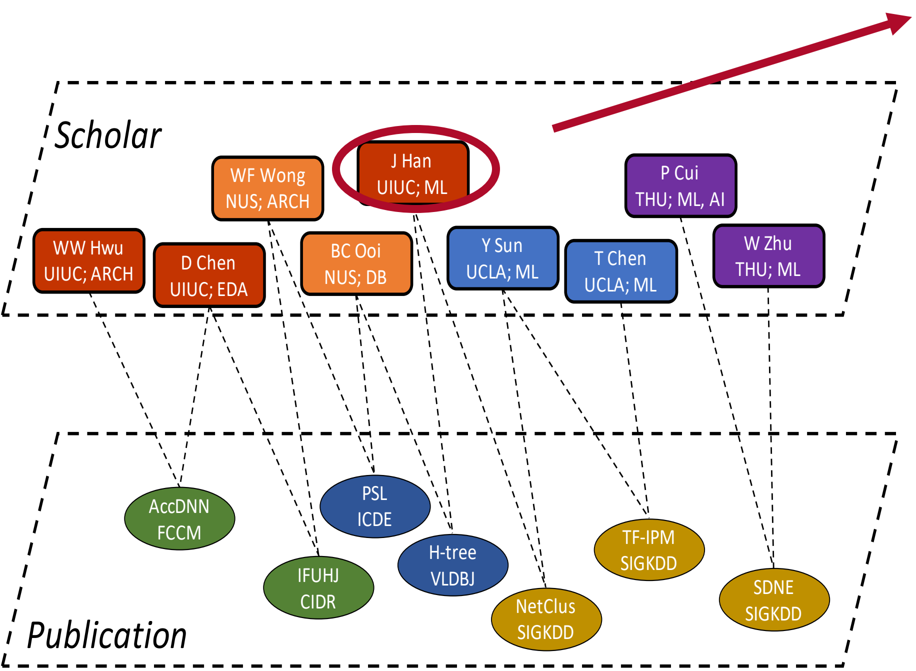
Methodology



$\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E}, \mathbf{X}_{\mathcal{U}}, \mathbf{X}_{\mathcal{V}})$ Intra-Partition Proximity Modeling Inter-Partition Proximity Modeling

Example

- Scholar-Publication Network



- Jiawei Han
- Gender: Male
- Institutions: UIUC, SFU
- Research Interests:
  - Data Mining
  - Database Systems
  - Data Warehousing
  - Information Networks

Scholar Partition:

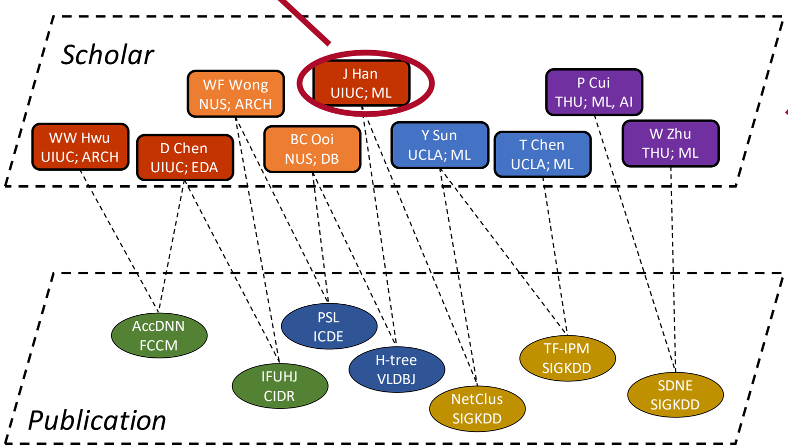
WW Hwu: Wen-mei W. Hwu	WF Wong: Weng-Fai Wong	BC Ooi: Beng Chin Ooi
D Chen: Deming Chen	W Zhu: Wenwu Zhu	T Chen: Ting Chen
Y Sun: Yizhou Sun	J Han: Jiawei Han	P Cui: Peng Cui

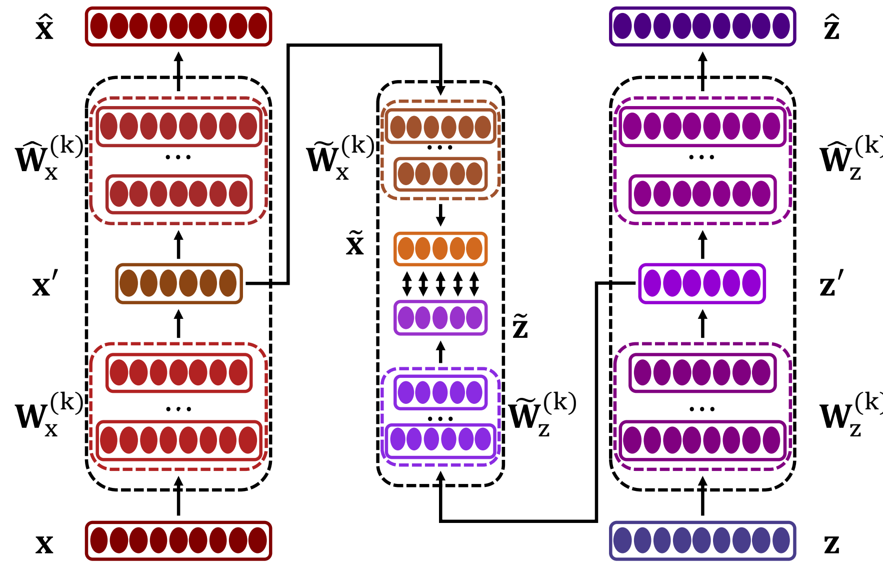
Publication Partition:

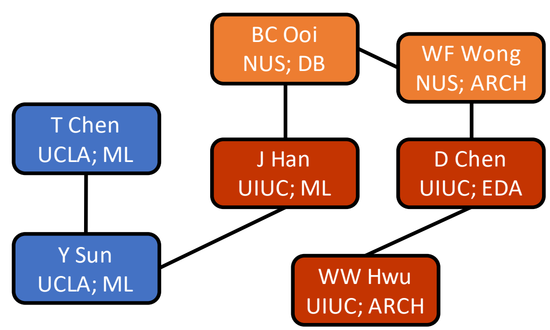
TF-IPM: Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network.
IFUHJ: Is FPGA Useful for Hash Joins?
AccDNN: An IP-Based DNN Generator for FPGAs.
PSL: Parallelizing Skip Lists for In-Memory Multi-Core Database Systems.
SDNE: Structural Deep Network Embedding.
H-tree: Index nesting – an efficient approach to indexing in object-oriented databases.
NetClus: Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema.

