Meta Structure: Computing Relevance in Large Heterogeneous Information Networks

Zhipeng Huang

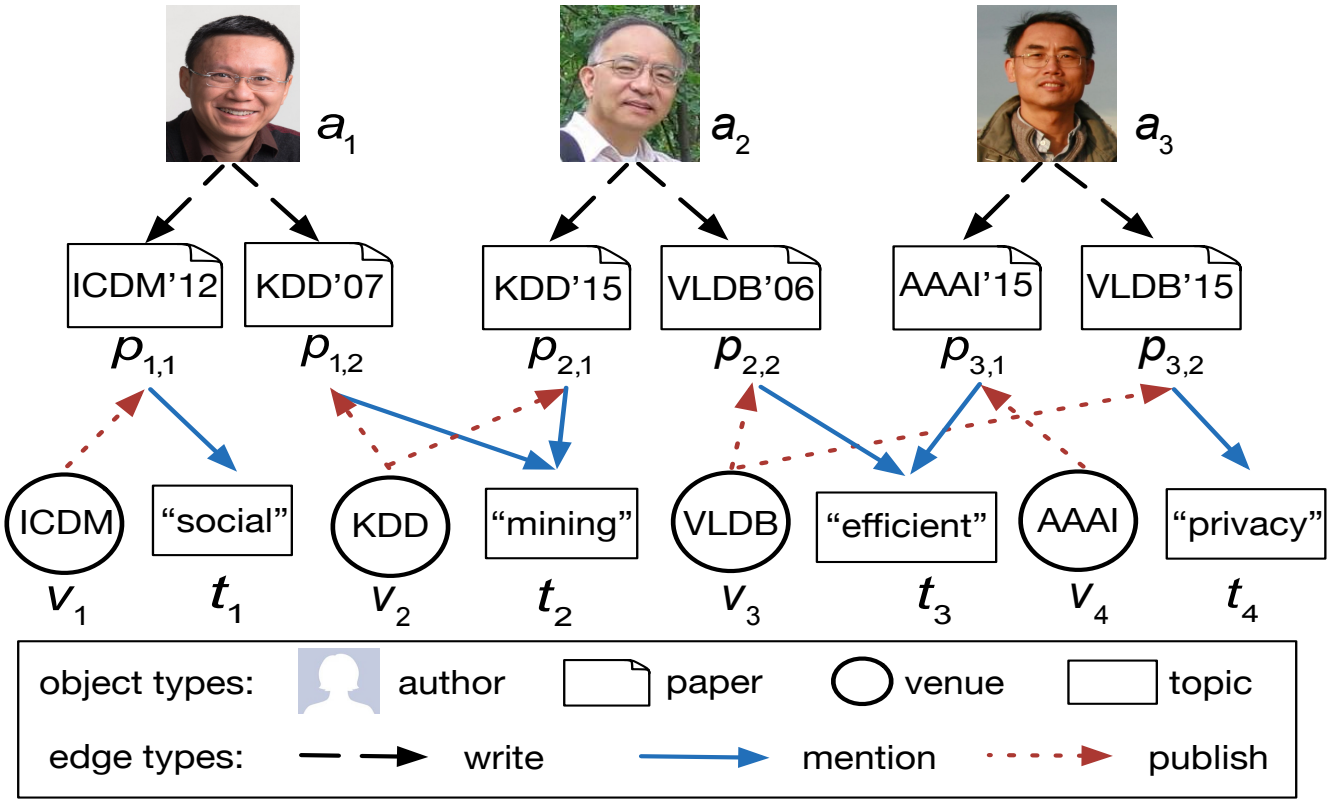
http://i.cs.hku.hk/~zphuang/

Introduction

- Computing relevance on networks (e.g., social network, co-author network) supports many applications:
  - similarity search
  - recommendation
- Many measures have been studied:
  - Jaccard coefficient, common neighbors, shortest path
  - Page Rank, Personalize Page Rank, SimRank, etc.

