HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning

Tao-yang Fu
The Pennsylvania State University
University Park, PA 16802, USA
txf225@cse.psu.edu

Wang-Chien Lee
The Pennsylvania State University
University Park, PA 16802, USA
wlee@cse.psu.edu

Zhen Lei
The Pennsylvania State University
University Park, PA 16802, USA
zlei@psu.edu

CIKM'17, November 6–10, 2017, Singapore.

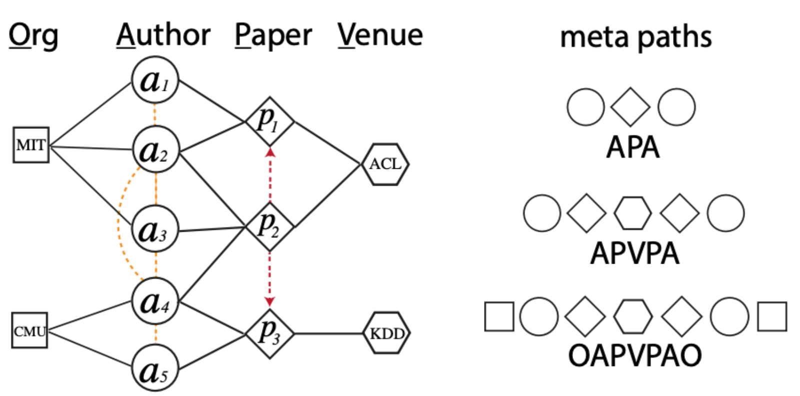
ISBN 978-1-4503-4918-5/17/11...$15.00
DOI: https://doi.org/10.1145/3132847.3132953

Definition

DEFINITION 1. *Information Network.* *An information network is a directed graph $G = (V, E, \Phi, \Psi)$, where V is the set of nodes; $E \subseteq V \times V$ is the set of edges in V. $\Phi : V \rightarrow A$ and $\Psi : E \rightarrow R$ are type mapping functions for nodes and edges, respectively. Here each node $v \in V$ is mapped to one particular node type in A, i.e., $\Phi(v) \in A$, and each link $e \in E$ belongs to a particular edge type in R, i.e., $\Psi(e) \in R$. When $|A| > 1$ or $|R| > 1$, the network is called a heterogeneous information network (HIN); otherwise, it is a homogeneous information network.*

DEFINITION 2. *Meta-path.* *Given a heterogeneous information network $G = (V, E, \Phi, \Psi)$, a meta-path π is a sequence of node types $a_1, a_2, ..., a_n$ and/or edge types $r_1, r_2, ..., r_{n-1}$:*

$$\pi = a_1 \xrightarrow{r_1} ...a_i \xrightarrow{r_i} ... \xrightarrow{r_{n-1}} a_n$$