Intra-Partition Proximity Modeling

- Jiawei Han
- Gender: Male
- Institutions: UIUC, SFU
- Research Interests:
  - Data Mining
  - Database Systems
  - Data Warehousing
  - Information Networks

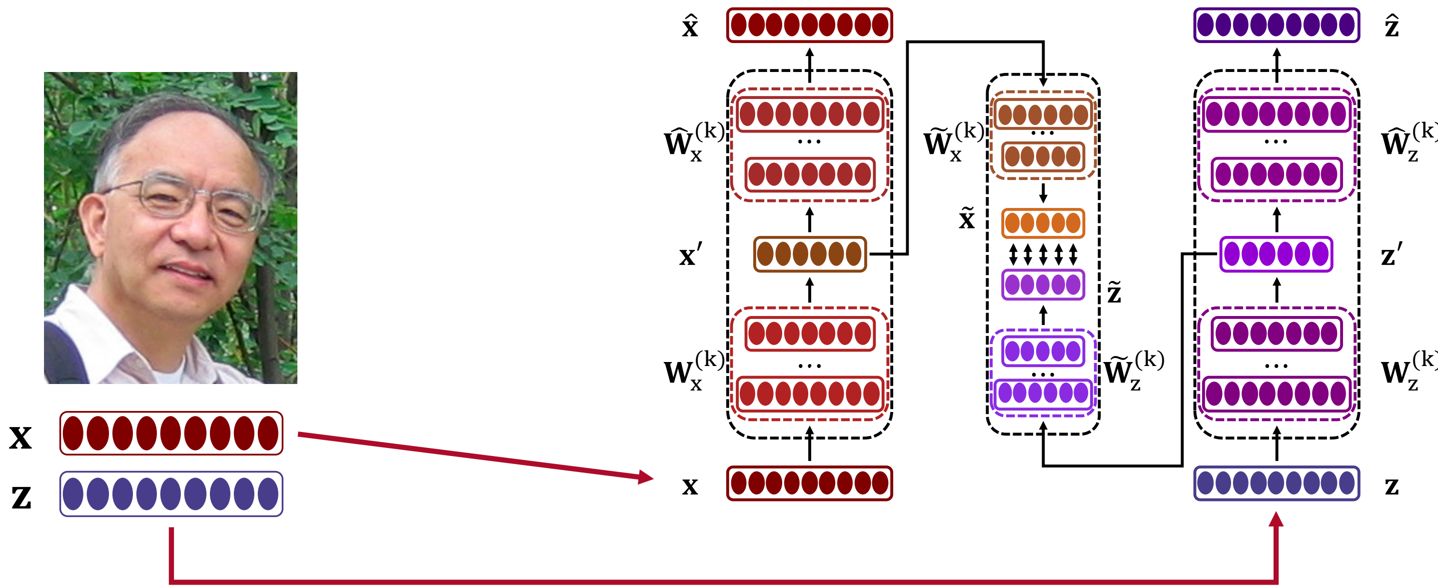




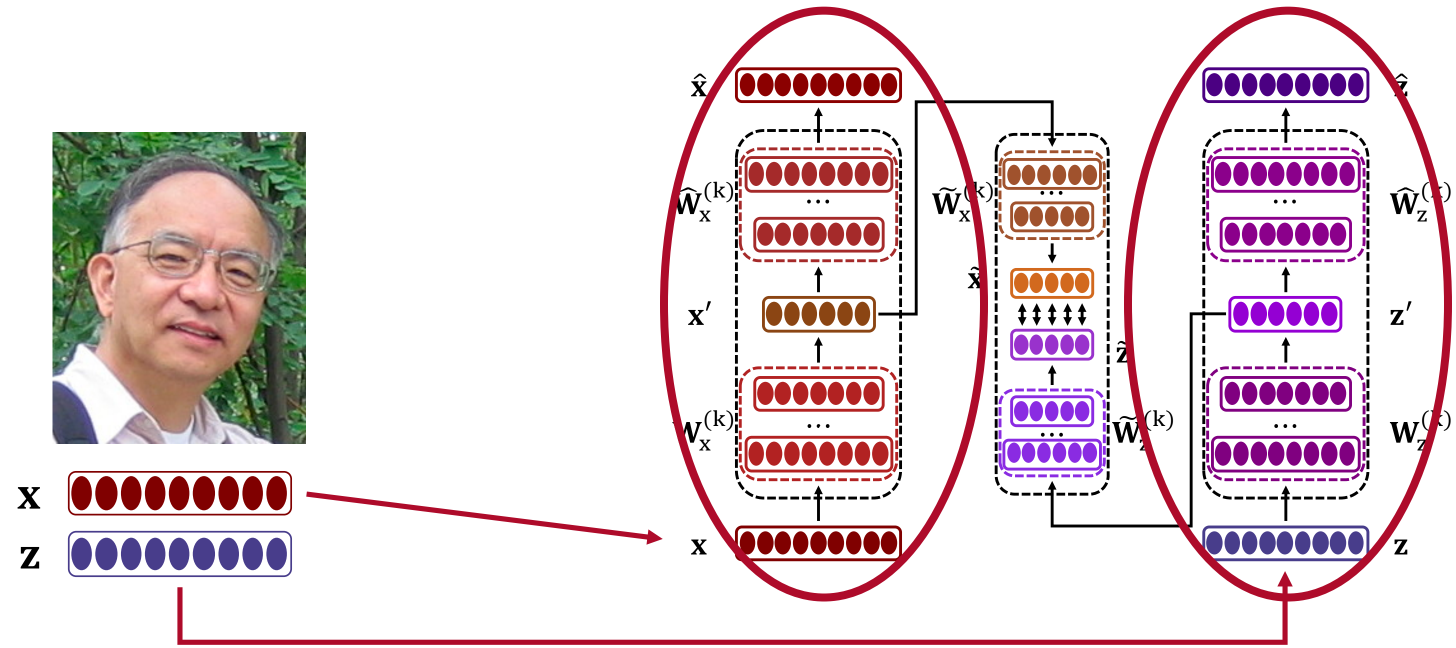


$$Z = \hat{A} + \hat{A}^2 + \cdots\cdots + \hat{A}^{k-1} + \hat{A}^k$$

Intra-Partition Proximity Modeling



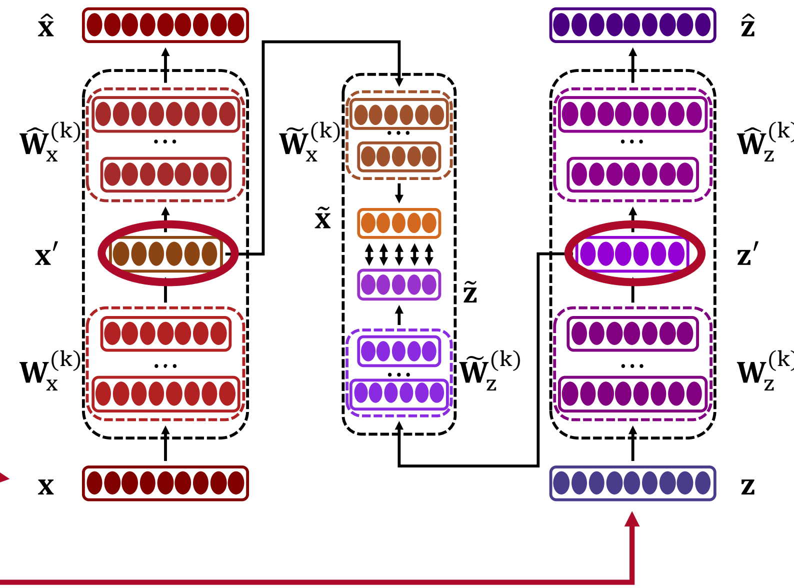
Intra-Partition Proximity Modeling



❑ Compact Feature Learning

$$L_2 = \sum_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 + \sum_i \|\hat{\mathbf{z}}_i - \mathbf{z}_i\|^2$$

