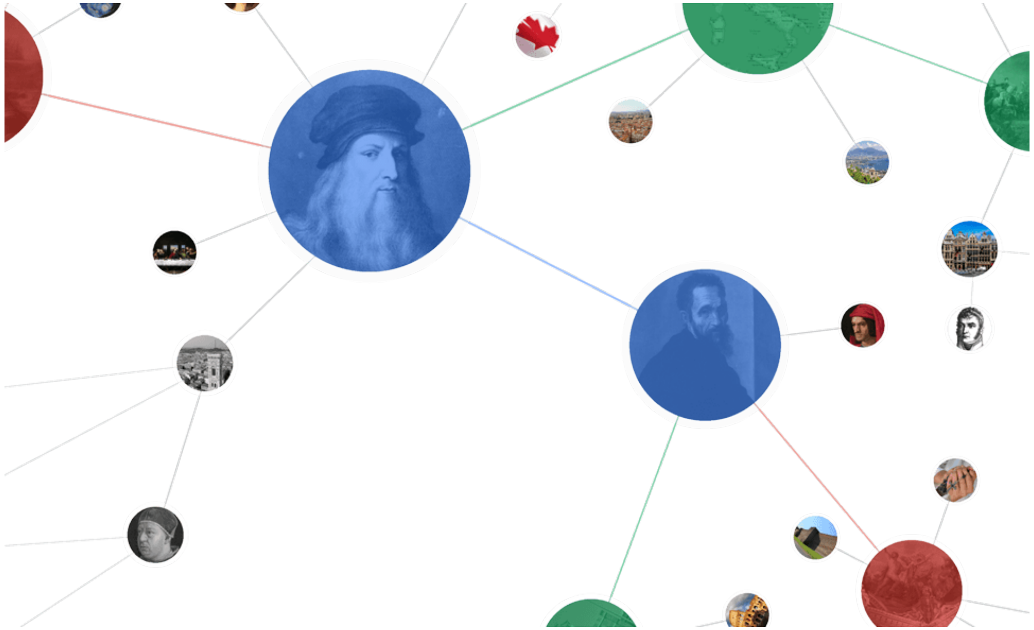
APWeb-WAIM Joint Conference on Web and Big Data 2017

# Meta Paths and Meta Structures: Analysing Large Heterogeneous Information Networks

**Reynold Cheng**
**Department of Computer Science**
**University of Hong Kong**
**http://www.cs.hku.hk/~ckcheng/**
**ckcheng@cs.hku.hk**

# Knowledge Graphs



# Social Networking Websites


https://makeawebsitehub.com/social-media-sites/

# Biological Network

http://serious-science.org/controlling-noisy-dynamics-in-biological-networks-to-fight-cancer-5376

# Research Collaboration Network

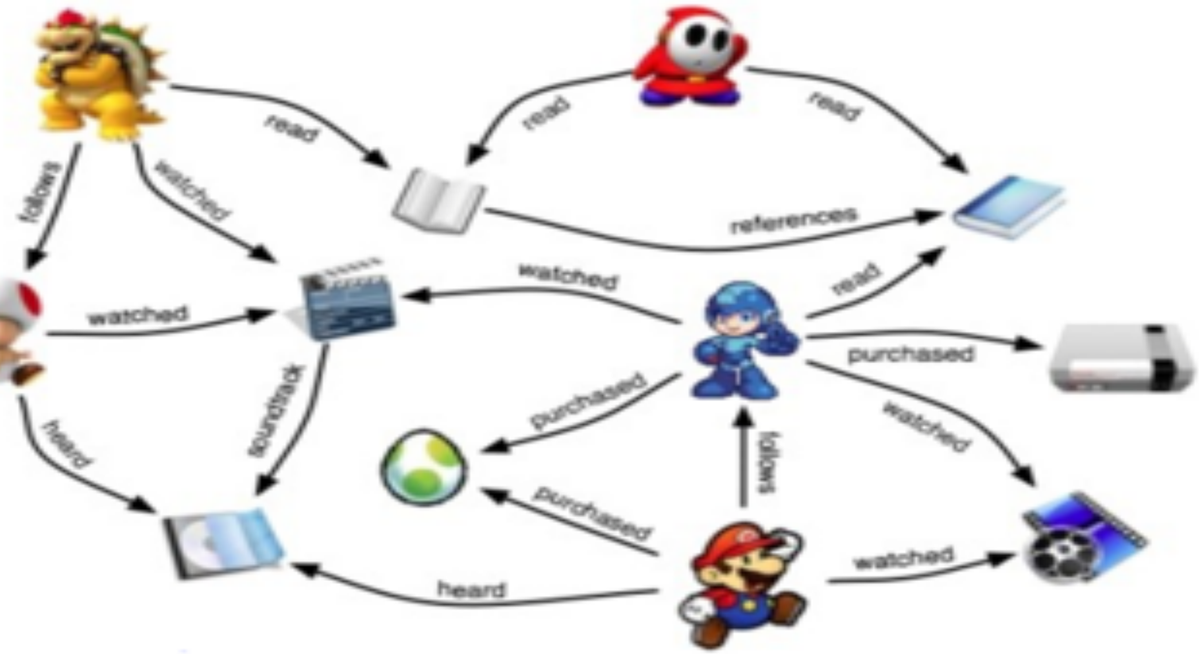
dblp
computer science bibliography

https://scholarlykitchen.sspnet.org/2017/04/07/updated-figures-scale-nature-researchers-use-scholarly-collaboration-networks/

# Product Recommendation Network


http://www.sciencedirect.com/science/article/pii/S0957417413006921
Byunghak Leem. Heuiju Chun. An impact of online recommendation network on demand

# Heterogeneous Information Network (HIN)





Yangqiu Song. Recent Development of Heterogeneous Information Networks: From Meta-paths to Meta-graphs

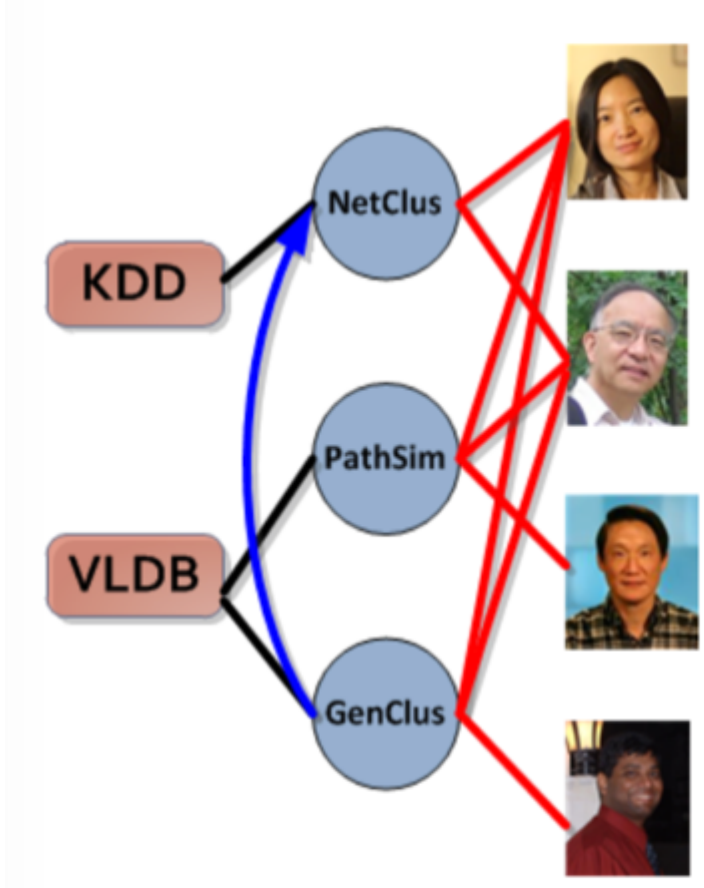
# HINs are Ubiquitous !

- Healthcare
  - Doctor, Patient, Disease
- Source Code Repository
  - Project, Developer, Repository
- E-Commerce
  - Seller, Buyer, Product
- News
  - Author, Organization


Jiawei Han. A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks.

# Example HINs

- DBLP Bibliographic Network
- Node (Type):
  - KDD (Venue)
  - Jiawei Han (Author)
- Link (Type):
  - Write (Author → Paper)
  - Publish (Paper → Venue)

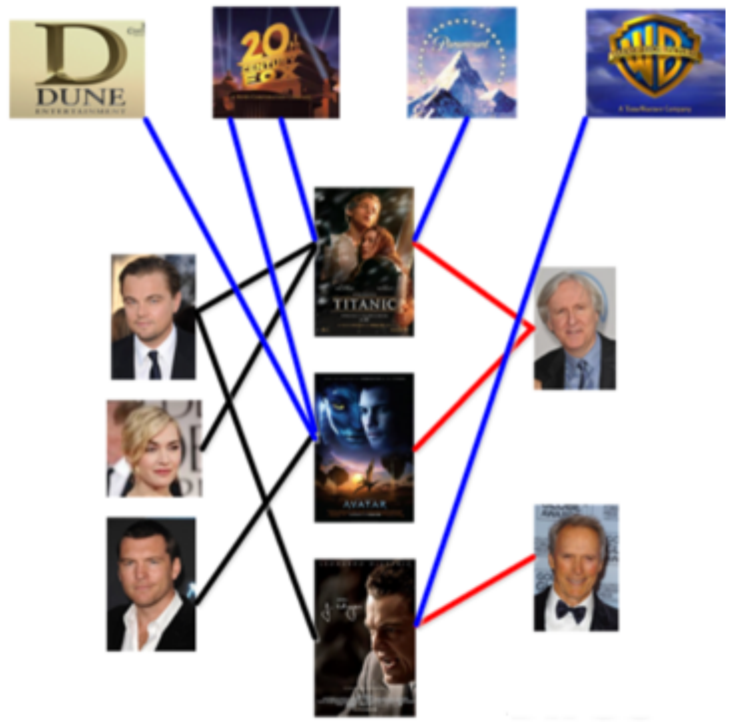



Jiawei Han. A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks.

# Example HINs

- The IMDB Movie Network

- Node (Type):
  - Forrest Gump (Movie)
  - Tom Cruise (Actor)
- Link (Type):
  - Make (Producer → Movie)
  - Act (Author → Movie)




Jiawei Han. A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks.

# Example HINs

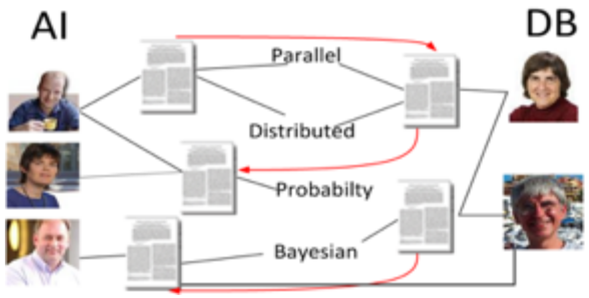
- The Facebook Network

- Node (Type):
  - Jimmy (User)
  - Coca Cola (Product)
- Link (Type):
  - Like (User → Product
  - Follow (User → User)


Jiawei Han. A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks.

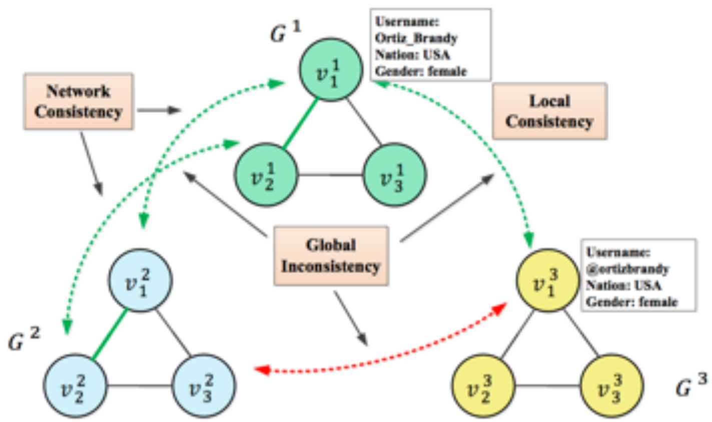
# HIN Applications

- Link Prediction




- Entity Profiling


- Data Integration



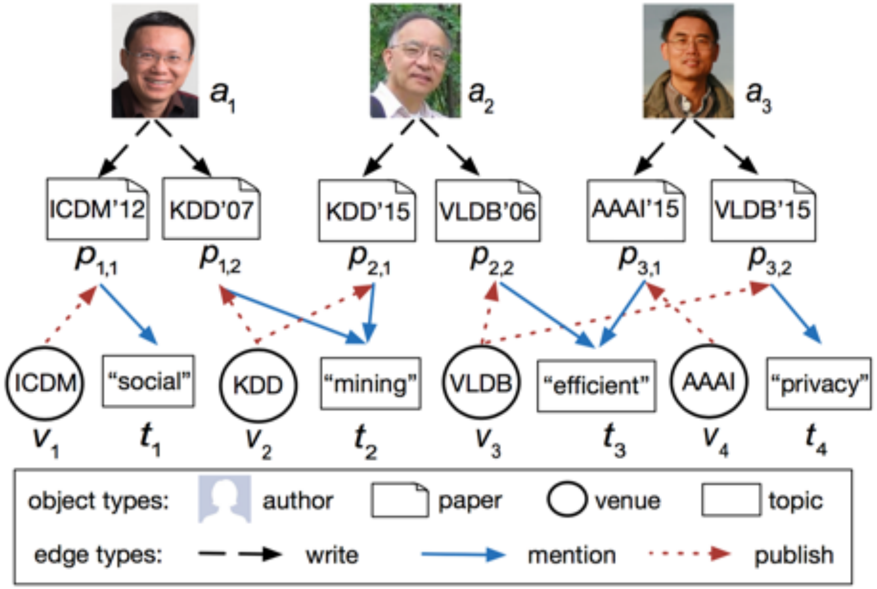

Yangqiu Song. Recent Development of Heterogeneous Information Networks: From Meta-paths to Meta-graphs
Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S. Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency, KDD 2015.