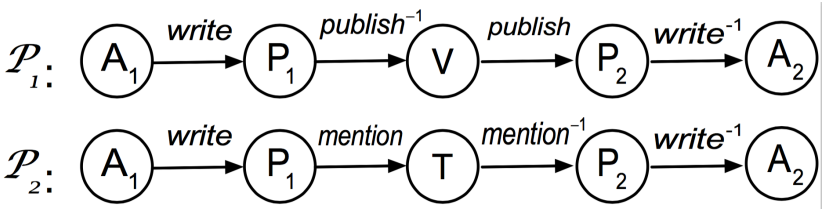
Heterogeneous Information Network

- HIN: Directed graph with multiple node types and edge types.



Meta Path-Based Relevance Measures

- *Meta Path:* a sequence of node and edge types.



- Measures*: PathCount* [1], *PathSi*m [1] and *PCRW* [2]
- *Source*: Automatically generate meta path(WWW'15)
- **Limitation:** Fail to discover common nodes.
  - Example: A researcher wants to search for some authors who have published papers in the same venue ***and*** in the same topic with his papers.

Linear Combination

- $R(a_1,a_2)$

$= R(a_1,a_2 \mid P_1)+R(a_1,a_2 \mid P_2)$

$= 1+1$

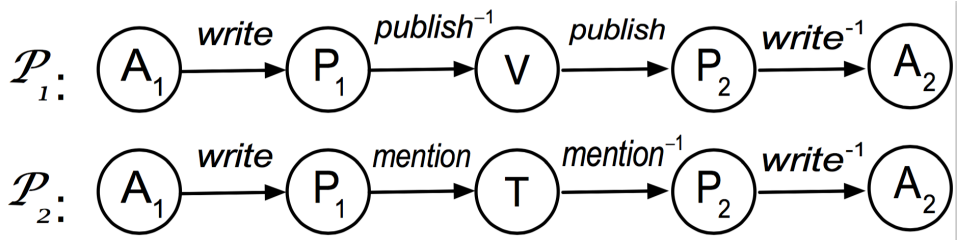
$= 2$

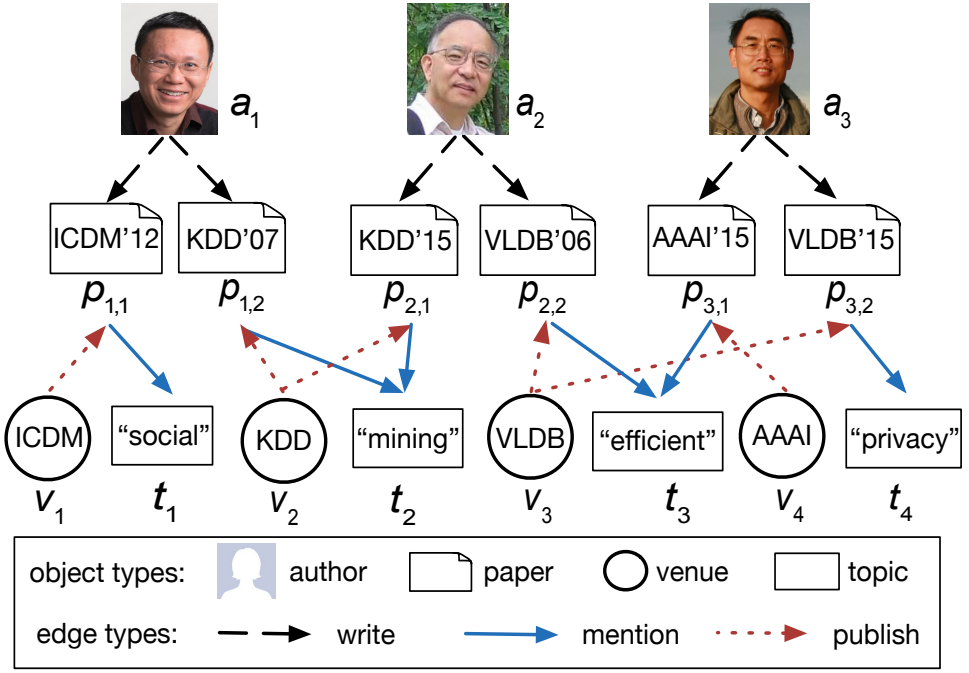
- $R(a_2,a_3)$

$= R(a_2,a_3 \mid P_1)+R(a_2,a_3 \mid P_2)$

$= 1+1$

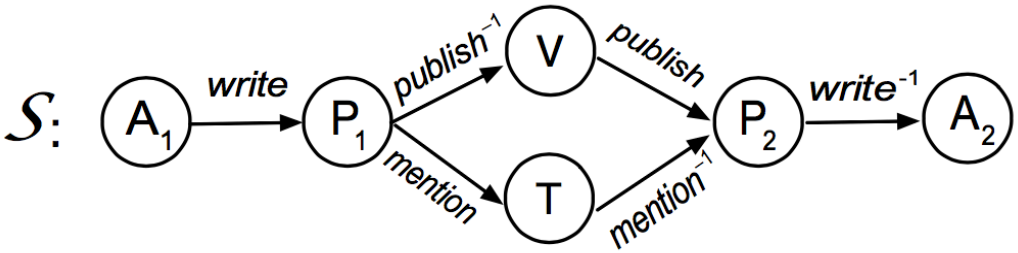
$= 2$





Meta Structure