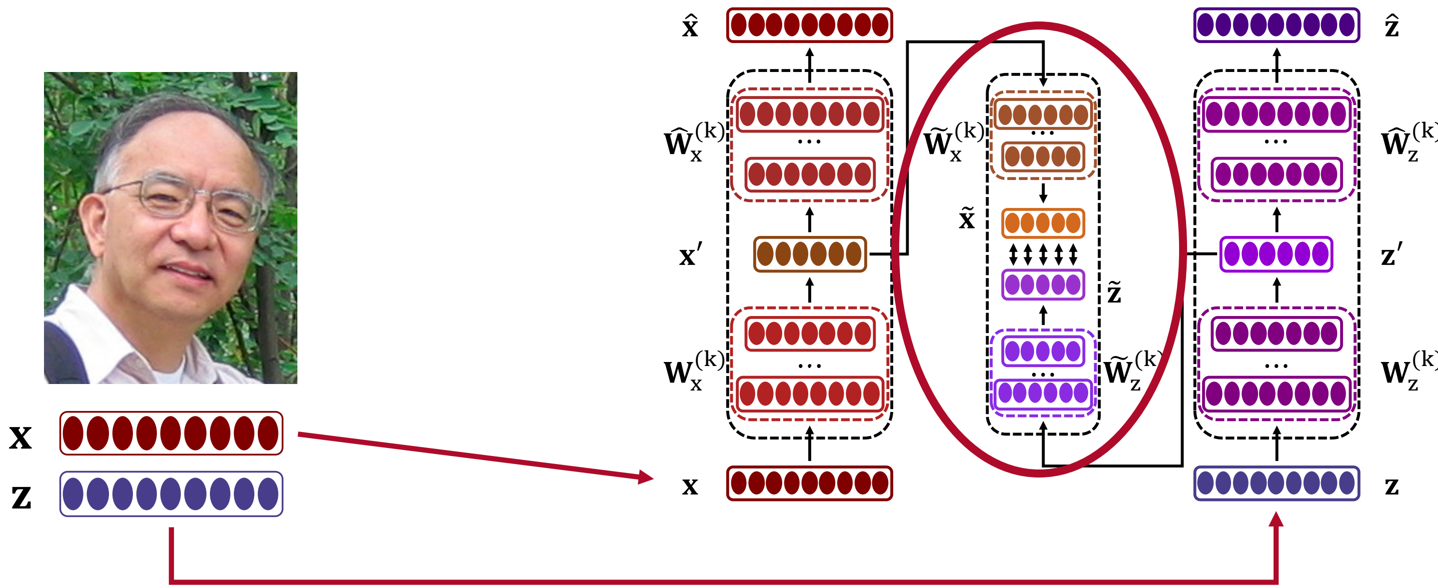
Intra-Partition Proximity Modeling



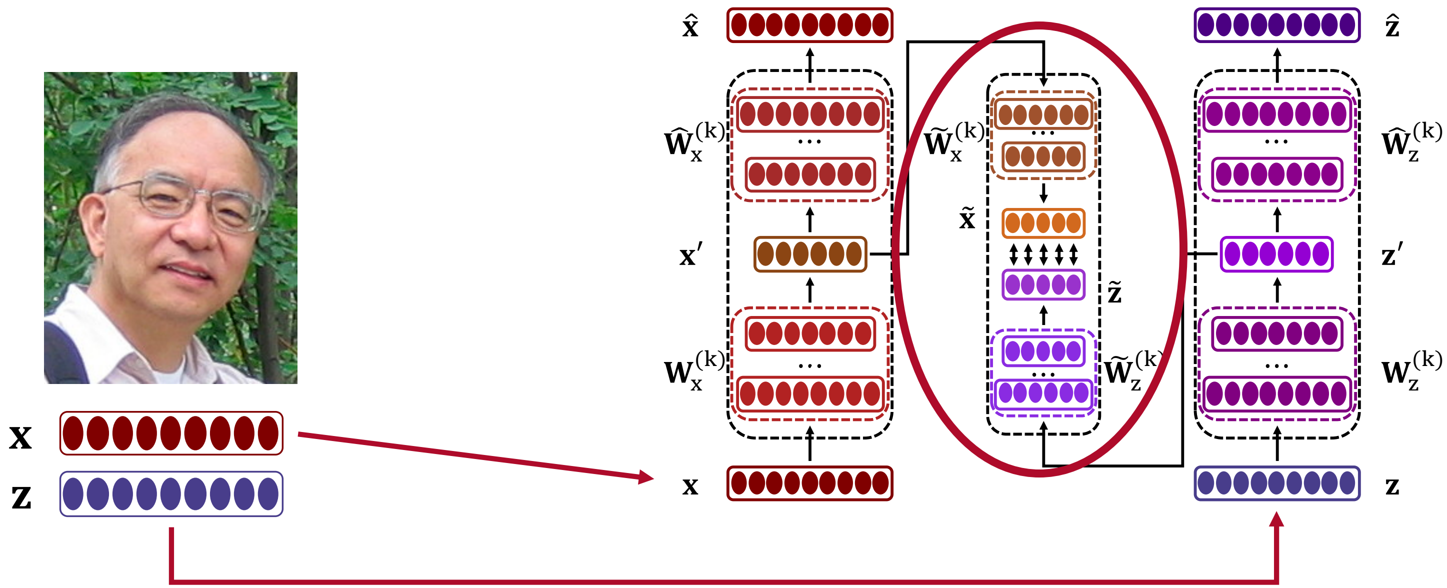
- ❑ Joint Modeling — Preserving the first-order proximity

$$
\begin{aligned}
L_3 = & - \sum_{a_{mn}>0} \log \sigma({\mathbf{x}'_m}^{\mathrm{T}} \cdot \mathbf{x}'_n) - \sum_{n'=1} \mathbb{E}_{v_{n'} \sim P'_n(v)} \log \sigma(-{\mathbf{x}'_m}^{\mathrm{T}} \cdot \mathbf{x}'_{n'}) \\
& - \sum_{a_{mn}>0} \log \sigma({\mathbf{z}'_m}^{\mathrm{T}} \cdot \mathbf{z}'_n) - \sum_{n'=1} \mathbb{E}_{v_{n'} \sim P'_n(v)} \log \sigma(-{\mathbf{z}'_m}^{\mathrm{T}} \cdot \mathbf{z}'_{n'})
\end{aligned}
$$