# Overview of the Tutorial

## Relevance Search

Find **Similar/Relevant** Objects in Networks

## Examples




$$A \xrightarrow{\text{writing}} P \xrightarrow{\text{written-by}} A$$

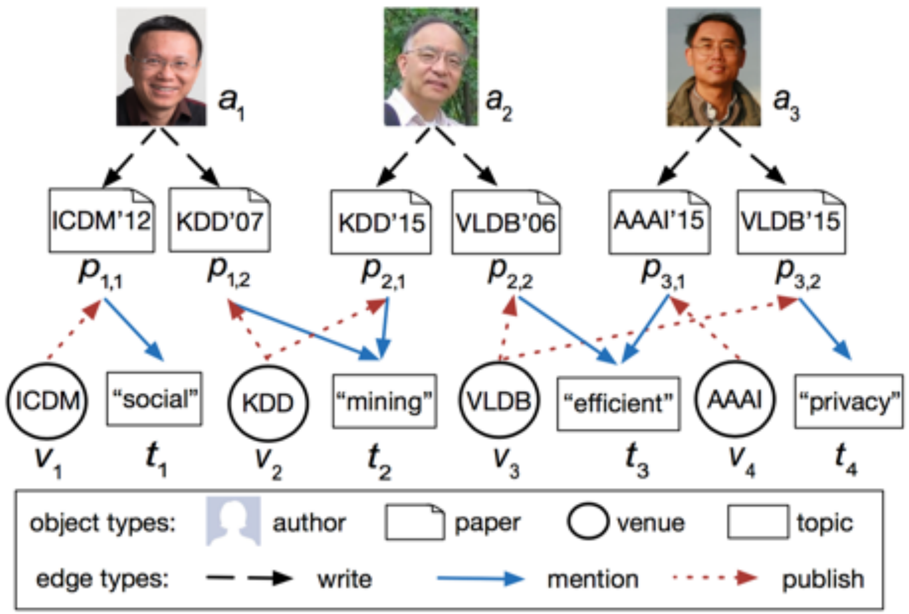
## DBLP[1]

- **Who** are most similar to *Jiawei Han* ?
- **Whose** recent publication is relevant with *Jiawei Han's research* ?

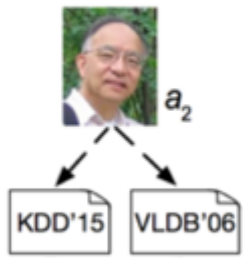
[1] http://dblp.uni-trier.de/

# Overview of the Tutorial

- Where do relations (meta-path) come from?
  - Provided by experts [Sun VLDB'11]
    - Not easy for a complex schema!




$$A \xrightarrow{\text{writing}} P \xrightarrow{\text{written-by}} A$$




Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, and Wangda Zhang. "Discovering Meta-Paths in Large Heterogeneous Information Networks", in WWW 2015.

# Overview of the Tutorial

- Query Recommendation: to suggest alternate relevant queries to a search engine user

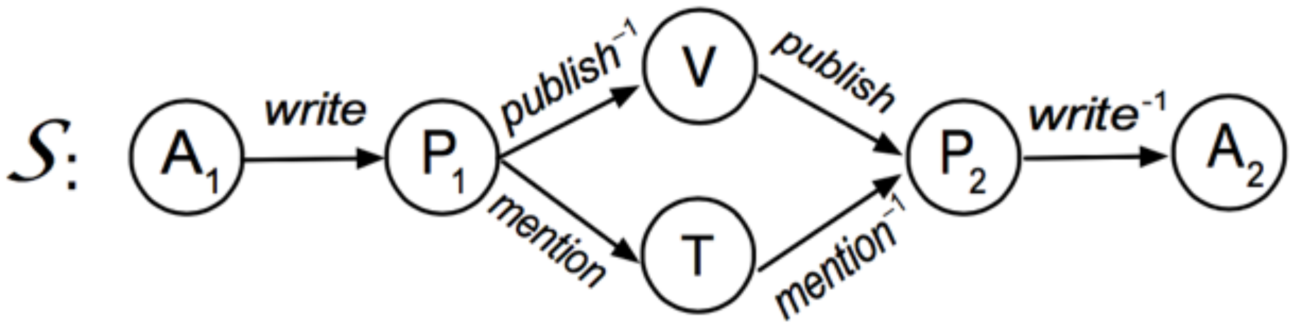
- How will HIN benefit query recommendation ?



Zhipeng Huang, Bogdan Cautis, Reynold Cheng, Yudian Zheng. KB-Enabled Query Recommendation for Long-Tail Queries. CIKM 2016.

# Overview of the Tutorial

- How can we express using more complex structure?




- More Expressive (i.e., contain more information) than a meta path.

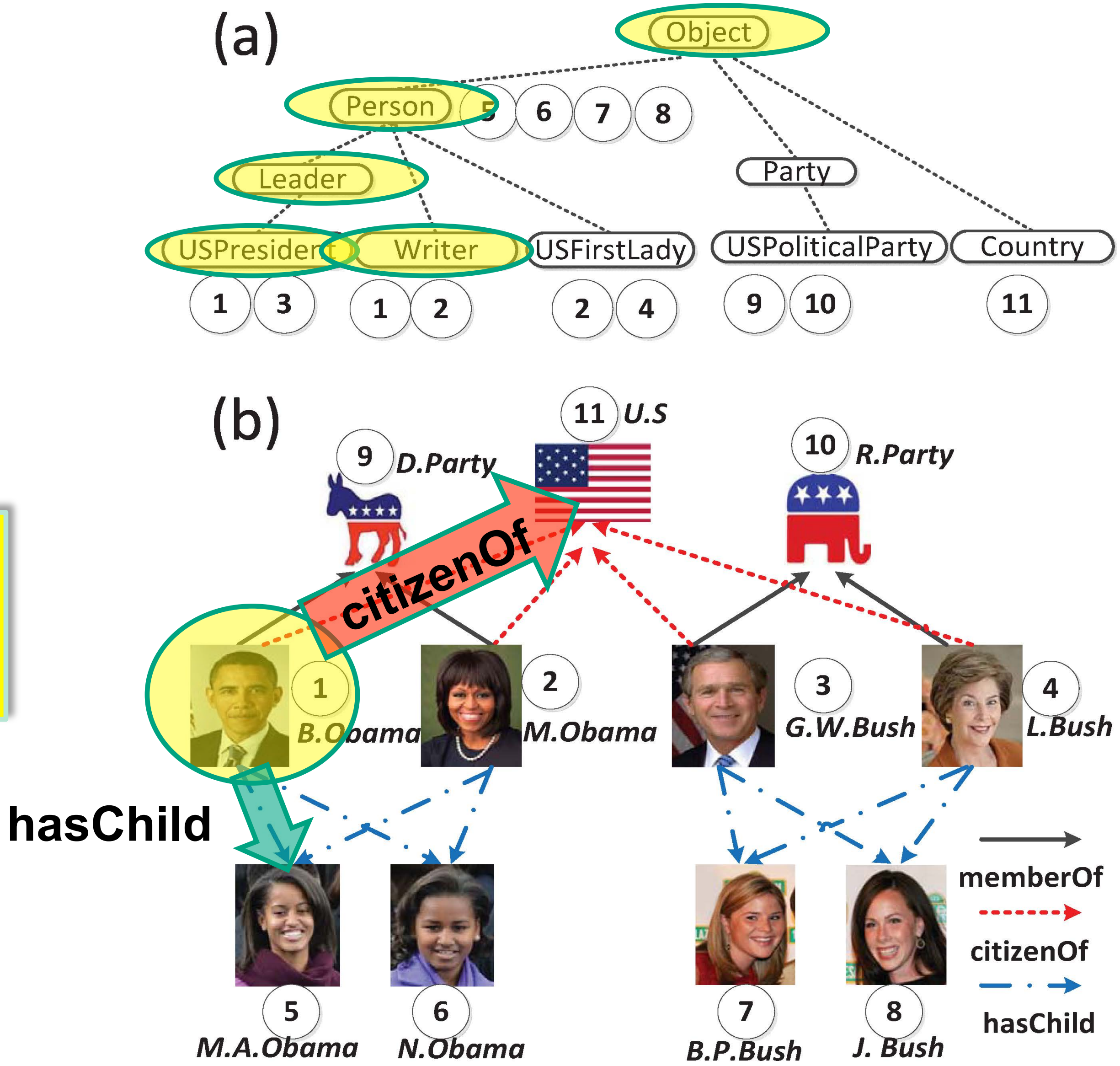
Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, Xiang Li. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. KDD 2016.

# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search
- Conclusions & Future Work

# Fundamental question: Relevance Computation

s B. Obama *relevant to* G. W. Bush?



# Relevance Search

- **How to measure the similarity?**
  - Define a **Effective Similarity Function** like *Cosine*, *Euclidean distance*, *Jaccard coefficient*.
- **Structure similarity or Semantic similarity?**
  - Structure Similarity: Based on structural similarity of **sub−network** structures. (like SimRank and PPR)
  - **Semantic Similarity:** **influenced** by **similar network** structures. This matters more for HIN! Semantic->edge relations

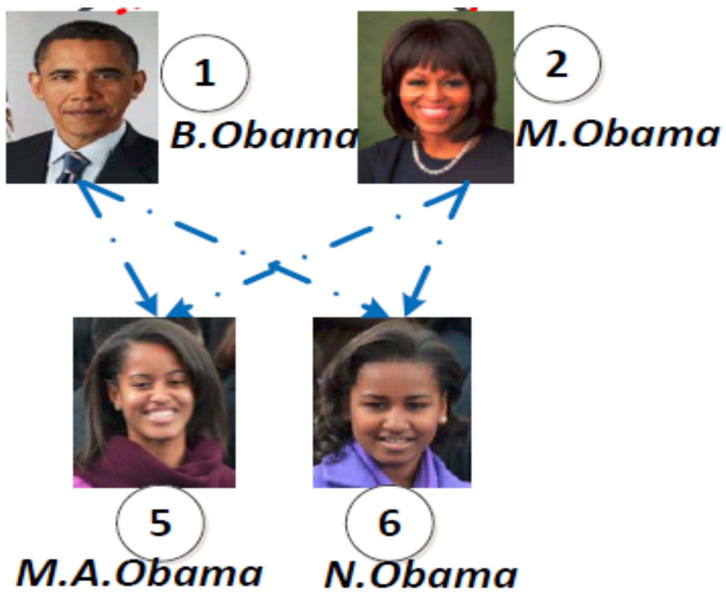
# Meta Path [Sun VLDB'11]

**Meta path: a sequence of node classes connected by edge types**

$m1$ :USPresident $\xrightarrow{\text{hasChild}}$ Person $\xrightarrow{\text{hasChild}^{-1}}$ USFirstLady,

$m2$ :USPresident $\xrightarrow{\text{memberOf}}$ USPoliticalParty $\xrightarrow{\text{memberOf}^{-1}}$ USFirstLady,

$m3$ :USPresident $\xrightarrow{\text{citizenOf}}$ Country $\xrightarrow{\text{citizenOf}^{-1}}$ USFirstLady.

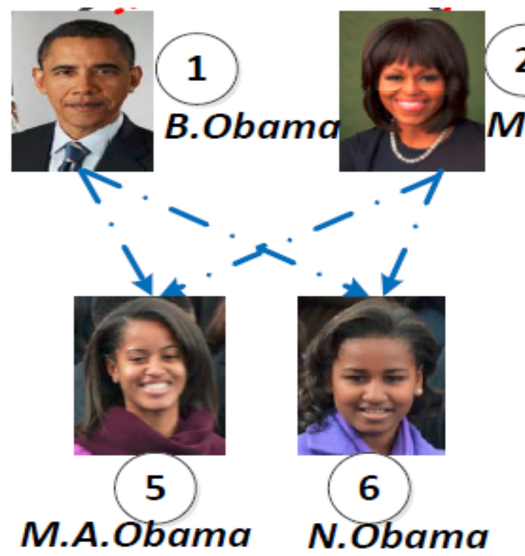



**Meta paths can be used to define relevance between 2 nodes.**

# Meta Path Relevance 1: Path Count (PC)

- **Path Count(PC)** [Sun VLDB'11]
  - **Number of the paths following a given meta path**
  - **PC(B.Obama, M.Obama) = 1+1=2, because there are two path instances.**

$$m1 : \text{USPresident} \xrightarrow{\text{hasChild}} \text{Person} \xrightarrow{\text{hasChild}^{-1}} \text{USFirstLady},$$




  - **PC biases popular objects with a large no. of links.**

# Meta Path Relevance 2: Path Constrained Random Walk

- **Model**

  Random walk on given paths.

- **Definition**
  - Performing random walks on given meta-paths between source and target node.
  - **PCRW**: Transition probability of the random walk following **a given meta-path**.

$$\mathrm{PCRW}(s,t|\mathbf{\Pi}) = \mathrm{P}(\boldsymbol{s} \rightarrow \boldsymbol{t}|\mathbf{\Pi})$$

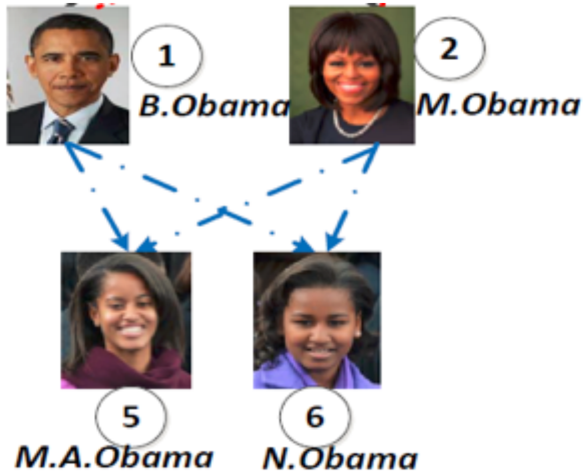
  - Between [0, 1].