(a) An academic network

Framework

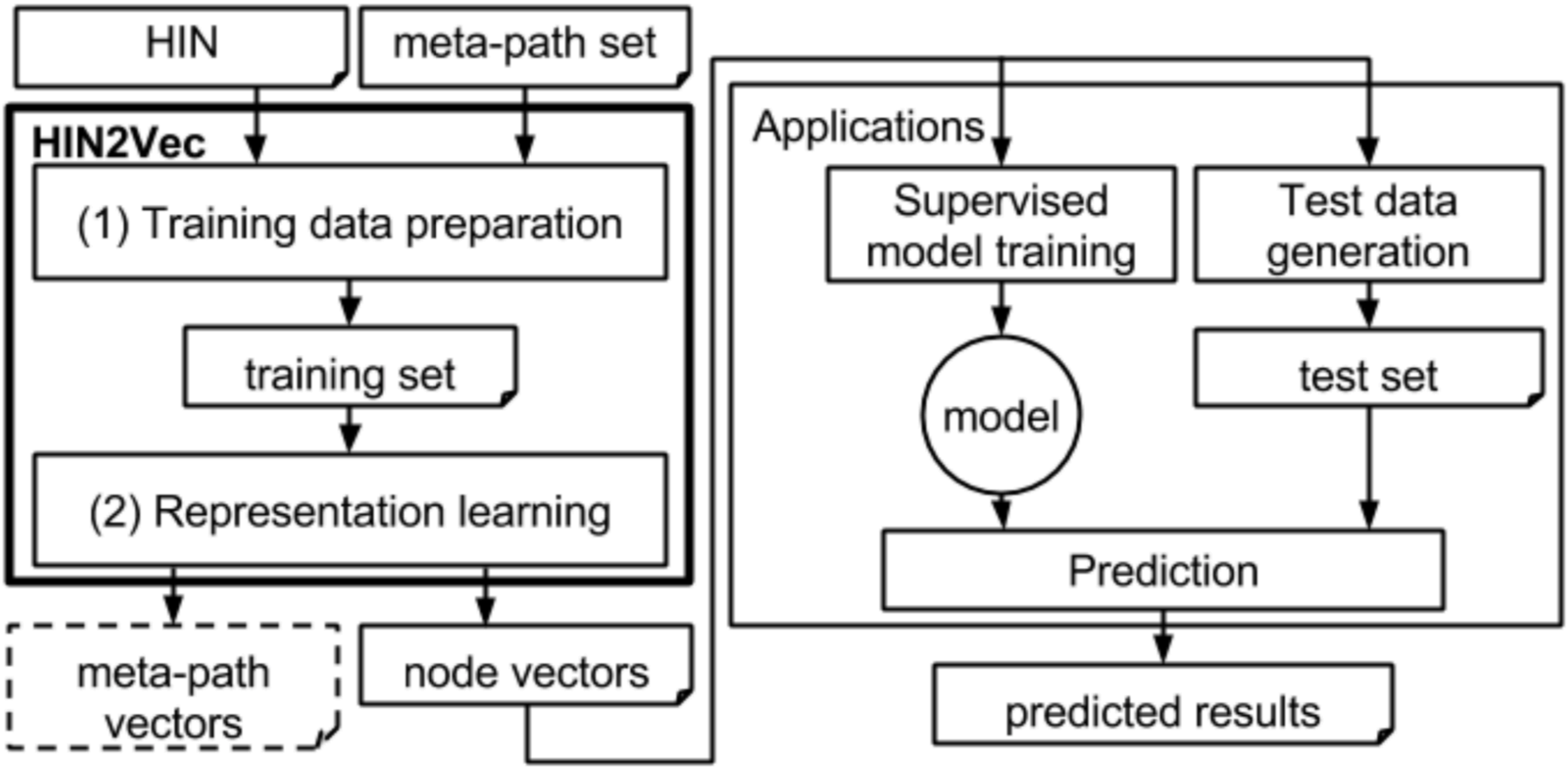


Figure 1: Overview of the HIN2Vec framework

Representation Learning

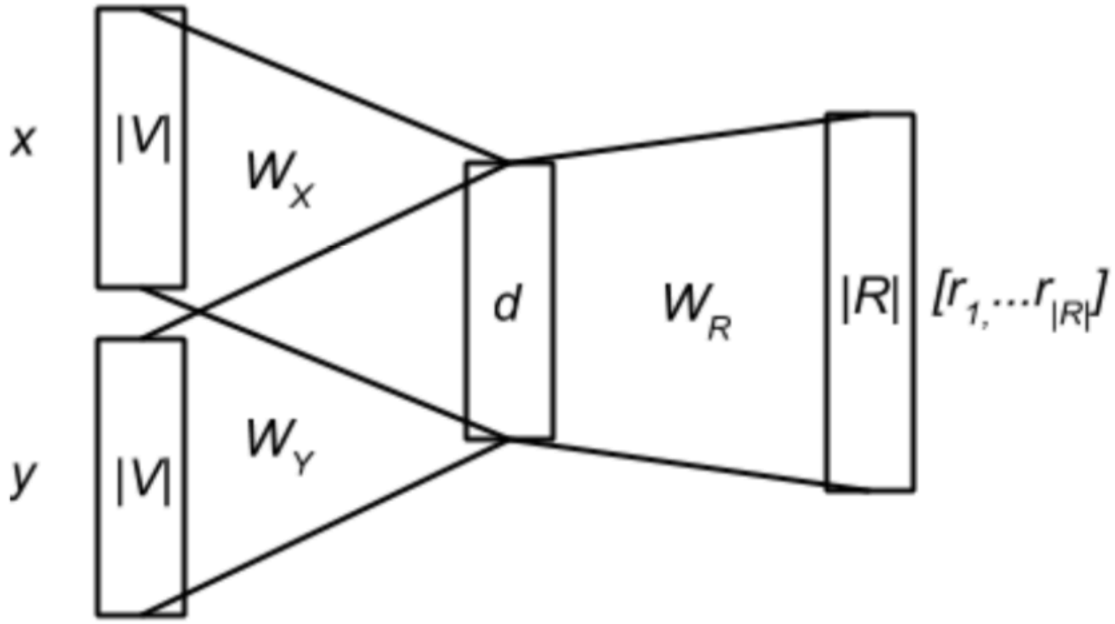


Figure 2: A conceptual model for HIN2Vec

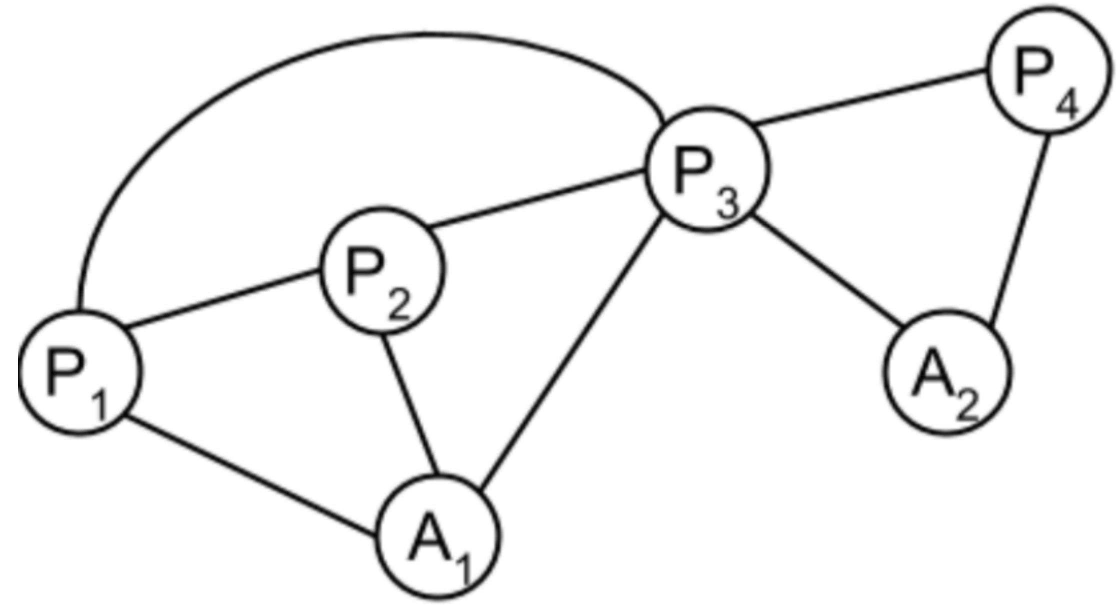


Figure 3: A paper-author HIN

$$R = \{P-P, P-A, A-P, P-P-A, P-P-P, P-A-P, A-P-P, A-P-A\}$$

Multi-label classification problem:

predict the probabilitiesP(r_i |x ,y)(i = 1..|R|)

For instance, P_1 and A_1 have two relationships, P-A and P-P-A. A training data entry is $\langle x : P_1, y : A_1, output : [0,1,0,0,1,0,0,0]\rangle$.

Disadvantages

- **For each pair of nodes, traverse the entire network as they prepare training data to find all possible relationship types R.**

- **W_X,W_Y and W_R, need to be updated in accordance with all the relationships**