Latent Correlation Training



Latent Correlation Training

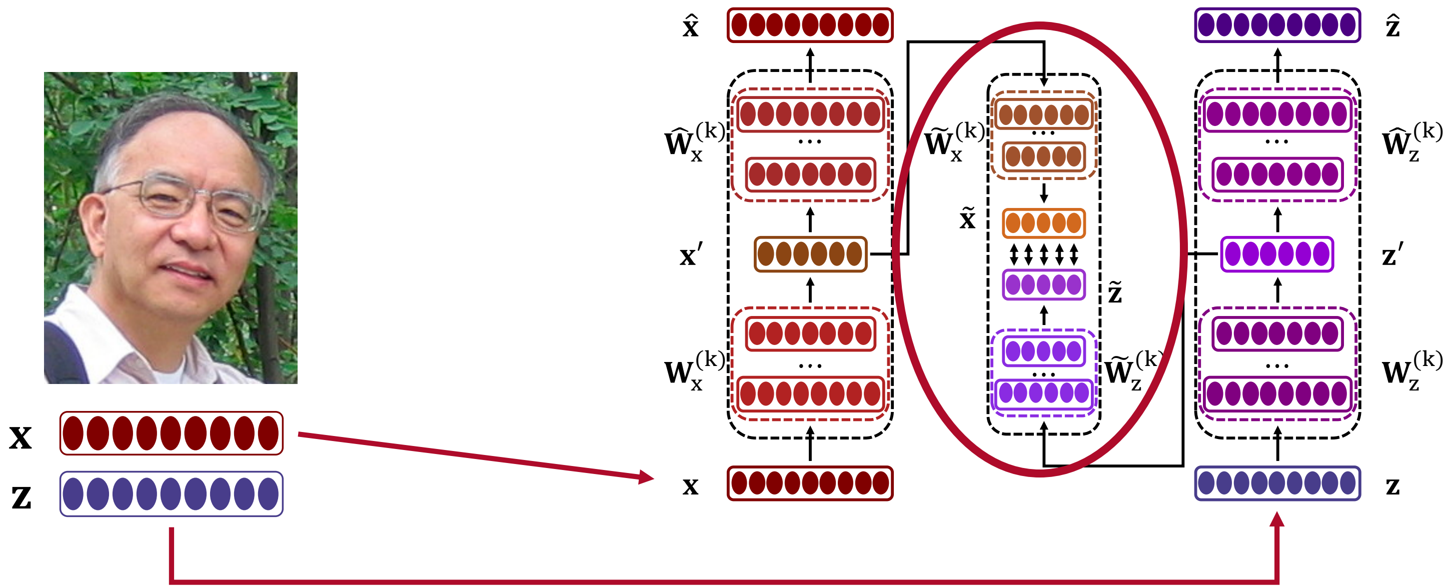


❑ Transform encodings to latent representations via auxiliary kernels.

$$\tilde{\mathbf{x}} = \delta^{(k)}(\widetilde{\mathbf{W}}_x^{(k)}(\cdots\delta^{(1)}(\widetilde{\mathbf{W}}_x^{(1)}\mathbf{x}' + \tilde{\mathbf{b}}_x^{(1)})\cdots) + \tilde{\mathbf{b}}_x^{(k)})$$

$$\tilde{\mathbf{z}} = \delta^{(k)}(\widetilde{\mathbf{W}}_z^{(k)}(\cdots\delta^{(1)}(\widetilde{\mathbf{W}}_z^{(1)}\mathbf{z}' + \tilde{\mathbf{b}}_z^{(1)})\cdots) + \tilde{\mathbf{b}}_z^{(k)})$$

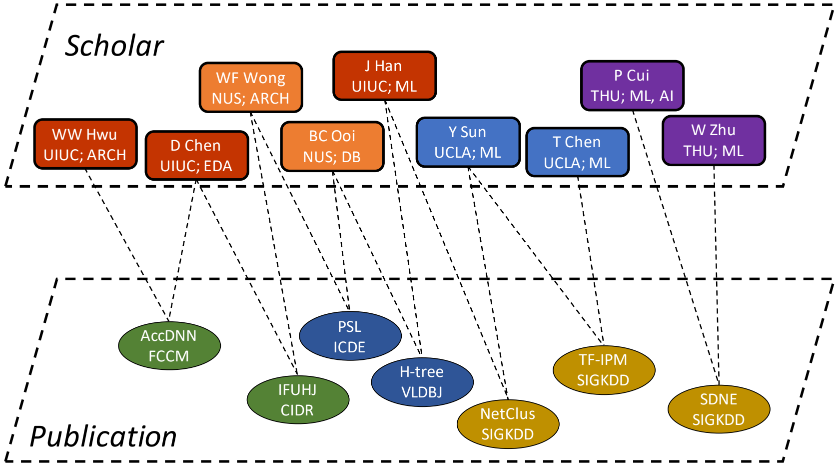
Latent Correlation Training



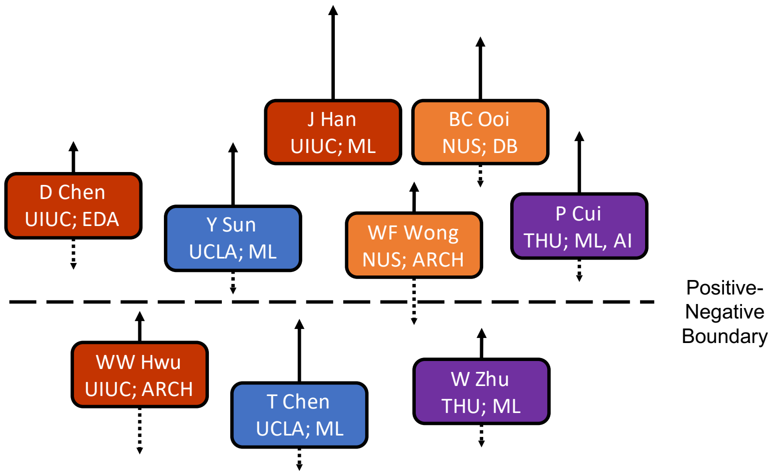
- ❑ Enhance the attribute-structure correlation

$$L_4 = -\sum_{m=n} \log \sigma(\tilde{\mathbf{x}}_m^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_n) - \sum_{n'=1} \mathbb{E}_{v_{n'} \sim P'_n(v)} \log \sigma(-\tilde{\mathbf{x}}_m{}^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_{n'})$$