[Cohen ECML'11]W. Cohen, N. Lao "Relational Retrieval Using a Combination of Path-Constrained Random Walks"

# PCRW

## ○ Example

$m_1$ Person $\xrightarrow{\text{hasChild}}$ Person $\xrightarrow{\text{hasChild-1}}$ Person




$m_1$ = P1 -> P2 -> P3

**PCRW(B. Obama, M. Obama) = 0.5**

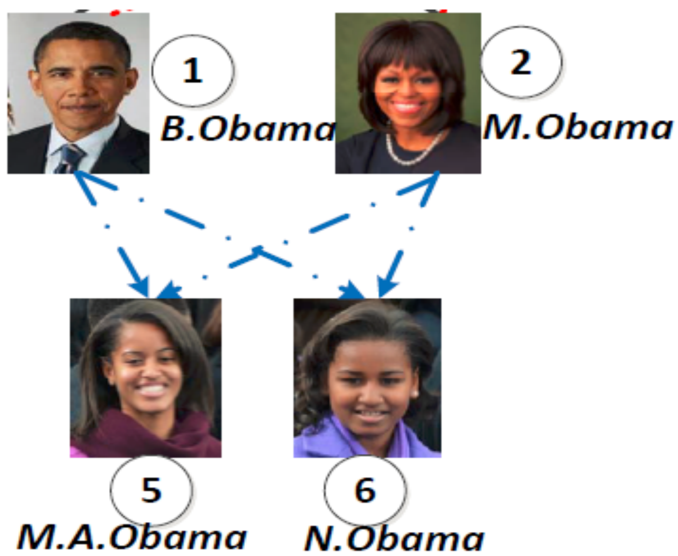
1. Pr(B.Obama | P1)=1
2. Pr(M.A. Obama | P2) = Pr(B.Obama | P1) / 2 = 0.5
   Pr(N.Obama | P2) = Pr(B.Obama | P1) / 2= 0.5
3. Pr(M.Obama | P3) = Pr(M.A. Obama | P2) /2 + Pr(N.Obama | P2) /2 = 0.5
   Pr(B.Obama | P3) = Pr(M.A. Obama | P2) 2 + Pr(N.Obama | P2) /2 = 0.5

[Cohen ECML'11]W. Cohen, N. Lao "Relational Retrieval Using a Combination of Path-Constrained Random Walks"

# Meta Path Relevance 3: BPCRW

- **Biased Path Constrained Random Walk(BPCRW)** [Meng WWW'15]
  - **Generalization of PC and PCRW.**
  - **Biased factor α in [0,1].**
    - **When α = 0, BPCRW becomes PC;**
    - **When α = 1, BPCRW becomes PCRW.**

$$m1 : \text{USPresident} \xrightarrow{\text{hasChild}} \text{Person} \xrightarrow{\text{hasChild}^{-1}} \text{USFirstLady},$$

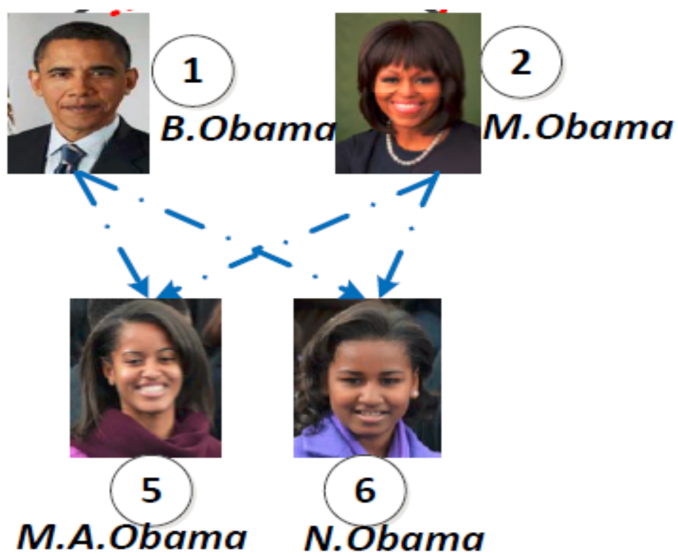


# Meta Path Relevance 4: PathSim (PS)

- **PathSim(PS)** **[Sun VLDB'11]**
  - **For symmetric meta paths only**
  - **PS is a normalized version of PC, with a value in [0, 1].**

$m1$ :USPresident $\xrightarrow{\text{hasChild}}$ Person $\xrightarrow{\text{hasChild}^{-1}}$ USFirstLady,

$m_2$ Person $\xrightarrow{\text{hasChild}}$ Person $\xrightarrow{\text{hasChild-1}}$ Person




- **PS(B.Obama, M.Obama | $m_2$)=1**

# Recent Developments

- **HeteSim** **(APWeb'14)**

  Enhanced version of SimRank

- **KnowSim** **(APWeb'14)**

  Based on given meta-path and the reverse meta-path

- **AvgSim** **(ICDM'16)**

  Measure the similarity of documents in HIN

- **RelSim** **(SDM'16)**

  Measure the similarity of relations in HIN

# Questions

- **Where do meta paths come from?**
  - **Provided by experts [Sun VLDB'11]**
    - **Not easy for a complex schema!**
  - **Enumeration within a given length of meta paths [Cohen ECML'11]**
    - **No clue about the length!**
  - **How do I know the weights ?**

# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Definition
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search

Conclusions & Future Work

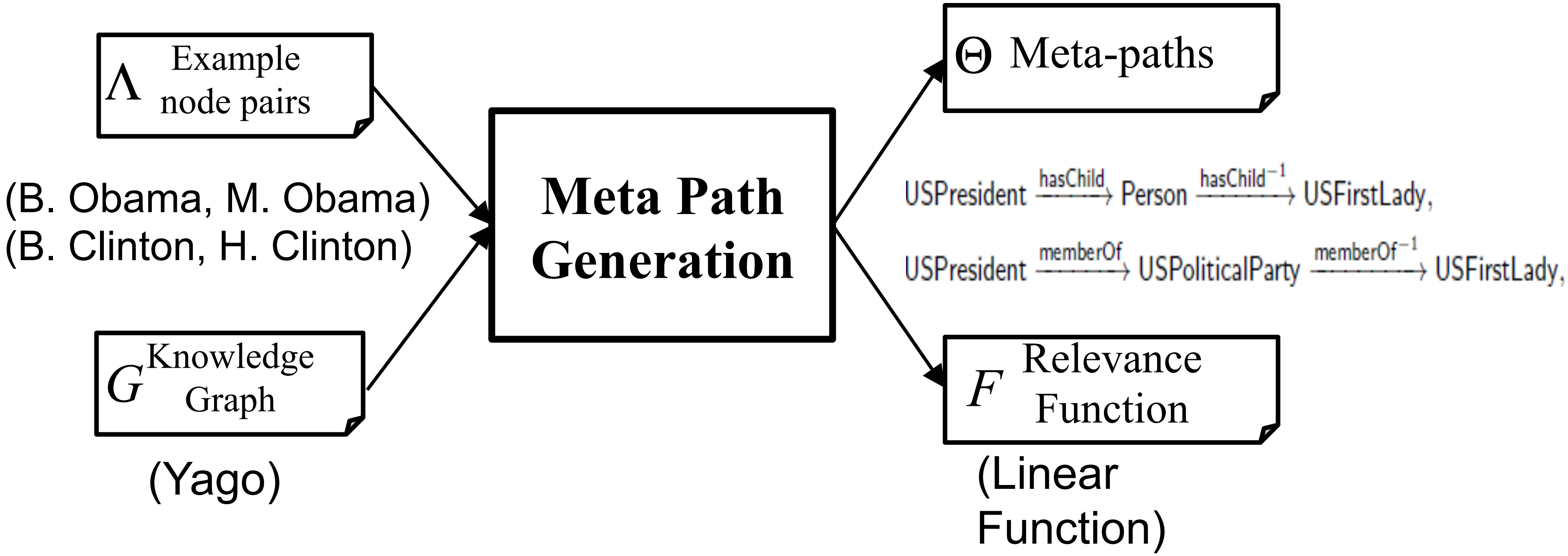
# Our Contributions (WWW'15)

- **Design a solution that:**
  - **(1) Discovers the best meta paths**
  - **(2) Learns the weights, without maximum weight specified.**

**[Meng WWW'15]** Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, and Wangda Zhang. "Discovering Meta-Paths in Large Heterogeneous Information Networks", in WWW 2015.

# Meta-Path Framework

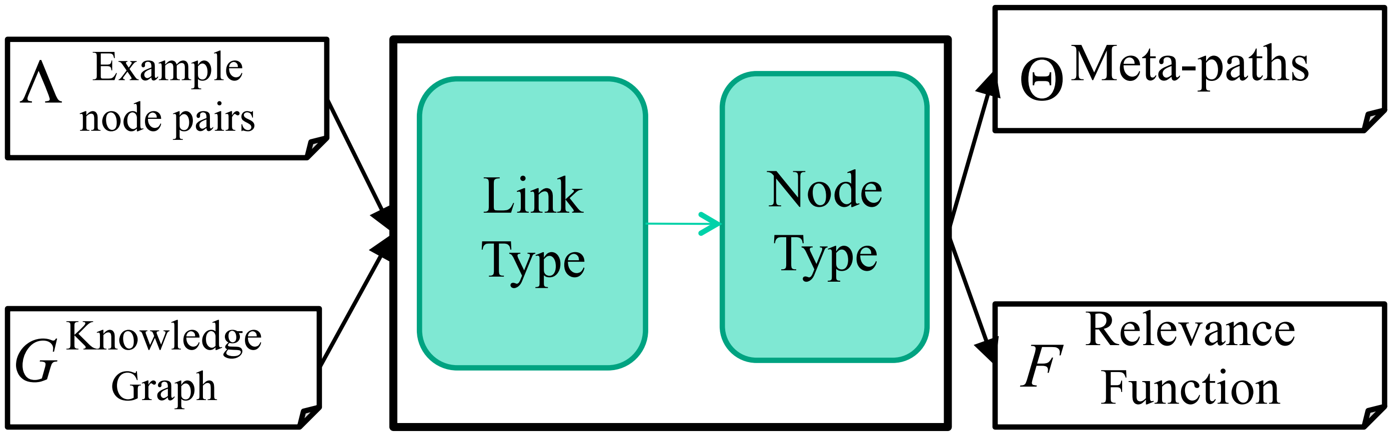
## ○ Framework




**Challenge**: Each node and edge can have many class labels. The number of candidate meta paths grows exponentially with their path lengths.

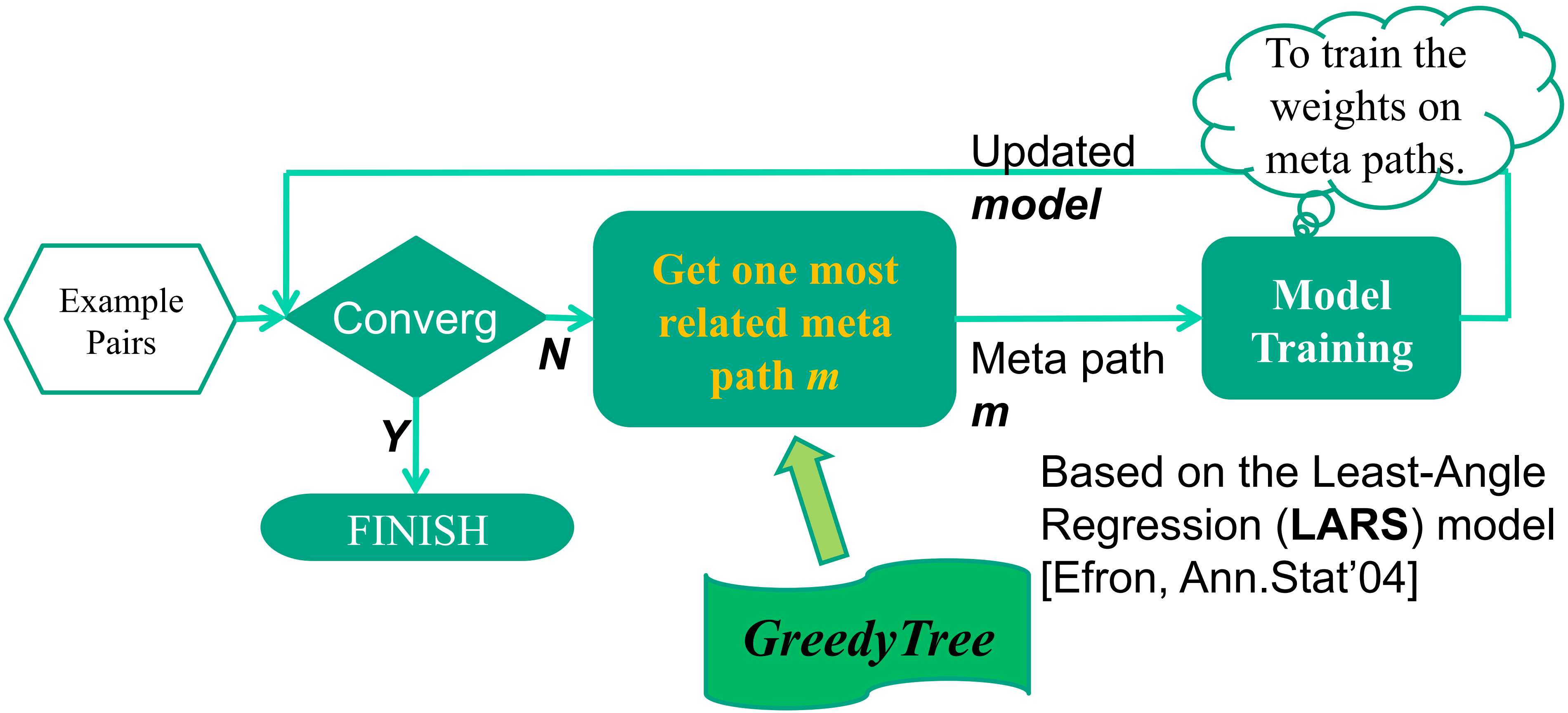
# Generating Meta-Paths

- In Two Phases



# Phase 1: Link-Only Path Generation

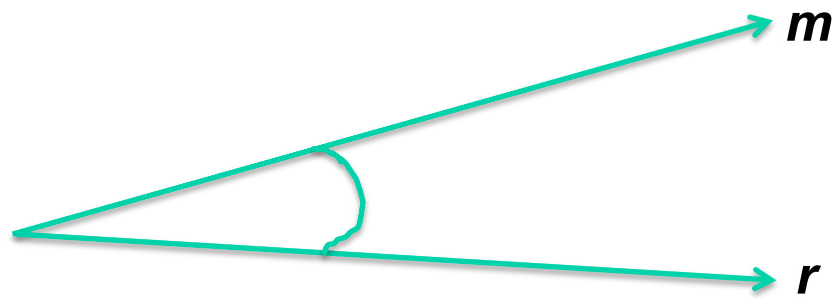
- **Forward Stage-wise Path Generation (FSPG)**
  - **iterratively generate the most related meta-paths and update the model**