HIN2vec

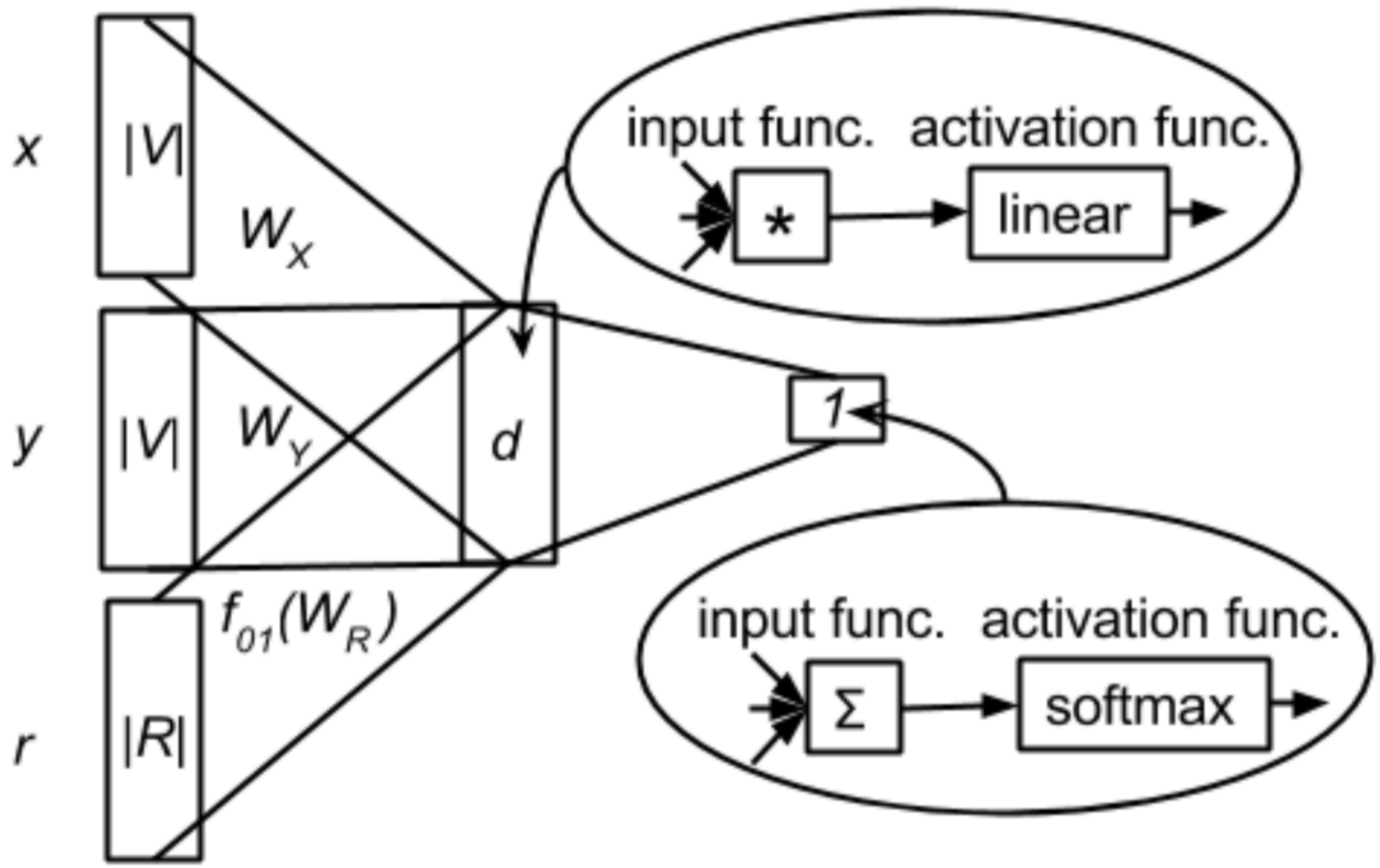


Figure 4: The HIN2Vec NN model

whether two nodes, x and y, have a specific relationship r.

Input layer:

1. The relationship type is no longer a prediction object, but appears as an input in the input layer.
2. The transformation matrix W_R of the relationship has a regularization function