Dynamic Positive Sampling

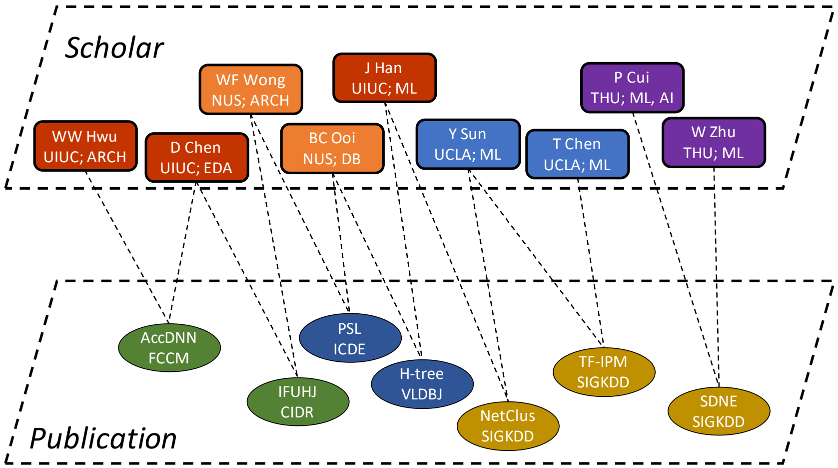


Scholar-Publication Network

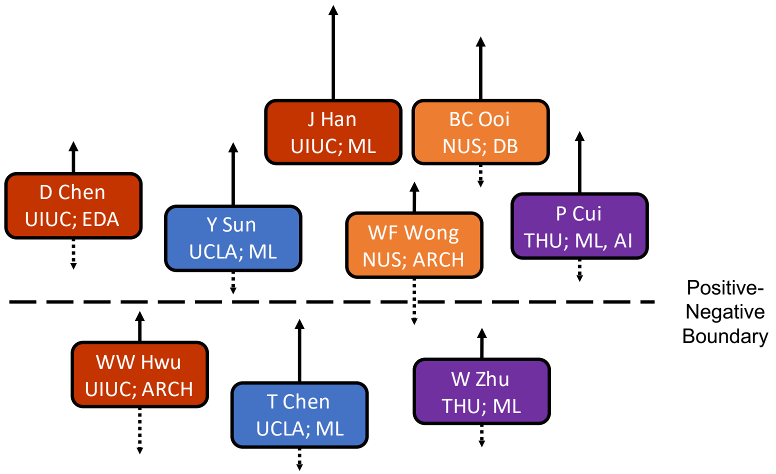


Dynamic Positive Sampling

Dynamic Positive Sampling



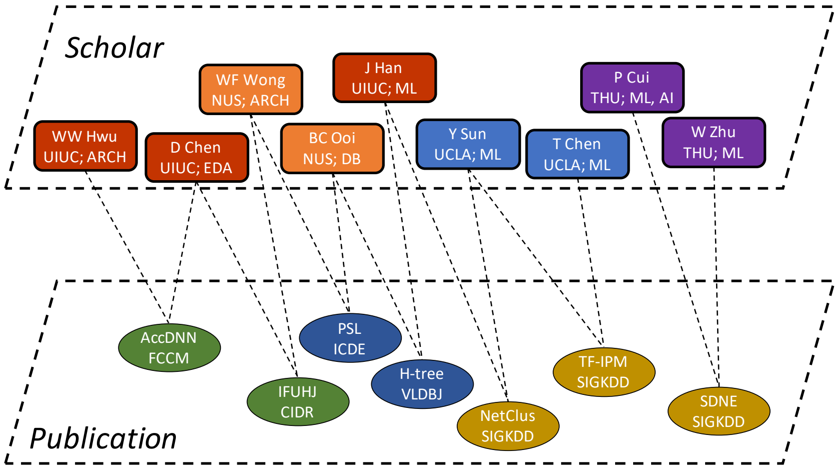
Scholar-Publication Network



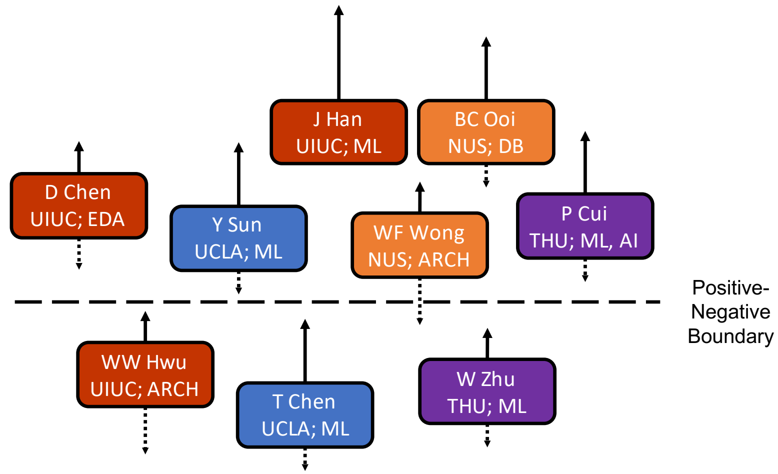
Dynamic Positive Sampling

- Build up HNSW index for each vector $(\tilde{x}, \tilde{z})$ in the latent space (time complexity: $O(n \log n)$)

Dynamic Positive Sampling



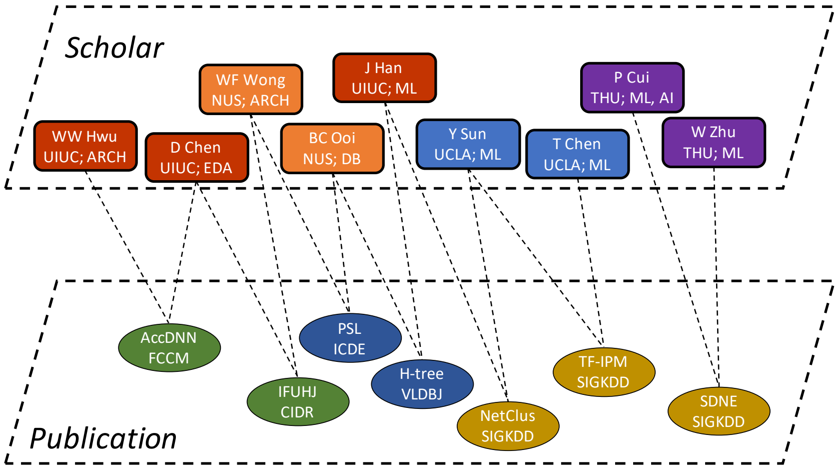
Scholar-Publication Network



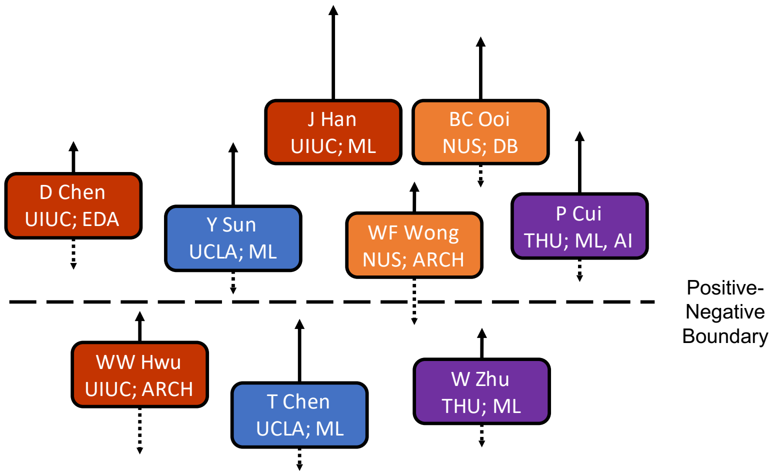
Dynamic Positive Sampling

- Build up HNSW index for each vector $(\tilde{x}, \tilde{z})$ in the latent space (time complexity: $O(n \log n)$)
- Perform kNN approximate search for each vector $(\tilde{x}, \tilde{z})$ via HNSW (time complexity: $O(n \log n)$)