Based on the Least-Angle Regression (**LARS**) model [Efron, Ann.Stat'04]

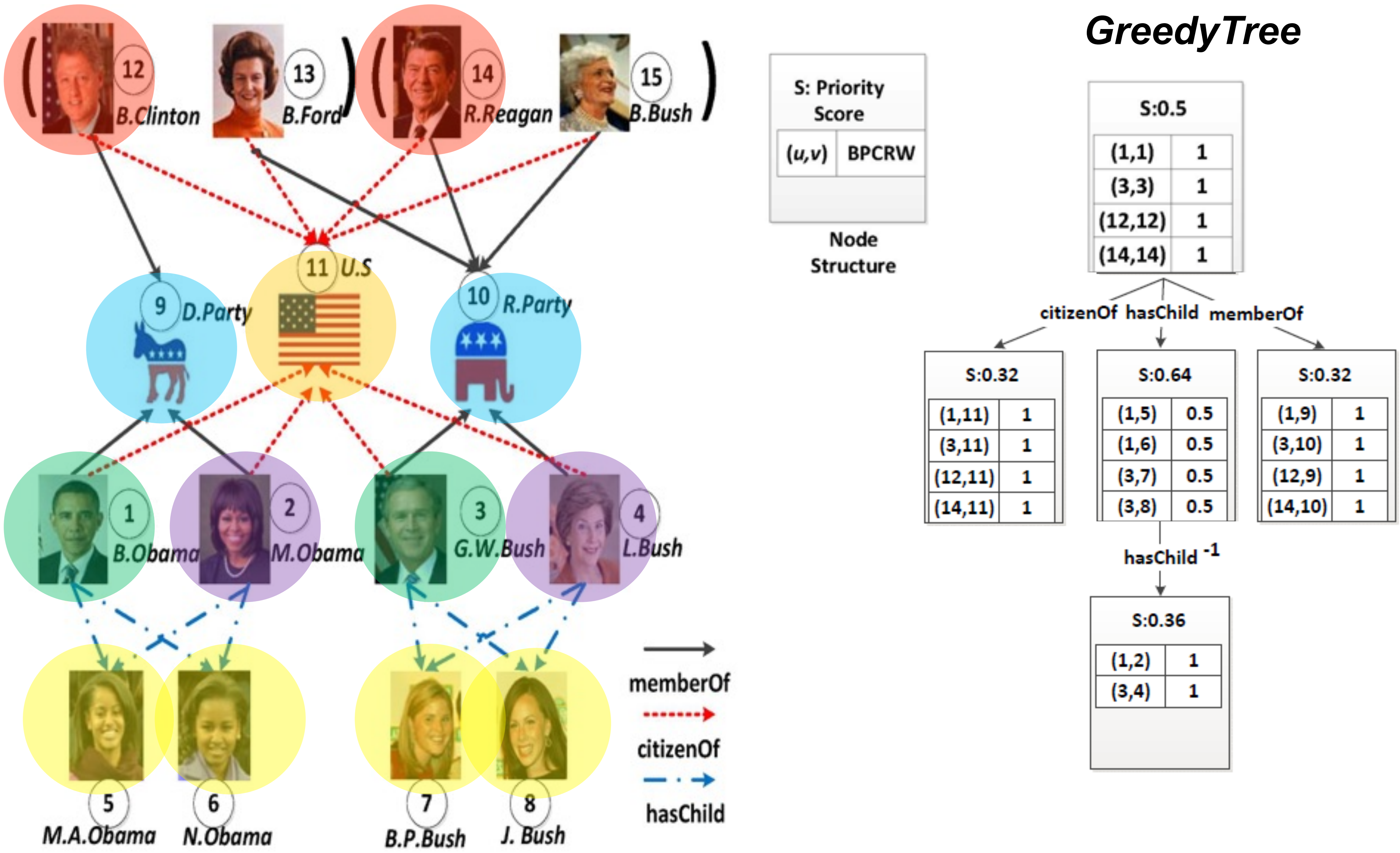
# Meta path Generation

- GreedyTree
  - A tree that greedily expands the node which has the largest priority score
  - Priority Score : related to the correlation between *m* and *r*
    - *m* is the vector expression of a meta path, *r* is the residual vector which evaluates the gap between the truth and current model




Details in WWW'05

# Phase 1: Link-Only Path Generation

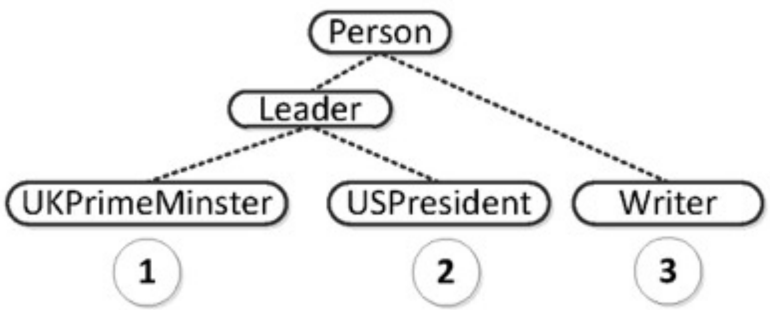


# Phase 2: Node Class Generation

- **Why node classes?**
  - **A link only meta path may introduce some unrelated result pairs**
  - **It is less specific**

$$? \xrightarrow{\text{liveIn}} ? \qquad \qquad \text{Scientist} \xrightarrow{\text{liveIn}} \text{CapitalCity}$$

  - **Solution : Lowest Common Ancestor (LCA)**
    - **Record the LCA in the node of GreedyTree**

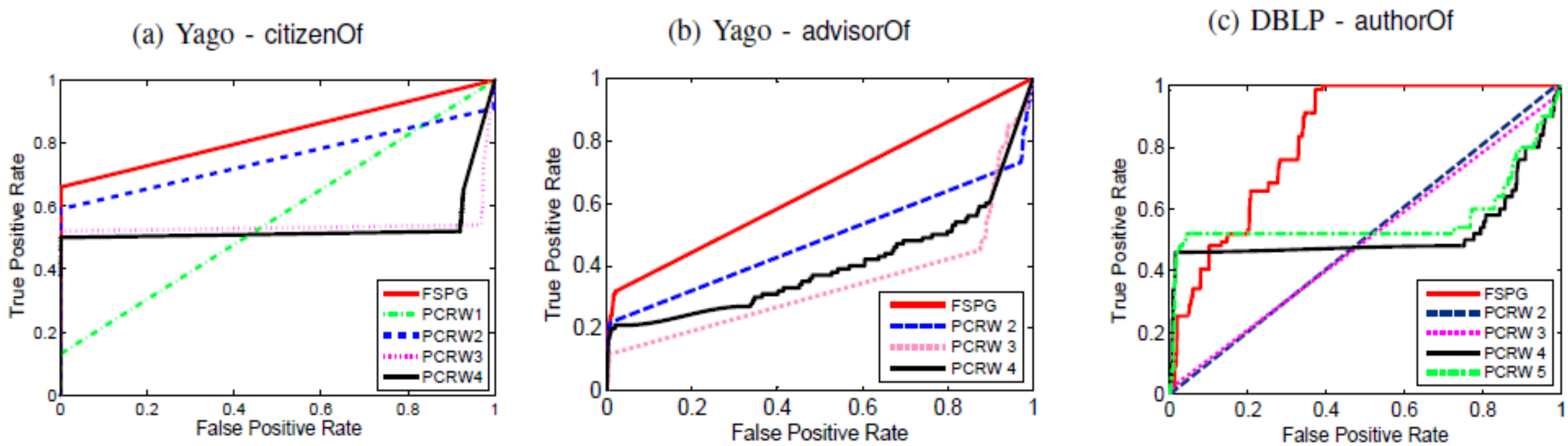


# Experiments

- **Datasets**
  - **DBLP (4 areas: DB, DM, AI, IR)**
    - **14K papers, 14K authors, 9K topics, 20 venues.**
  - **Yago**
    - **A KG derived from Wikipedia, WordNet and GeoNames.**
    - **CORE Facts: 2.1 million nodes, 8 million edges, 125 edge types, 0.36 million node types**
- **Link-prediction evaluation**
  - **Select n pairs of certain relationships as example pairs**
  - **Randomly select another m pairs to predict the links**

# Experiment 1: Effectiveness

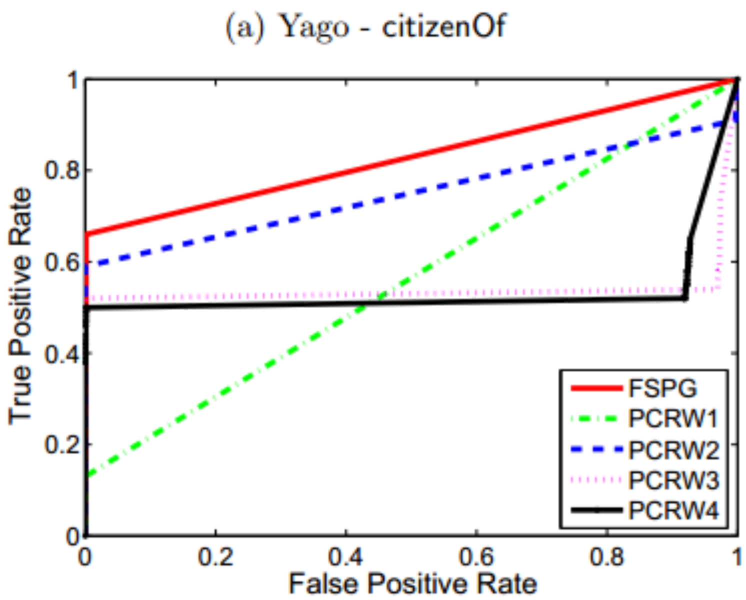
- **Baseline: enumerate all meta paths within a given max length L = 1, 2, 3, 4**
  - **L is small → low recall.**
  - **L is large → low precision.**




ROC for link prediction

# Experiment 2

- Case study: Yago citizenOf
  - Better than direct link (PCRW 1)
  - Better than best PCRW 2
  - Better than PCRW 3,4

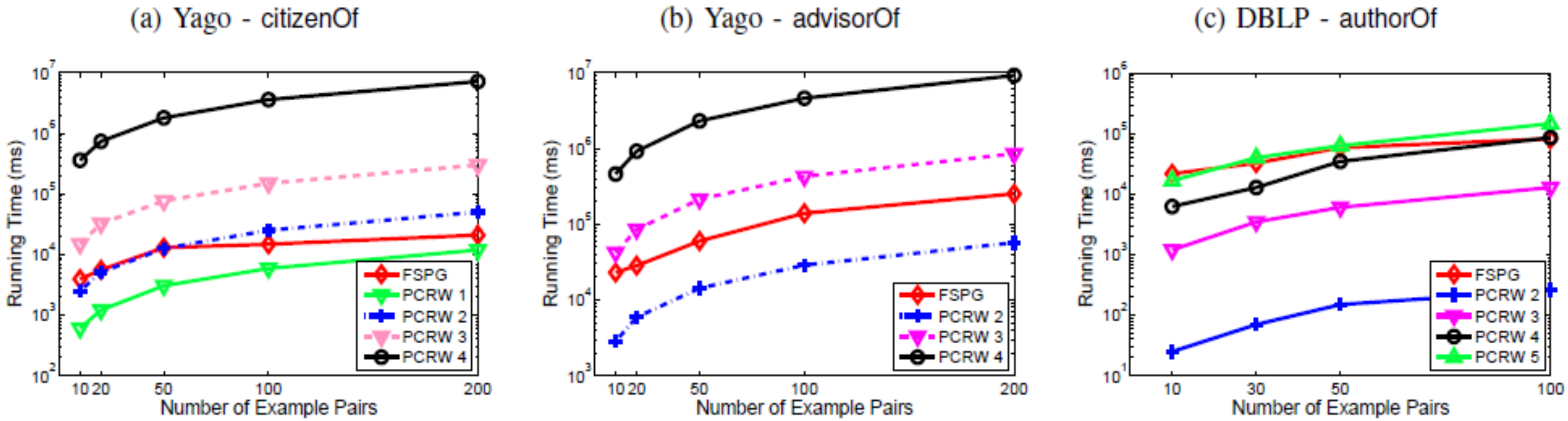



| meta-path | w |
|---|---|
| Person $\xrightarrow{\text{bornIn}}$ City $\xrightarrow{\text{locatedIn}}$ Country | 5.477 |
| Person $\xrightarrow{\text{livesIn}}$ Country | 0.361 |
| Person $\xrightarrow{\text{graduateOf}}$ University $\xrightarrow{\text{locatedIn}}$ Country | 0.023 |
| Person $\xrightarrow{\text{diedIn}}$ City $\xrightarrow{\text{locatedIn}}$ Country | 0.245 |
| Person $\xrightarrow{\text{bornIn}}$ City $\xrightarrow{\text{happenedIn}^{-1}}$ Event $\xrightarrow{\text{happenedIn}}$ Country | 0.198 |