- A powerful extension of meta path, a directed acyclic graph (DAG).



- More Powerful.
  - Contain more information than a meta path. Can express more semantic meaning.
- Challenges:
  - How to define measures based on meta structure?
  - More complex leads to high computational cost.
  - How to derive a meta structure? (Not yet studied well)

Relevance Measures

- StructCount: extension of PathCount

$$StructCount(x_0, y_0 \mid S) = \left| GraphIns(x_0, y_0 \mid S) \right|$$

- Structure Constrained Random Walk

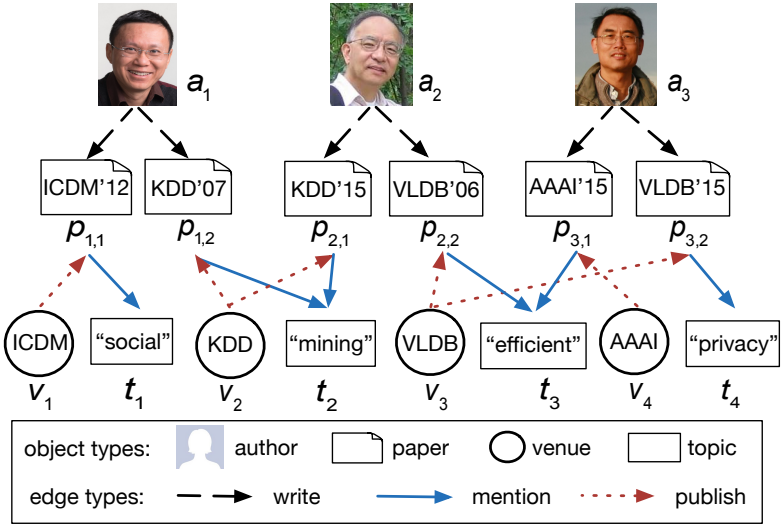
$$SCSE(g, i \mid \mathcal{S}, o_t) = \frac{\sum_{g' \in \sigma(g, i \mid \mathcal{S}, G)} SCSE(g', i+1 \mid \mathcal{S}, o_t)}{\mid \sigma(g, i \mid \mathcal{S}, G) \mid},$$

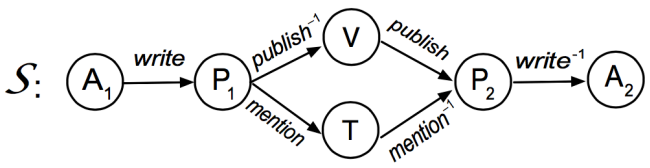
- Biased Structure Constrained Random Walk, a combination of the previous two measures.