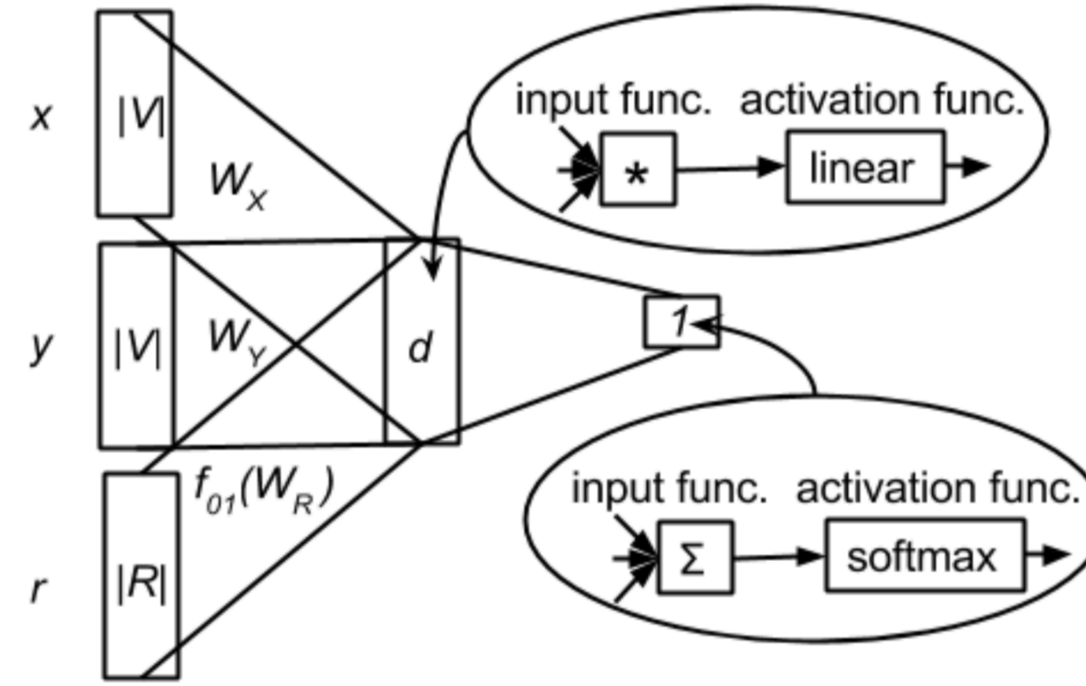


Figure 4: The HIN2Vec NN model

Hidden layer:

1. Input is $W_x^T\vec{x}, W_y^T\vec{y}, f_{01}(W_R^T\vec{r})$
2. Output is $W_x^T\vec{x} \odot W_y^T\vec{y} \odot f_{01}(W_R^T\vec{r})$
3. Hadamard, mean, difference, and absolute value of the difference are used respectively. Finally, the Hadamard function works best.

Output layer:

1. Input is $\sum W_x^T\vec{x} \odot W_y^T\vec{y} \odot f_{01}(W_R^T\vec{r})$
2. The activation function is sigmoid function.

Regularization of W_R

Binary Step function outperforms Sigmoid.

Optimizations

Training data form

$$< x, y, r, L(x, y, r) >$$

The objective function O

$$\mathcal{O} \propto log\mathcal{O} = \sum_{x,y,r \in D} log\mathcal{O}_{x,y,r}(x, y, r)$$

$$\mathcal{O}_{x,y,r}(x, y, r) = \begin{cases} P(r|x, y), & if L(x, y, r) = 1 \\ 1 - P(r|x, y), & if L(x, y, r) = 0 \end{cases}$$

$$log\mathcal{O}_{x,y,r}(x, y, r) = L(x, y, r)logP(r|x, y) + [1 - L(x, y, r)]log[1 - P(r|x, y)]$$

Whether there is a relationship r between nodes x and y

$$P(r|x, y) = sigmoid(\sum W_x^T \vec{x} \odot W_y^T \vec{y} \odot f_{01}(W_R^T \vec{R}))$$