5 most relevant meta paths for “citizenOf”

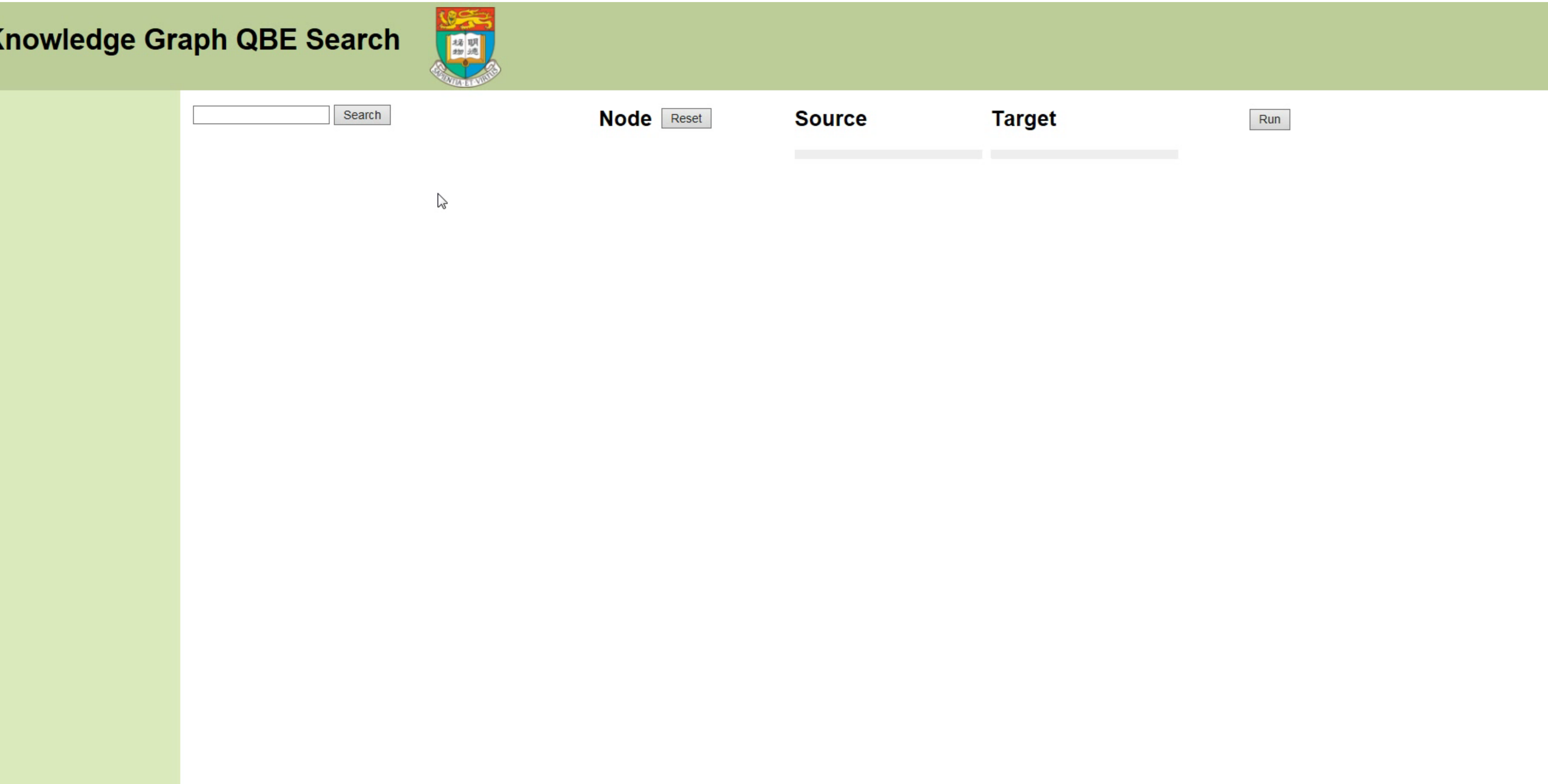
# Experiment 3: Efficiency

- **Findings:**
  - **In Yago, 2 orders of magnitude better than paths with lengths more than 2.**
  - **In DBLP, the running time is comparable to PCRW 5, but the accuracy is much better.**




Running time of FSPG

# Demo



nowledge Graph QBE Search
Search
Node
Reset
Source
Target
Run

# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Definition
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search
- Conclusions & Future Work

# Query Recommendation

- Suggest relevant queries to a search engine user
  - 1) As you type;
  - 2) *Related queries*



Zhipeng Huang, Bogdan Cautis, Reynold Cheng, Yudian Zheng. KB-Enabled Query Recommendation for Long-Tail Queries. CIKM 2016.

# Query Log

- Existing methods rely on query logs to analyze the flow among queries.
- A set of user log <q, u, t, C>
  - q: the query
  - u: user id
  - t: time stamp
  - C: the clicked URLs

Boldi, Paolo, et al. "The query-flow graph: model and applications." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

Bonchi, Francesco, et al. "Efficient query recommendations in the long tail via center-piece subgraphs." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

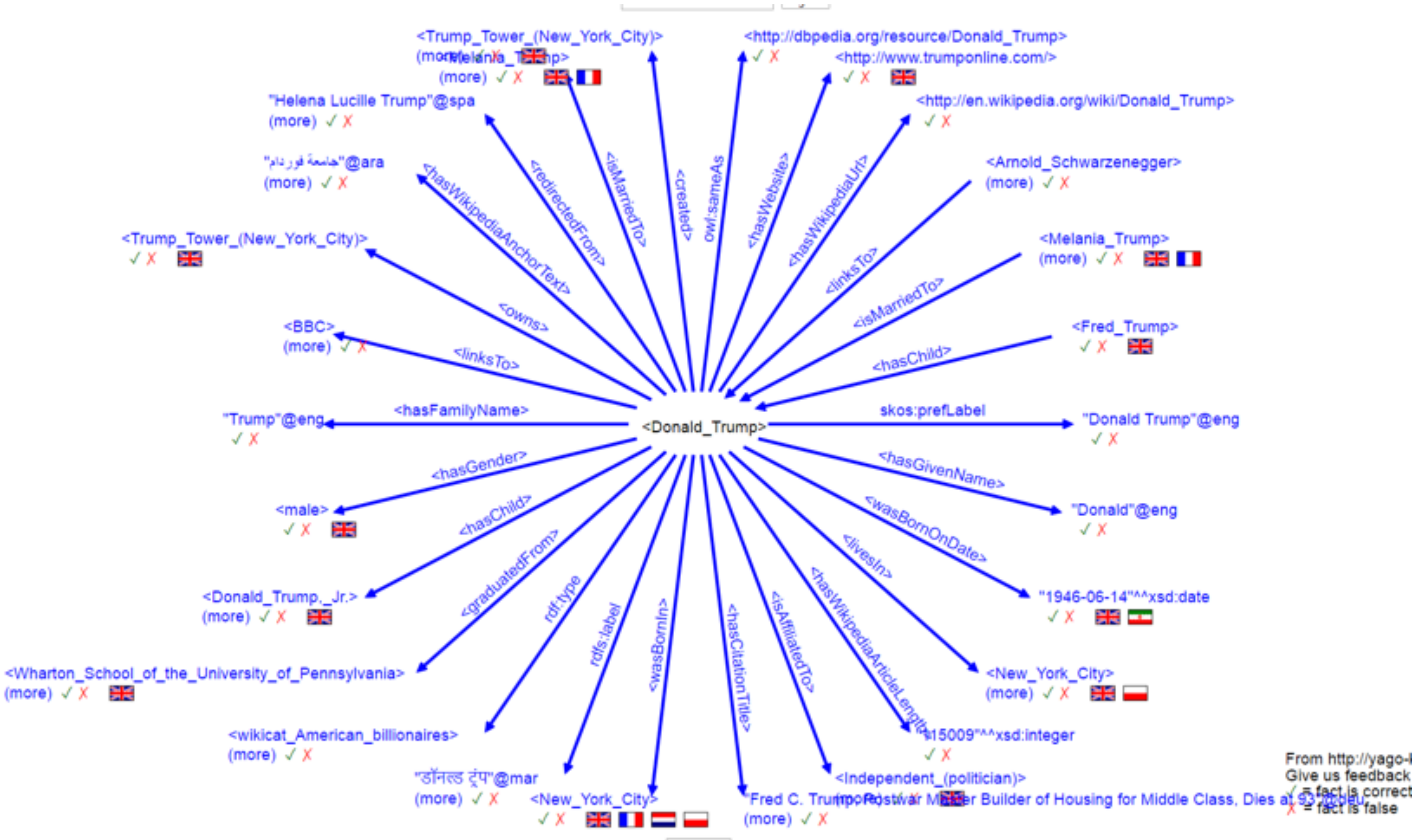
# Long Tail Distribution

- Long-tail queries: queries that are not commonly requested by users
  - "*akira kurosawa influence george lucas*"

# Motivation

- **Ubiquity:**
  - **84% of 10M queries appear no more than 3 times.**
- **Necessity:**
  - **Existing works often only rely on <u>query log</u> alone**

# Knowledge Graph




Hoffart, Johannes, et al. "Yago2: a spatially and temporally enhanced Knowledge Graph from wikipedia." (2012).

# Relationship in KG

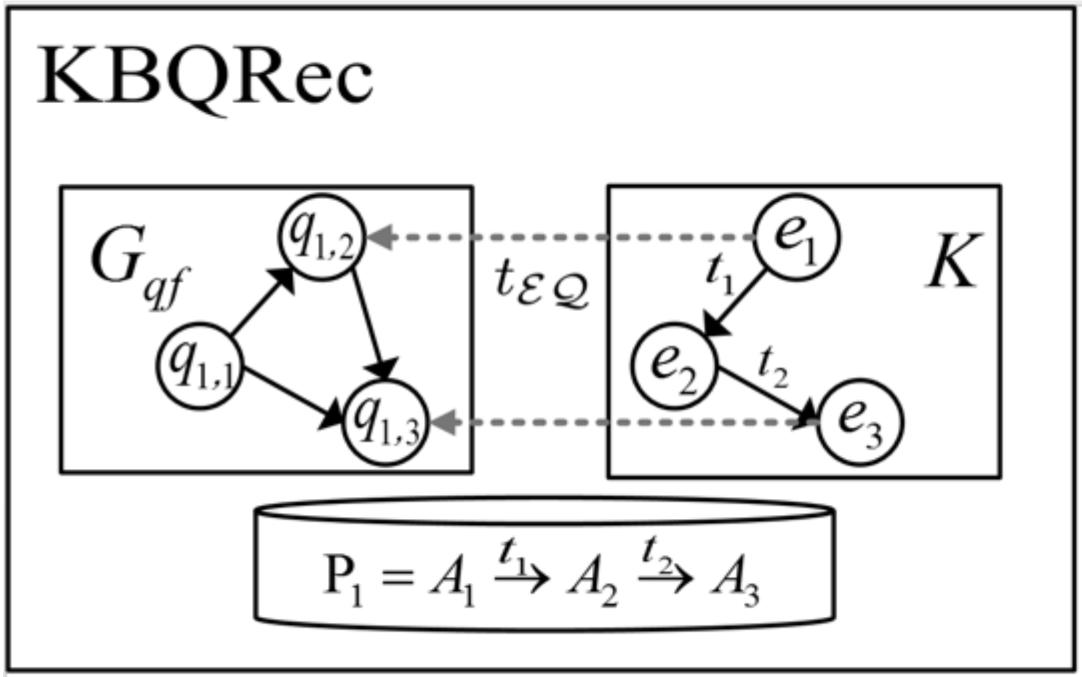
- **Meta path representation:**
  - **P: city nextTo city** $\longrightarrow$

- **Q: "weather Los Angeles"**
  - **Rec:**
    - **"weather Las Vegas"**
    - **"weather San Diego"**

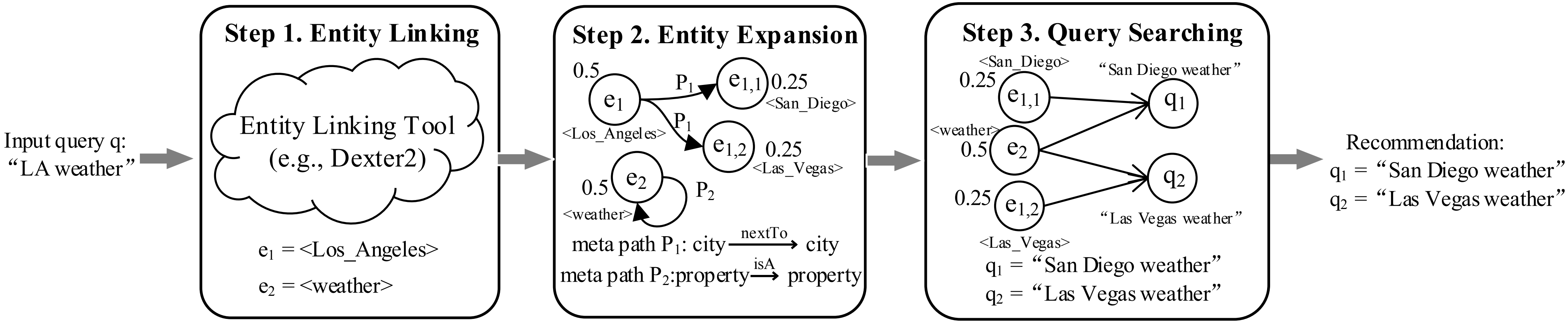
[Sun, Han VLDB'11] Y. Sun, J. Han, el "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks

# System Overview

- $G = (G_{qf}, K, t_{eq}, P)$
  - $G_{qf}$ is a query-flow graph
  - K is a Knowledge Graph
  - $t_{EQ}$ is a set of entity-query links
  - P is a set of meta path to be extracted from query log

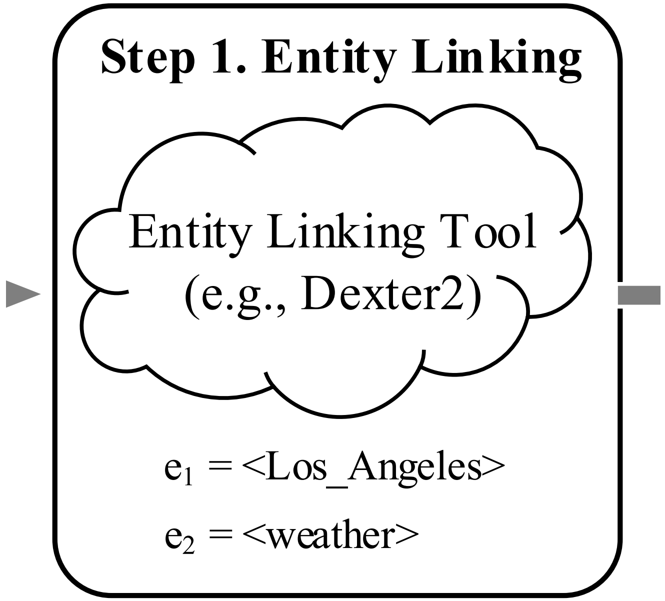


# Online Process



# Step 1: Entity Linking

- **Given**
  - **q = “weather Los Angeles”**
- **Return:**
  - **$e_1$ = Los_Angeles**

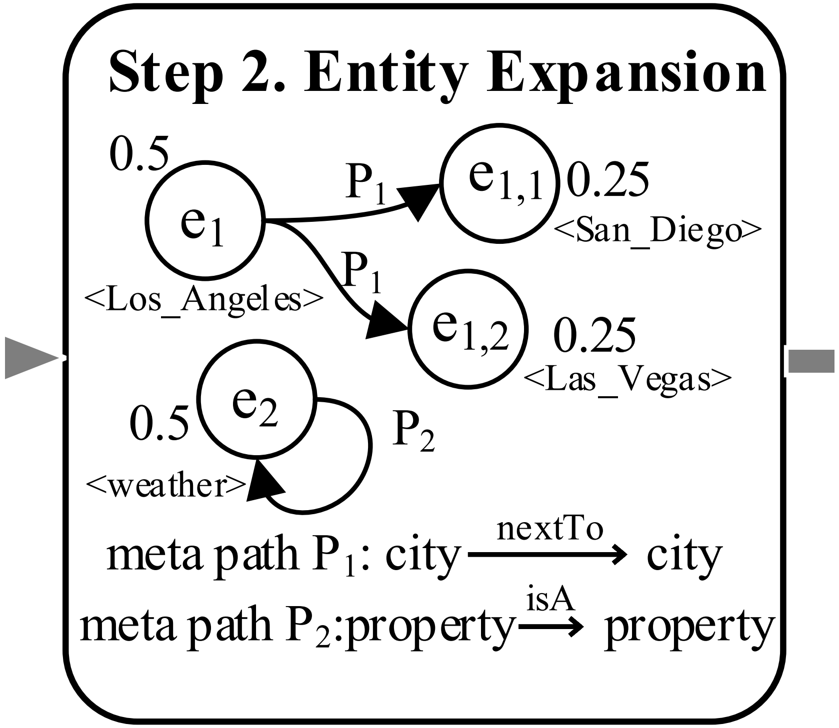



Ceccarelli, Diego, et al. "Dexter: an open source framework for entity linking." Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval. ACM, 2013.

# Step 2. Entity Expansion

- Given
  - $e_1$ = Los_Angeles
- Using P:
  - city $\xrightarrow{\text{NextTo}}$ city

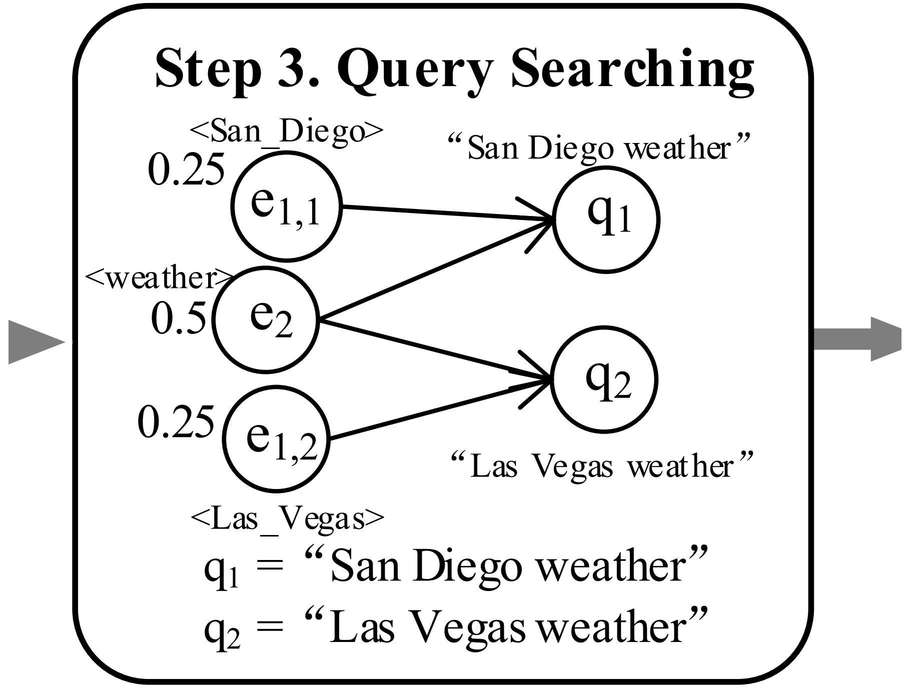
- Return
  - $e_2$ = Las_Vegas
  - $e_3$ = San_Diego



# Step 3. Query Searching

- Given:
  - $e_2$ = Las_Vegas
  - $e_3$ = San_Diego

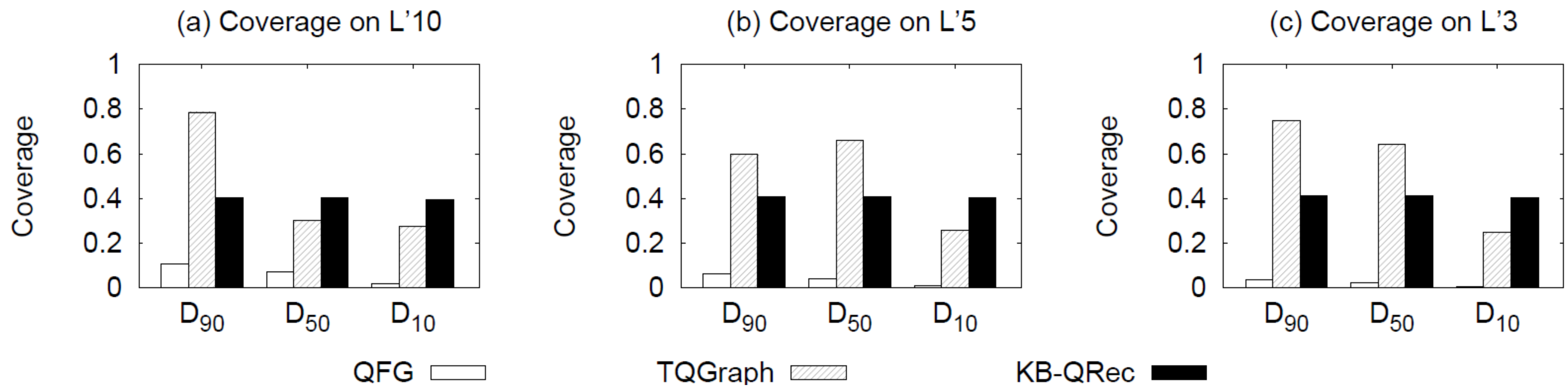
- Return:
  - $q_1$ = “weather las vegas”
  - $q_2$ = “weather san diego”

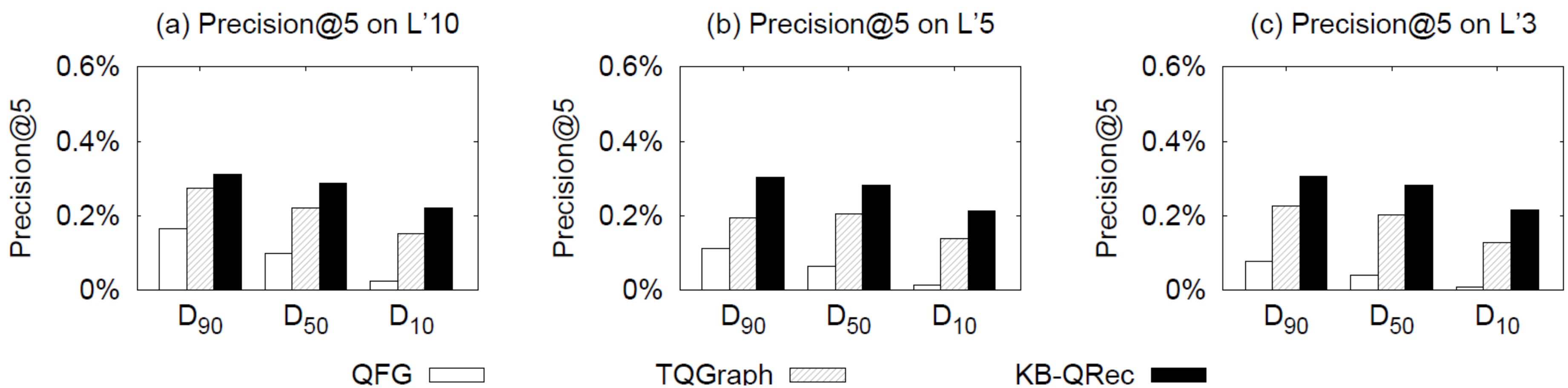


# Experiments

- Dataset: AOL. 20M query instances from 9M distinct queries.
- Use 10%, 50%, 90% for building the query log, and 10% for testing.
- Testing sets: We use 3, 5, 10 as the threshold for long-tail queries. We name them L'3, L'5 and L'10.
- Measures:
  - Coverage
  - Precision@5

# Experimental Results



# Efficiency

- **Time for offline:**

Table 4: Efficiency for building KB-QREC's index.

| | $D_{10}$ | $D_{50}$ | $D_{90}$ |
|---|---|---|---|
| Building Time | 14 min | 56 min | 132 min |

- **Time for entity linking:**
  - **60ms for Dexter2; can be reduced to 0.4ms if we use FEL method.**

Table 5: Efficiency (in ms)

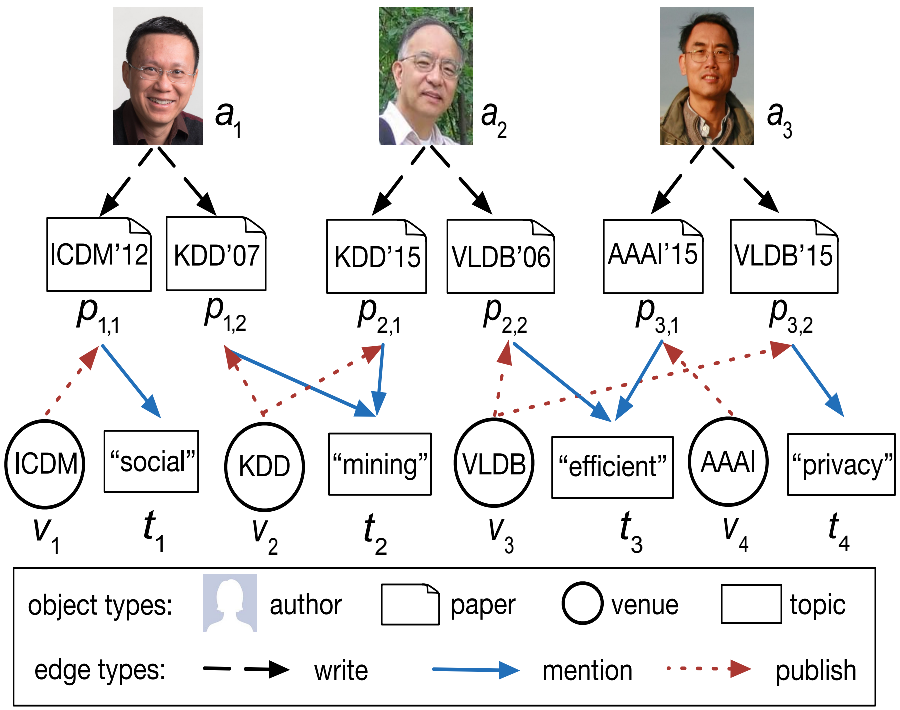
| | entity expansion | PPR (no cache) | PPR (cache) | KB-QREC (no cache) | KB-QREC (cache) |
|---|---|---|---|---|---|
| $D_{90}$ | 34 ms | 91 ms | 9 ms | 143 ms | 60 ms |
| $D_{50}$ | 34 ms | 55 ms | 5 ms | 100 ms | 47 ms |
| $D_{10}$ | 33 ms | 13 ms | 1 ms | 59 ms | 37 ms |