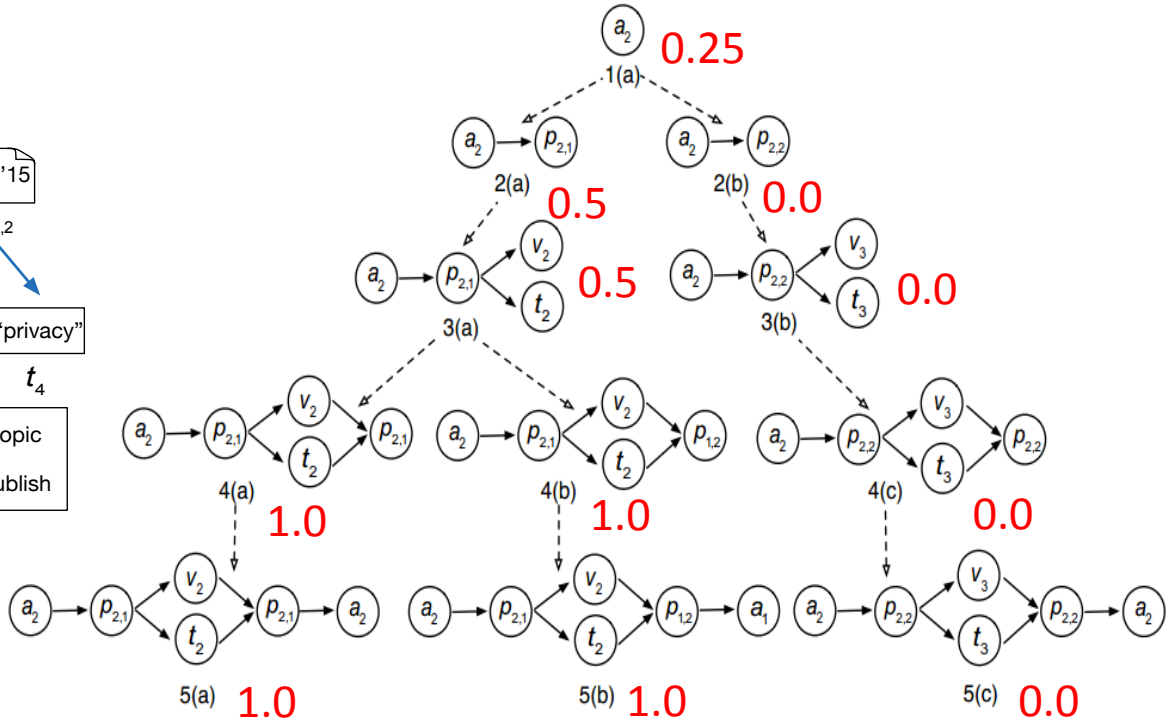
$$BSCSE(g, i \mid \mathcal{S}, o_t) = \frac{\sum_{g' \in \sigma(g, i \mid \mathcal{S}, G)} BSCSE(g', i+1 \mid \mathcal{S}, o_t)}{\mid \sigma(g, i \mid \mathcal{S}, G) \mid^{\alpha}},$$

Recursive Tree

- To calculate the BSCSE relevance of a_2 and a_1:

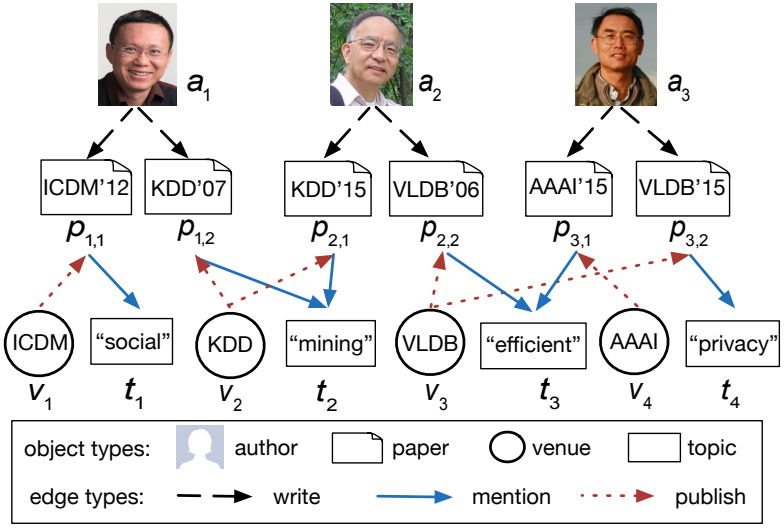


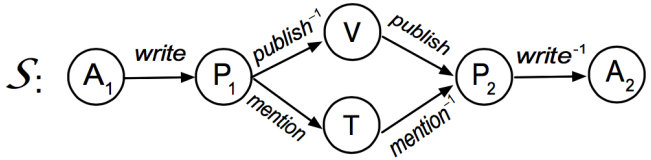




i-LTable

- Index the probability distribution starting from the i-th level of a meta structure.