Training Data Preparation

Remove cycles during random walk

$$(A_1, A_1, A - P - A)$$

Sample negative data

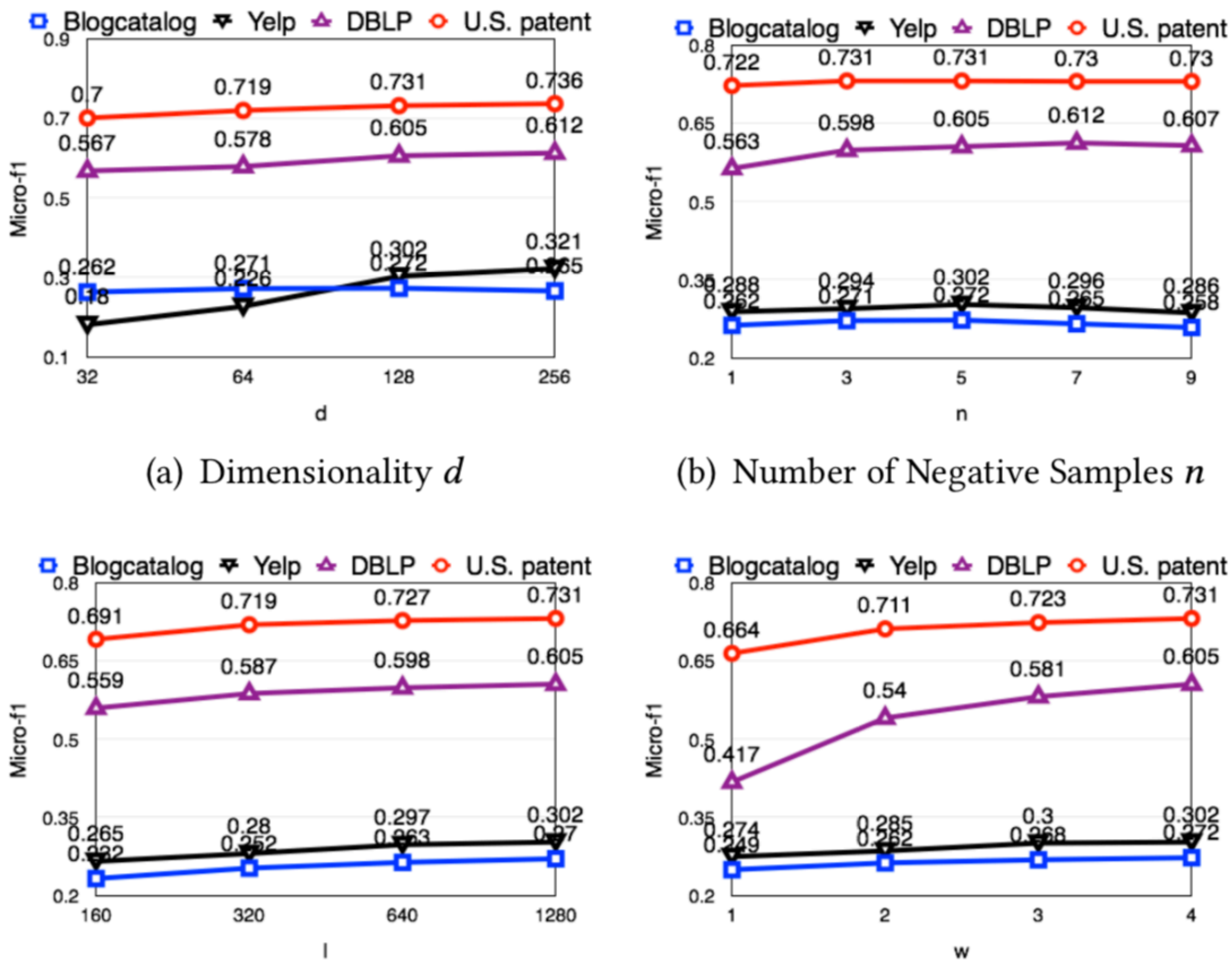
For each sampled positive entry, $\langle x, y, r \rangle$,

replacing one of the three values

negative data entry, $\langle x'', y'', r'' \rangle$

x'' and y'' are not expected to have a certain relationship r''.

EXPERIMENTS



(a) Dimensionality d

(b) Number of Negative Samples n

(c) Length of Random Walks l

(d) Length of Meta-paths w

Figure 5: Parameter Tuning

how the micro-f1 changes in node classification in the networks.

Table 2: Performance Evaluation of Node Classification

	Blogcatalog		Yelp		DBLP		U.S. Patents	
	micro-f1	macro-f1	micro-f1	macro-f1	micro-f1	macro-f1	micro-f1	macro-f1
DeepWalk	0.244	0.140	0.276	0.165	0.481	0.463	0.675	0.676
LINE	0.239	0.128	0.270	0.163	0.449	0.429	0.66	0.663
node2vec	0.246	0.141	0.276	0.166	0.491	0.470	0.676	0.677
PTE	0.179	0.096	0.222	0.130	0.417	0.394	0.547	0.555
HINE	*0.250	*0.144	*0.278	*0.169	0.475	0.461	*0.681	*0.685
ESim	0.207	0.102	0.229	0.132	*0.514	*0.496	0.610	0.562
HIN2Vec	**0.272(9.9%)**	**0.158(11.3%)**	**0.302(7.9%)**	**0.192(12.0%)**	**0.605(23.8%)**	**0.594(20.1%)**	**0.729(6.6%)**	**0.732(6.4%)**