Dynamic Positive Sampling



Scholar-Publication Network

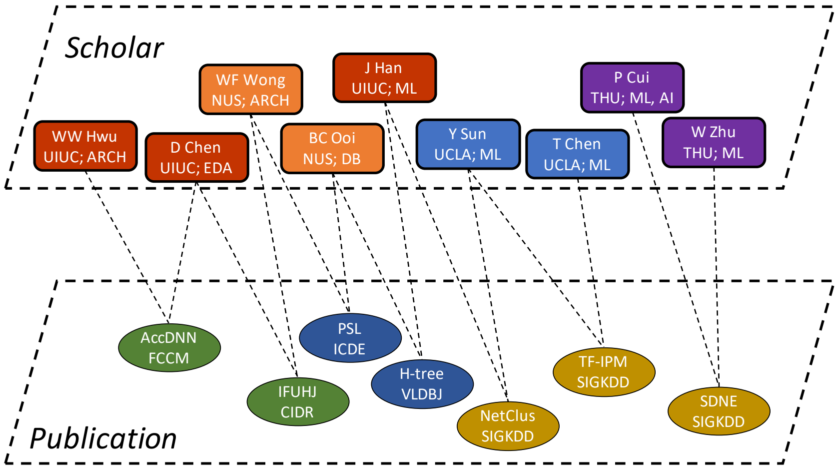


Dynamic Positive Sampling

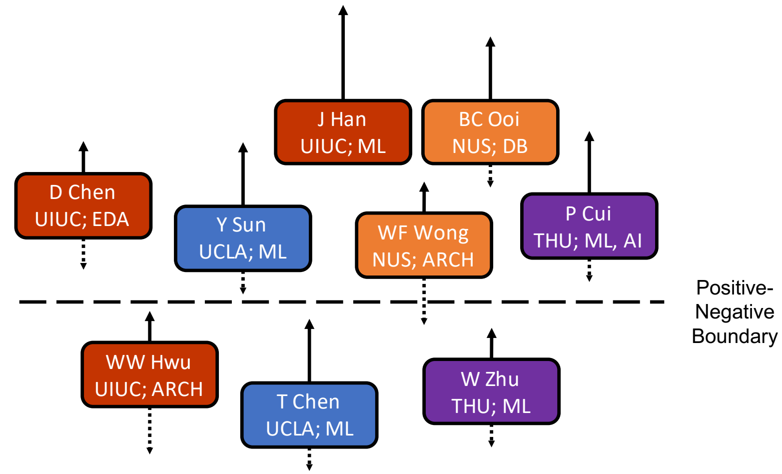
- Build up HNSW index for each vector $(\tilde{x}, \tilde{z})$ in the latent space (time complexity: $O(n \log n)$)
- Perform kNN approximate search for each vector $(\tilde{x}, \tilde{z})$ via HNSW (time complexity: $O(n \log n)$)

$$L_4 = -\sum_{\substack{m=n \\ or \\ u_m,u_n \sim \tilde{p}(m,n)}} \log \sigma(\tilde{\mathbf{x}}_m^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_n) - \sum_{n'=1} \mathbb{E}_{v_{n'} \sim P'_n(v)} \log \sigma(-\tilde{\mathbf{x}}_m{}^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_{n'})$$

Dynamic Positive Sampling



Scholar-Publication Network



Dynamic Positive Sampling

- Build up HNSW index for each vector $(\tilde{x}, \tilde{z})$ in the latent space (time complexity: $O(n \log n)$)
- Perform kNN approximate search for each vector $(\tilde{x}, \tilde{z})$ via HNSW (time complexity: $O(n \log n)$)

$$L_4 = -\sum_{\substack{m=n \\ \text{or} \\ \boxed{u_m, u_n \sim \tilde{p}(m,n)}}} \log \sigma(\tilde{\mathbf{x}}_m^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_n) - \sum_{n'=1} \mathbb{E}_{v_{n'} \sim P'_n(v)} \log \sigma(-\tilde{\mathbf{x}}_m{}^{\mathrm{T}} \cdot \tilde{\mathbf{z}}_{n'})$$

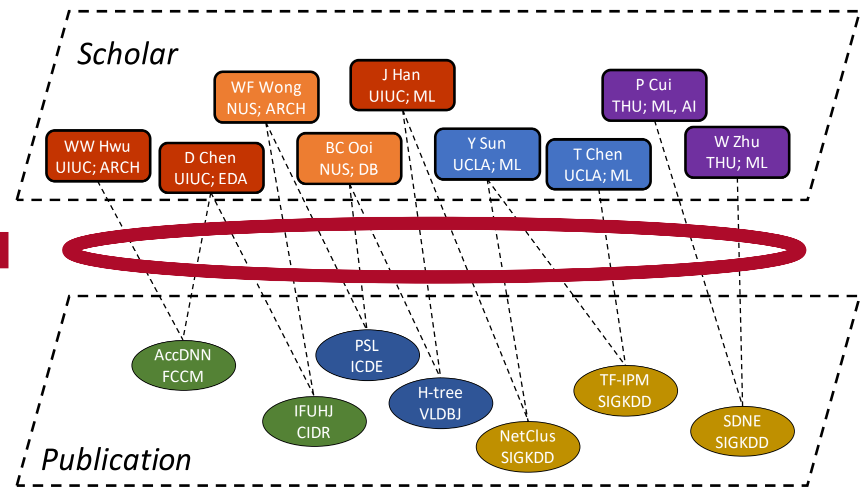
HNSW positive sampling probability distribution

Inter-Partition Proximity Modeling

❑ **Inter-Partition Proximity Modeling**

$$h = [x', \ z']$$

$$L_{Inter} = \sum_{E} log(\sigma(h_u^T \cdot h_v))$$



Inter-Partition Proximity Modeling

❑ **Inter-Partition Proximity Modeling**

$$h = [x',\ z']$$

$$L_{Inter} = \sum_{E} log(\sigma(h_u^T \cdot h_v))$$

❑ **Intra-Partition Proximity Modeling**

$$L_{Intra} = L_2 + L_3 + L_4$$

❑ **Joint Training**

$$L = L_{Intra} + L_{Inter}$$

Experimental Setup

- Tasks:
  - Link Prediction & Node Classification
- Metrics:
  - AUC-ROC, AUC-PR
  - Micro-F1, Macro-F1
- Datasets:

Dataset	#user	#item	#link	#user-attr	#item-attr	sparsity(%)
MovieLens	6,000	3,069	225,344	3	1	0.9878
AMiner	80,461	66,107	168,525	1	1	0.9999
Alibaba	38,140	7,913	59,237	18	28	0.9998

$$sparsity = 1 - \frac{\#link}{\#user \times \#item}$$

Experimental Setup

- Compared Methods:

Homogeneous Network Methods:

- DeepWalk
  [Perozzi et al SIGKDD 2014]
- node2vec
  [Grover et al SIGKDD 2016]
- SDNE
  [Wang et al SIGKDD 2016]

Heterogeneous Network Methods:

- metapath2vec++
  [Dong et al KDD 2017]
- BiNE
  [Gao et al SIGIR 2018]
- NGCF
  [Wang et al SIGIR 2019]

Attributed Network Methods:

- AANE
  [Huang et al SDM 2017]
- ANRL
  [Zhang et al IJCAI 2018]
- FeatWalk
  [Huang et al AAAI 2019]
- STAR-GCN
  [Zhang et al IJCAI 2019]

Efficacy Study

- Link Prediction

Model	MovieLens		AMiner		Alibaba	
	AUC (ROC)	AUC (PR)	AUC (ROC)	AUC (PR)	AUC (ROC)	AUC (PR)
DeepWalk	0.6583	0.6229	0.7730	0.8378	0.8074	0.8353
node2vec	0.6597	0.6296	0.8169	0.8649	0.8605	0.8846
SDNE	0.7454	0.7393	0.5638	0.5646	0.5863	0.6267
metapath2vec++	0.7243	0.6736	0.6935	0.7480	0.8188	0.8346
BiNE	0.7616	0.7297	0.5997	0.5812	0.6886	0.6411
NGCF	0.7547	0.7117	0.7692	0.8290	0.8574	0.8856
ANRL	0.5554	0.5449	0.8350	0.8251	0.6639	0.6429
AANE	0.7010	0.6670	0.5943	0.5924	0.7142	0.6852
FeatWalk	0.7117	0.7007	0.7589	0.8086	0.7948	0.8180
STAR-GCN	0.7621	0.7405	0.6455	0.6587	0.5924	0.5721
BiANE	**0.7711**	**0.7409**	**0.8972**	**0.9054**	**0.8903**	**0.8997**

Efficacy Study

- Node Classification

Model	AMiner				Alibaba			
	60%		80%		60%		80%	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
DeepWalk	0.4427	0.2689	0.4440	0.2736	0.3800	0.1194	0.3866	0.1026
node2vec	0.4587	0.2907	0.4583	0.2898	0.3661	0.0992	0.3879	0.1001
SDNE	0.2833	0.1141	0.2842	0.1132	0.3800	0.0919	0.3765	0.0907
metapath2vec++	0.3926	0.2183	0.3925	0.2199	0.3809	0.1122	0.3860	0.1030
BiNE	0.2648	0.1074	0.2648	0.1067	0.4011	0.0828	0.3999	0.0828
NGCF	0.3417	0.0968	0.3408	0.1094	0.4005	0.0818	0.3986	0.0850
ANRL	0.7772	0.6777	0.7778	0.6779	0.4015	0.0818	0.3992	0.0815
AANE	0.7574	0.6651	0.7550	0.6616	0.3986	0.0912	0.3967	0.0913
FeatWalk	0.3779	0.1977	0.3819	0.2009	0.3759	0.1581	0.3910	0.1554
STAR-GCN	0.2951	0.1278	0.2938	0.1276	0.4008	0.0818	0.3980	0.0814
BiANE	**0.8000**	**0.7137**	**0.7976**	**0.7115**	**0.4078**	**0.1866**	**0.4245**	**0.1795**

Ablation Setup

- BiANE-ATTR: BiANE without structure information
- BiANE-STRUC: BiANE without attribute information
- BiANE-INTER: BiANE with inter-partition proximity modeling only
- BiANE-CONCAT: Integrating attribute and structure encoding by concatenation
- BiANE-LAYER: Integrating attribute and structure encoding by sharing neural layers
- BiANE-IS: BiANE with the sampling distribution $\frac{\exp(\tilde{p}(m,n))}{\sum_{n'} \exp(\tilde{p}(m,n'))}$
- BiANE-ISL: BiANE with the sampling distribution $\frac{\exp(\tilde{p}(m,n))}{\sum_{n'} \exp(\tilde{p}(m,n'))}$ in the latent space