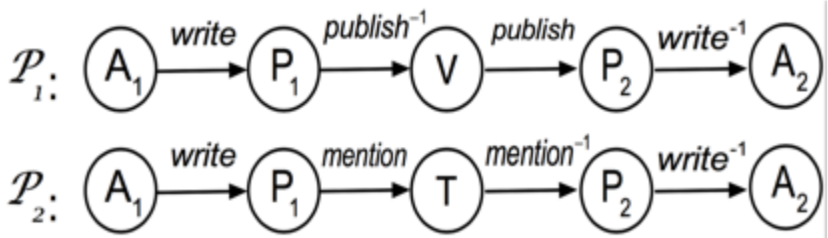
# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search
- Conclusions & Future Work

# Limitations of Meta Paths

- Fail to discover common nodes in different meta paths!
  - E.g., a researcher wants to search for some authors who have published papers in the same venue *and* in the same topic with his

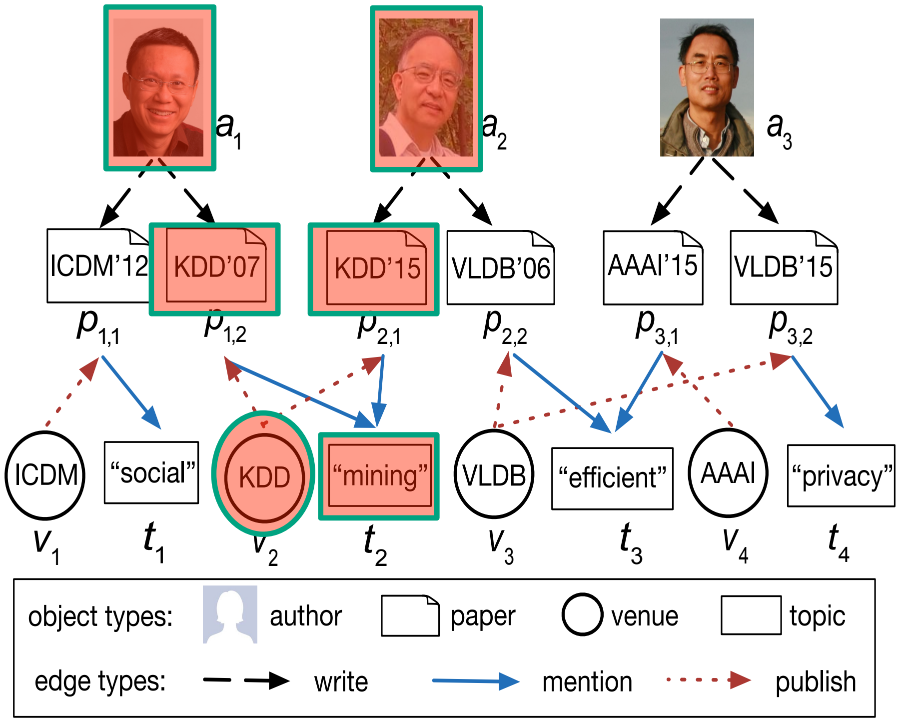


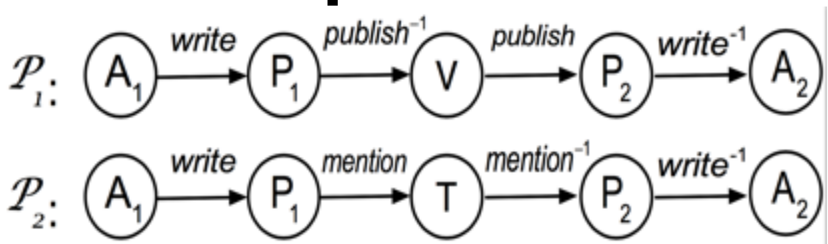




| Pair | Meta Path Measures | | |
|---|---|---|---|
| | PathCount | PathSim | PCRW |
| $a_2, a_1$ | 2 | 0.5 | 0.25 |
| $a_2, a_3$ | 2 | 0.5 | 0.25 |

# Limitations of Meta Paths

- **Fail to discover common nodes in different meta paths!**
  - **E.g., a researcher wants to search for some authors who have published papers in the same venue *and* in the same topic with his**

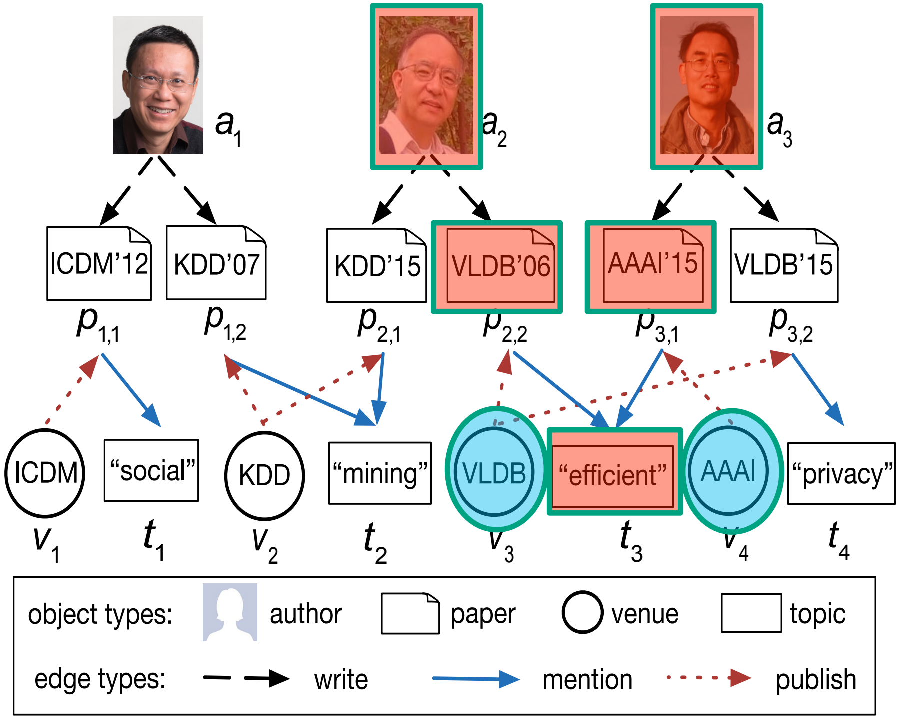


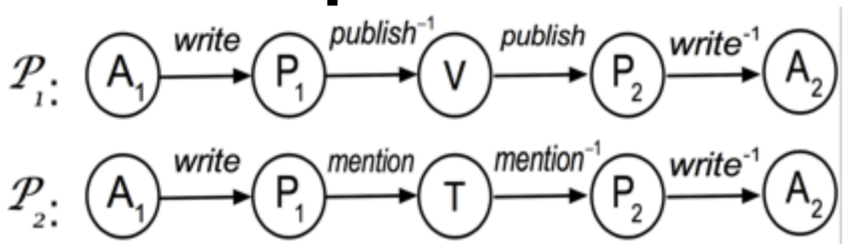




| Pair | Meta Path Measures | | |
|---|---|---|---|
| | PathCount | PathSim | PCRW |
| $a_2, a_1$ | 2 | 0.5 | 0.25 |
| $a_2, a_3$ | 2 | 0.5 | 0.25 |

# Limitations of Meta Paths

- Fail to discover common nodes in different meta paths!
  - E.g., a researcher wants to search for some authors who have published papers in the same venue *and* in the same topic with his

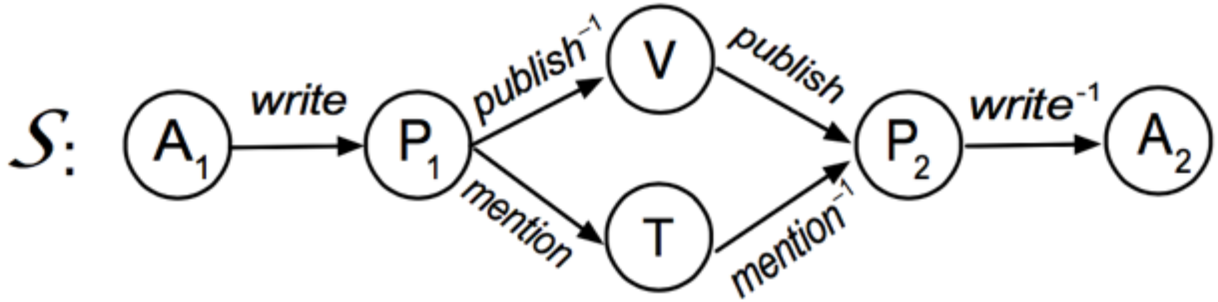



| Pair | Meta Path Measures | | |
|---|---|---|---|
| | PathCount | PathSim | PCRW |
| $a_2, a_1$ | 2 | 0.5 | 0.25 |
| $a_2, a_3$ | 2 | 0.5 | 0.25 |

# Meta Structure

- A meta structure is a directed acyclic graph (DAG) with a single source and sink (target) node




- More Expressive (i.e., contain more information) than a meta path.

[Huang KDD'16] ZP. Huang "Meta Structure: Computing Relevance on Large Heterogeneous Information Networks" KDD 2016

# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Definition
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search
- Conclusions & Future Work

# Relevance Measure 1: StructCount

- ***StructCount*: extension of *PathCount***

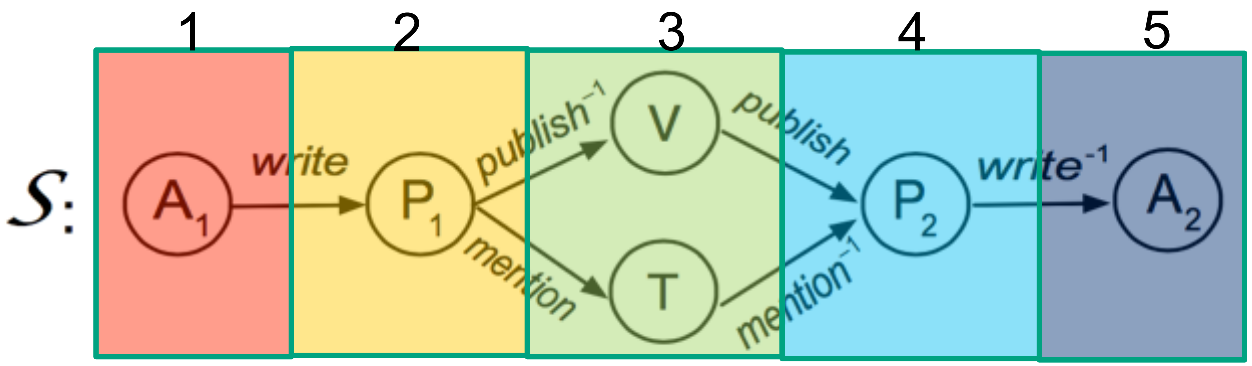
$$StructCount(x_0, y_0 \mid S) = \left|GraphIns(x_0, y_0 \mid S)\right|$$

- **StructCount biases towards popular objects with a large number of links.**

[Huang KDD'16] ZP. Huang "Meta Structure: Computing Relevance on Large Heterogeneous Information Networks" KDD 2016

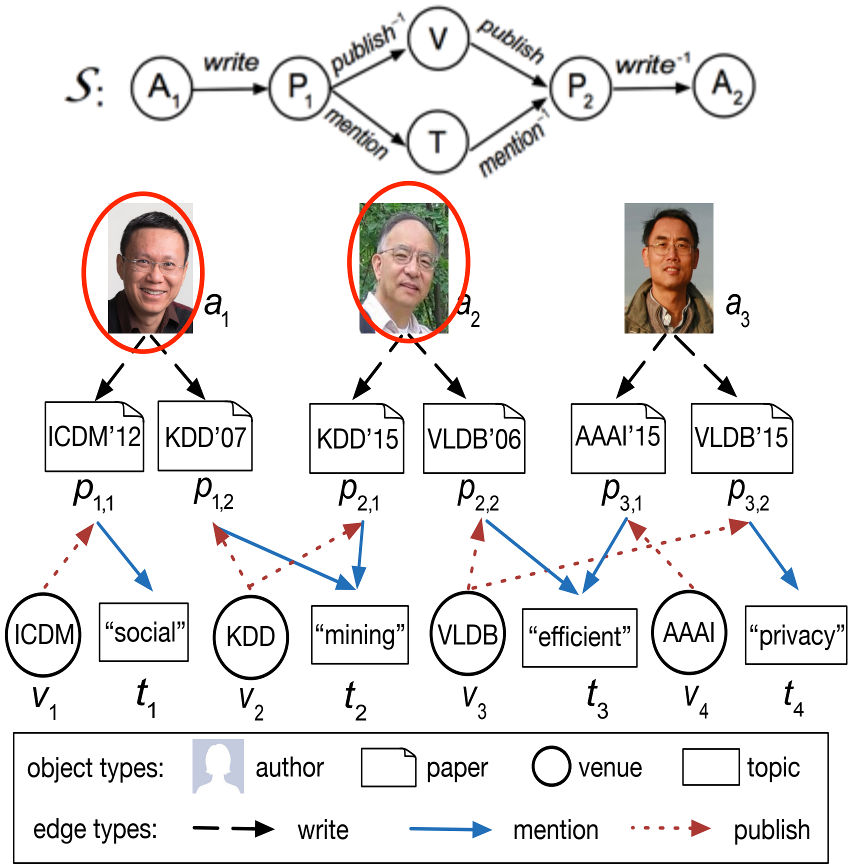
# Layers of Meta Structure

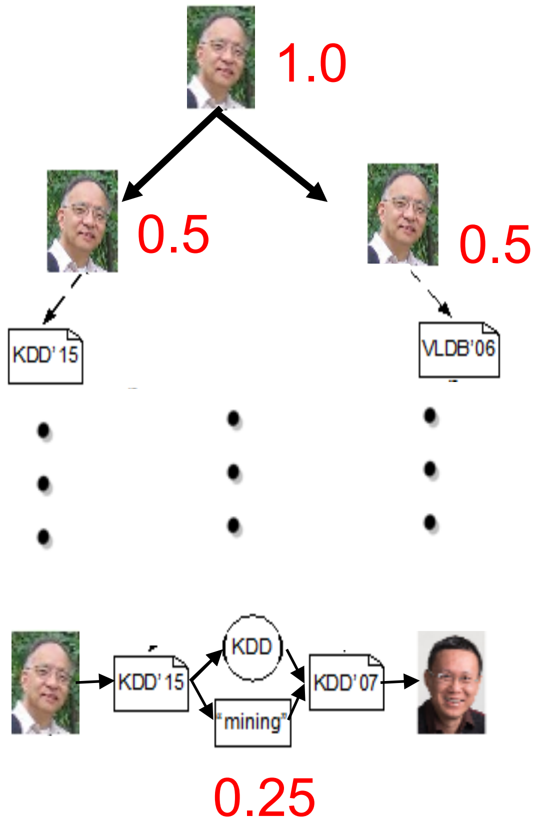
- **The layer of meta structure is a topological ordering of a DAG**