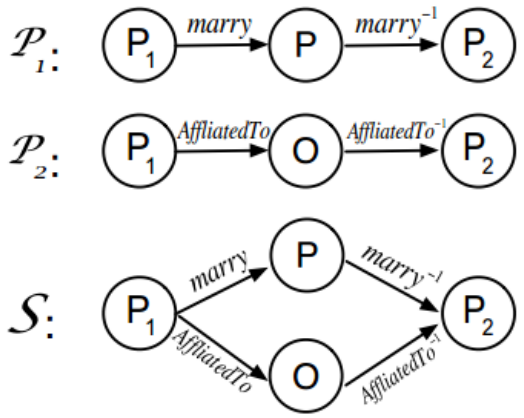




key	*value*
$<v_1, t_1>$	$<a_1, 1.0>$
$<v_2, t_2>$	$<a_1, 0.5>$
	$<a_2, 0.5>$
$<v_3, t_3>$	$<a_2, 1.0>$
$<v_3, t_4>$	$<a_3, 1.0>$
$<v_4, t_3>$	$<a_3, 1.0>$

Experiment: Entity Resolution

- To find duplicated entities in YAGO
  - *Barack_Obama* and *Presidency_Of_Barack_Obama*
- Metric: AUC

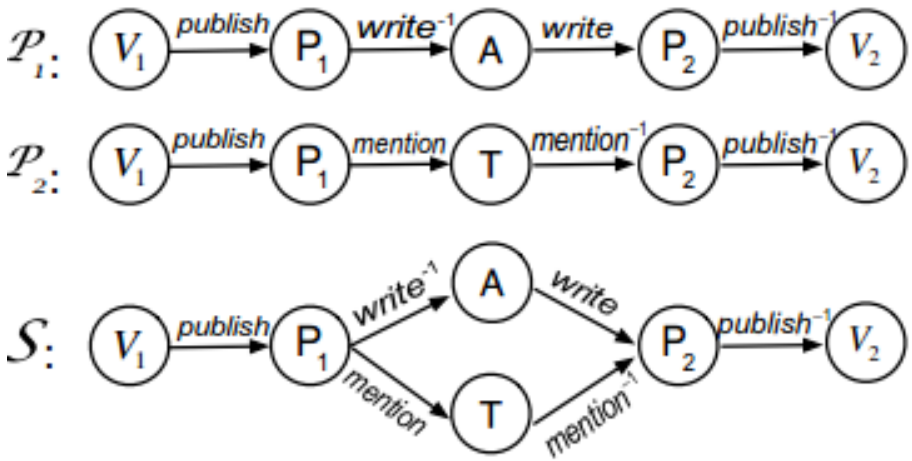


	P1			P2		
Measure	PathCount	PCRW	PathSim	PathCount	PCRW	PathSim
AUC	0.1324	0.0120	0.0097	0.0003	0.0014	0.0002
	Linear Combination(optimal)			Meta Structure S		
Measure	PathCount	PCRW	PathSim	SC	SCSE	BSCSE*
AUC	0.2898	0.2606	0.2920	0.5556	0.5640	**0.5640**

Relevance Ranking

- We label the relevance of venues in DBLP_4_Area.
- 0 for not relevant, 1 for relevant and 2 for strongly relevant.
- Consider both **scope** and **level** of the venues. (like SIGMOD and VLDB are 2)
- Normalized Discounted Cumulative Gain (NDCG)

Relevance Ranking



	P1			P2		
Measure	PathCount	PCRW	PathSim	PathCount	PCRW	PathSim
nDCG	0.9004	0.9047	0.9083	0.8224	0.8901	0.8834
	Linear Combination(optimal)			Meta Structure S		
Measure	PathCount	PCRW	PathSim	SC	SCSE	BSCSE*
nDCG	0.9004	0.9100	0.9083	0.9056	0.9104	0.9130

Reference

- [1] Sun Yizhou, et al. “Pathsim: Meta path-based top-k similarity search in heterogeneous information networks." VLDB’11 (2011).
- ［2] Lao, Ni, and William W. Cohen. "Relational retrieval using a combination of path-constrained random walks." Machine learning 81.1.010): 53-67
- [3] Meng, Changping, et al. "Discovering meta-paths in large heterogeneous information networks." **WWW’15**.
- [4] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, Xiang Li, “Meta Structure: Computing Relevance in Large Heterogeneous Information Networks”, **SIGKDD’ 16**