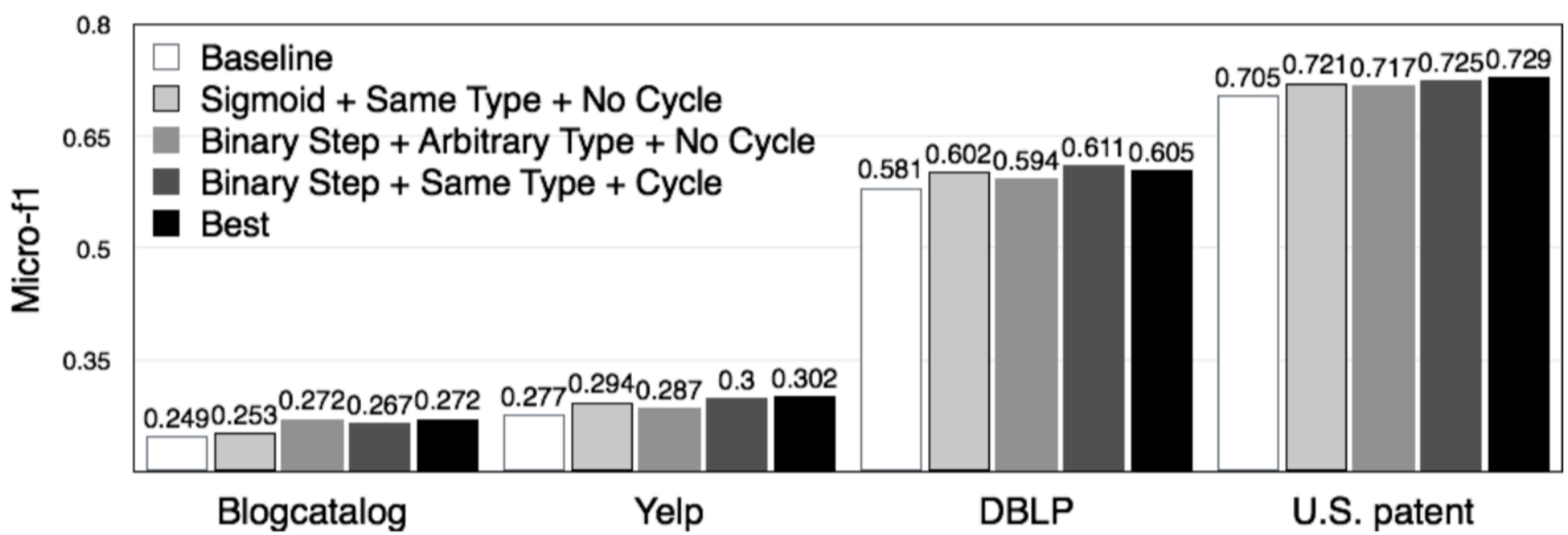


Figure 6: Comparison of approaches to issues in HIN2Vec

Table 4: Vector Functions of Node Pairs

Functions	Hadamard	Average	Minus	Abs. Minus
Description	$\vec{v_1}_i * \vec{v_2}_i$	$\frac{\vec{v_1}_i + \vec{v_2}_i}{2}$	$\vec{v_1}_i - \vec{v_2}_i$	$\lvert\vec{v_1}_i - \vec{v_2}_i\rvert$

Table 5: Performance Evaluation of Vector Functions

	Blogcatalog		Yelp		DBLP		U.S. Patents	
	MAP	recall@100	MAP	recall@100	MAP	recall@100	MAP	recall@100
Hadamard	**0.141**	**0.279**	**0.028**	**0.138**	**0.265**	**0.751**	**0.176**	**0.602**
Average	0.074	0.245	0.004	0.033	0.005	0.124	0.008	0.063
Minus	0.050	0.171	0.004	0.030	0.004	0.114	0.009	0.059
Abs. minus	0.130	0.238	0.023	0.119	0.249	0.750	0.130	0.540

Table 6: Performance Evaluation of Link Prediction

	Blogcatalog		Yelp		DBLP		U.S. Patents	
	MAP	recall@100	MAP	recall@100	MAP	recall@100	MAP	recall@100
DeepWalk	0.124	0.227	*0.021	0.110	0.230	*0.710	0.093	0.500
LINE	*0.134	*0.249	0.017	0.104	0.086	0.580	0.091	0.400
node2vec	0.125	0.229	*0.021	*0.111	*0.231	*0.710	0.095	*0.503
PTE	0.067	0.139	0.004	0.034	0.071	0.324	0.030	0.243
HINE	0.085	0.179	0.016	0.097	0.205	0.697	*0.103	0.495
ESim	0.132	0.185	x	x	0.179	0.633	x	x
MPE	**0.141(5.0%)**	**0.279(10.8%)**	**0.028(31.8%)**	**0.138(24.3%)**	**0.265(12.8%)**	**0.751(5.8%)**	**0.176(70.8%)**	**0.602(19.9%)**