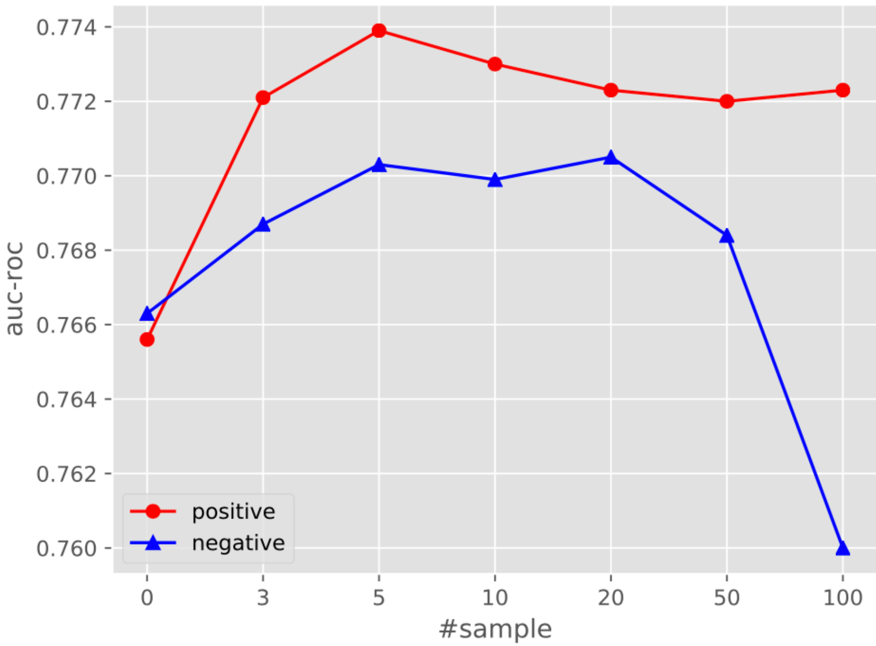
Ablation Study

- Node Classification on AMiner and Alibaba Dataset

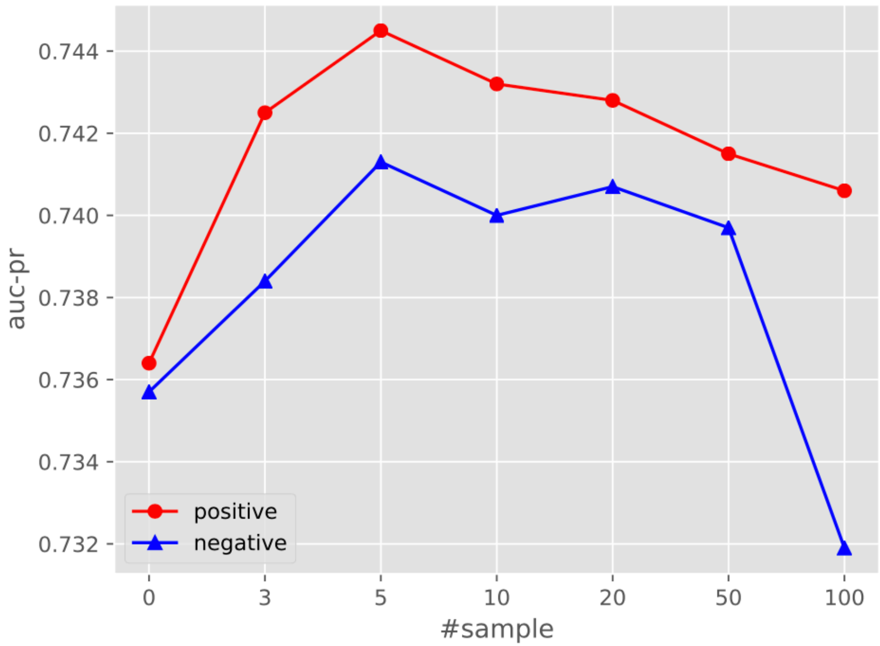
Model	AMiner				Alibaba			
	60%		80%		60%		80%	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BiANE-ATTR	0.7931	0.7089	0.7925	0.7069	0.4062	0.1423	0.4024	0.1327
BiANE-STRUC	0.3818	0.2047	0.3841	0.2077	0.3958	0.0888	0.3961	0.0851
BiANE-INTER	0.7961	0.7083	0.7924	0.7059	0.3977	0.1691	0.4144	0.1673
BiANE-CONCAT	0.7973	0.7063	0.7949	0.7032	0.4065	0.1798	0.4125	0.1646
BiANE-LAYER	0.7967	0.7093	0.7947	0.7051	0.3986	0.1754	0.4087	0.1701
BiANE-IS	0.7970	<u>0.7118</u>	0.7939	<u>0.7075</u>	0.4015	0.1786	<u>0.4201</u>	<u>0.1755</u>
BiANE-ISL	<u>0.7985</u>	0.7079	<u>0.7966</u>	0.7057	**0.4087**	<u>0.1849</u>	0.4131	0.1726
BiANE	**0.8000**	**0.7137**	**0.7976**	**0.7115**	<u>0.4078</u>	**0.1866**	**0.4245**	**0.1795**

Performance w.r.t. #Sample

- Link Predcition on MovieLens Dataset



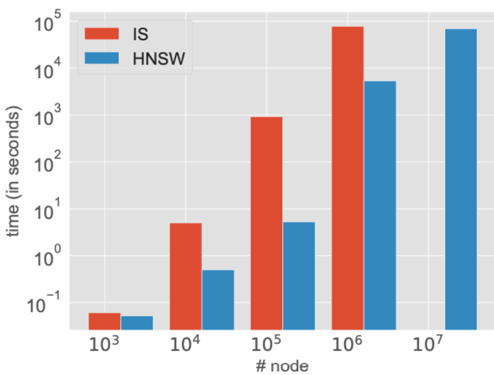
(a) AUC-ROC w.r.t. #Sample



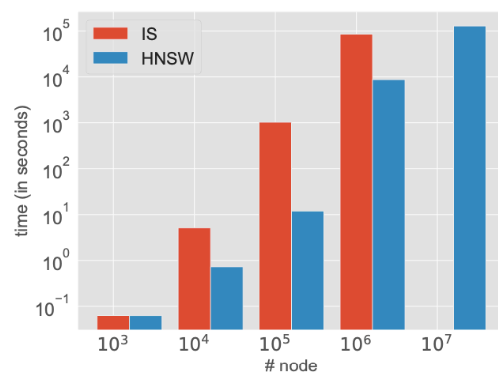
(b) AUC-PR w.r.t. #Sample

Efficiency Study

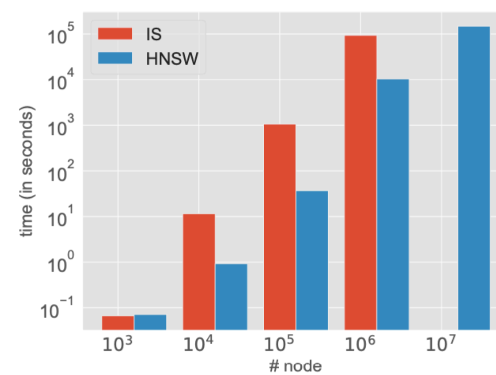
- The Time Cost of a Single Round of Sampling



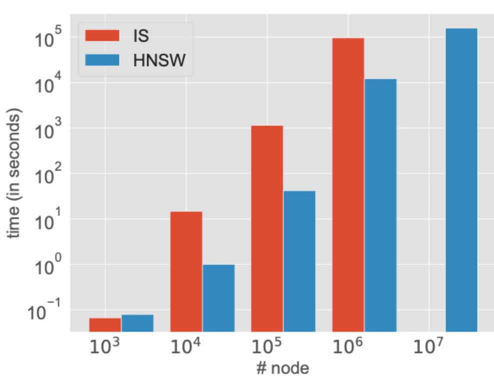
(a) dimension-16



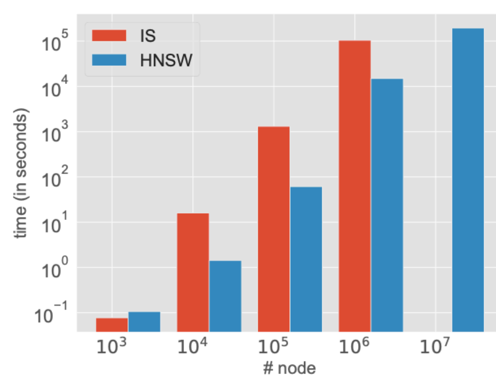
(b) dimension-64



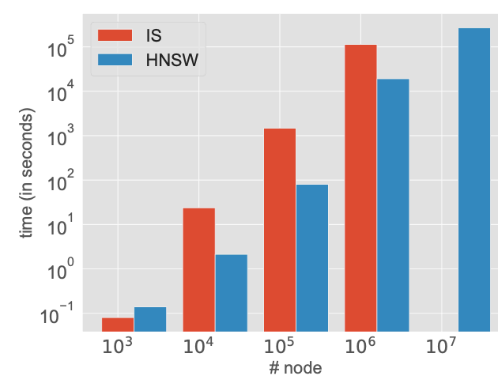
(c) dimension-100



(d) dimension-128



(e) dimension-100



(f) dimension-300

Conclusion & Future Work

- Conclusion
  - Propose a model for embedding bipartite attributed networks, which simultaneously preserves the intra-partition proximity and the inter-partition proximity
  - Introduce a dynamic positive sampling strategy to ameliorate the representation learning process without loss of model scalability.

- Future Work
  - Reduce the space complexity for representation learning model.
  - Extend the current work to model dynamic bipartite attributed networks.

THANK YOU FOR YOUR ATTENTION!

Q&A

中國人民大學
RENMIN UNIVERSITY OF CHINA

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Alibaba Group
阿里巴巴集团