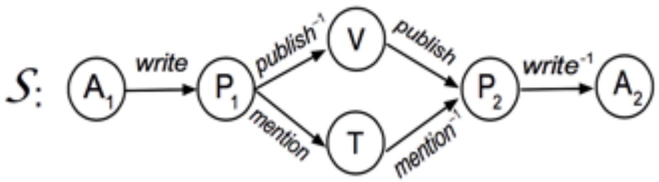
# Relevance Measure 2: SCSE

- **Structure Constrained Random Walk (SCSE): extension of PCRW.**

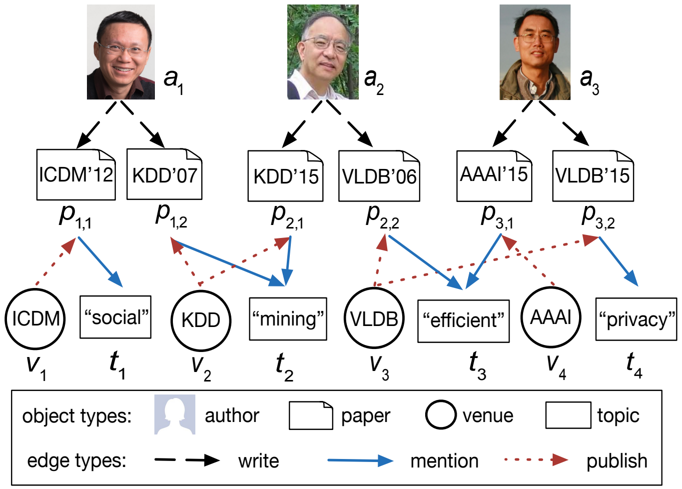


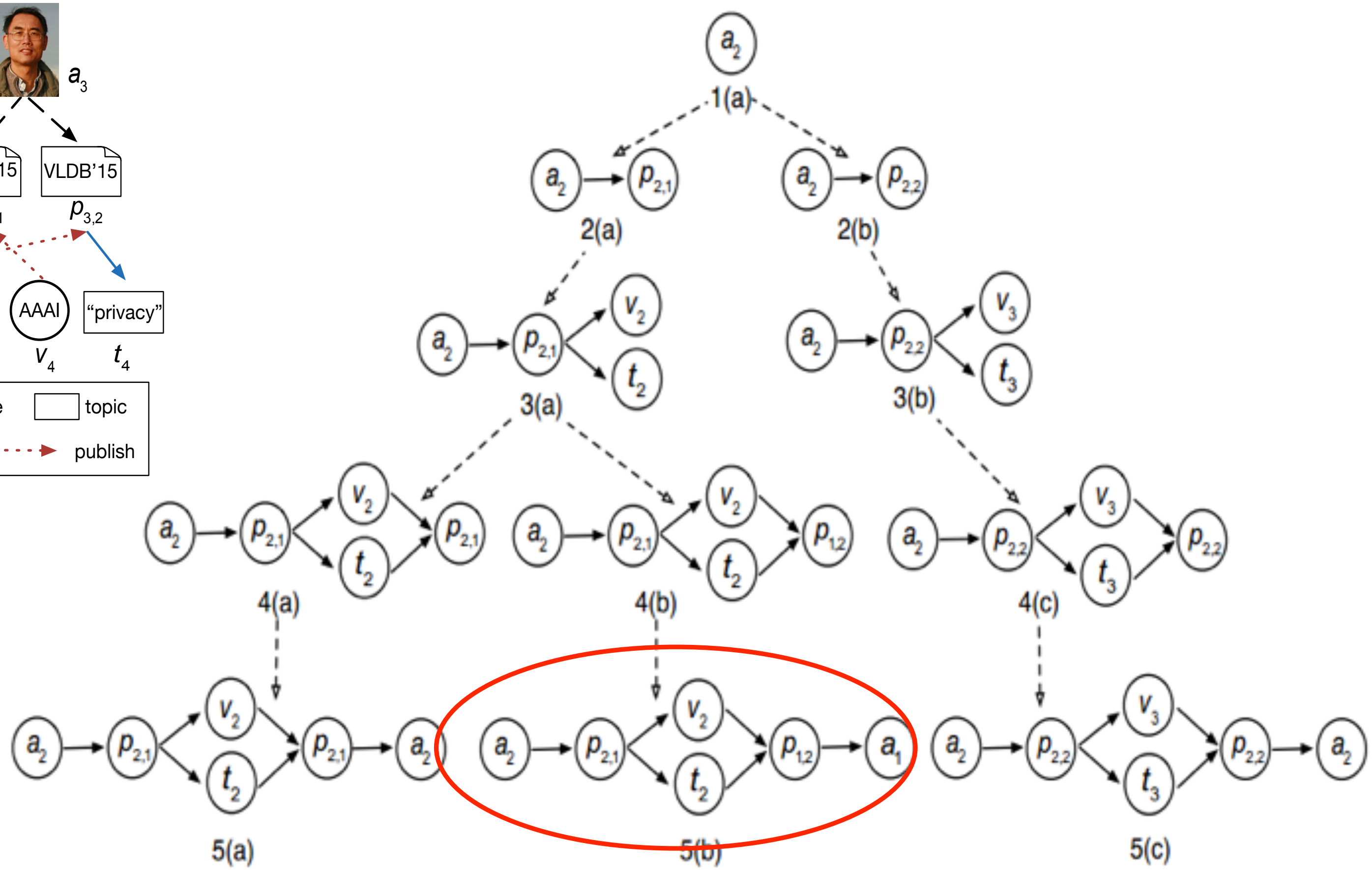



# Relevance Measure 2: SCSE




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



# Relevance Measure 3: BSCSE

- **Biased Structure Constrained Random Walk (BSCSE): extension of BPCRW.**
  - **A combination of SC and SCSE**
  - **SC 0 ← → 1 SCSE**

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

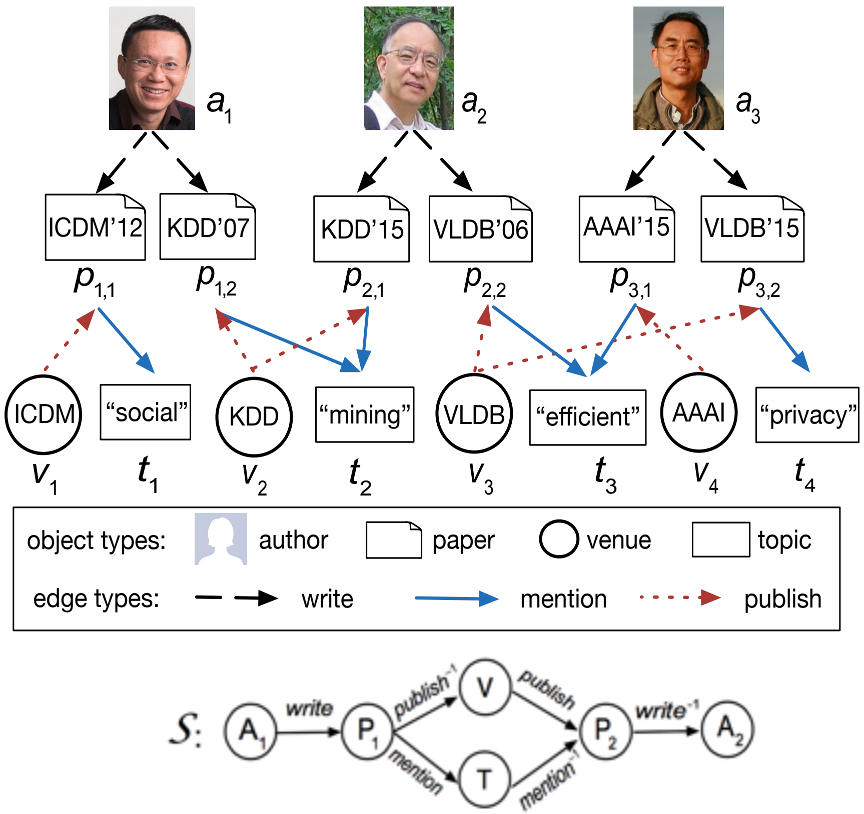
[Huang KDD'16] ZP. Huang "Meta Structure: Computing Relevance on Large Heterogeneous Information Networks" KDD 2016

# Relevance Measures: Summary

| Meta Path | Meta Structure | Meaning |
|---|---|---|
| PathCount | **StructCount** | # of meta-path/structure instances |
| PCRW | **SCSE** | Random walk probability on meta-path/structure |
| BPCRW | **BSCSE** | Combination of count and probability |

# i-LTable

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

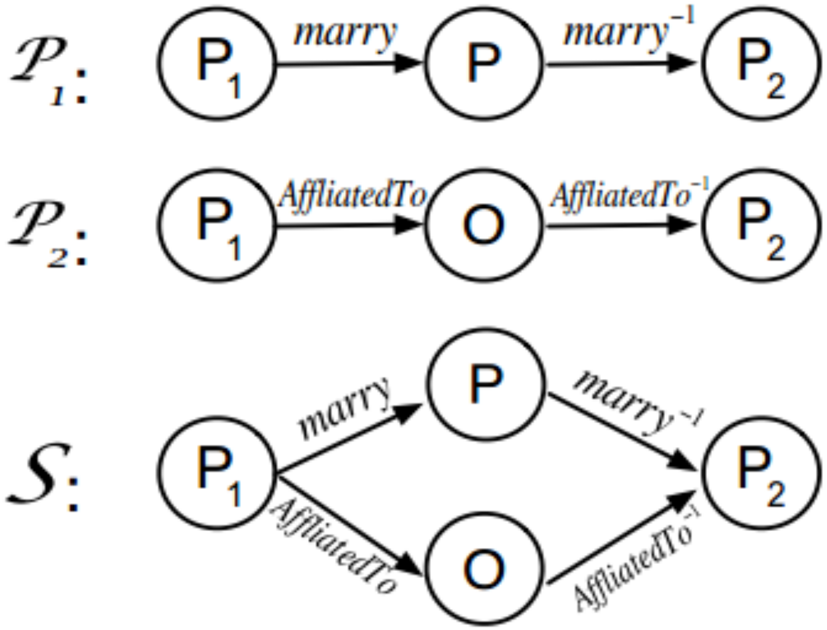



| Key / layer 3 | Value |
|---|---|
| <ICDM, social> | <Pei, 1.0> |
| <KDD, mining> | <Pei, 0.5> |
| | <Han, 0.5> |
| <VLDB, efficient> | <Han, 1.0> |
| <VLDB, privacy> | <Yang, 1.0> |
| <AAAI, efficient> | <Yang, 1.0> |

# Experiment: Entity Resolution

- **On YAGO, we have duplicated entities, e.g., *Barack_Obama* and *Presidency_Of_Barack_Obama***
- **We retrieve the top-k pairs; the high relevance of the node pairs indicates that the nodes are duplicated**
- **Area under PR-Curve (AUC)**

# Experiment: Entity Resolution

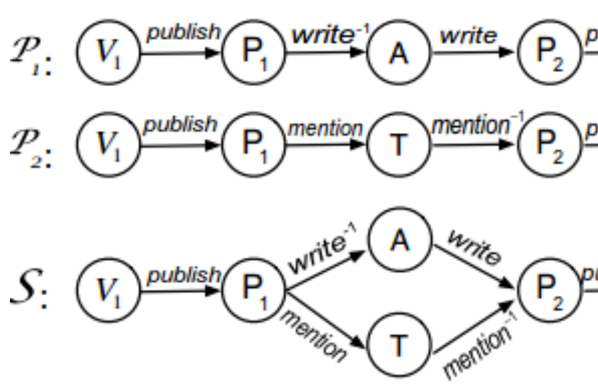



| | P1 | | | P2 | | |
|---|---|---|---|---|---|---|
| Measure | PathCount | PCRW | PathSim | PathCount | PCRW | PathSim |
| AUC | 0.1324 | 0.0120 | 0.0097 | 0.0003 | 0.0014 | 0.0002 |
| | Linear Combination(optimal ) | | | Meta Structure S | | |
| Measure | PathCount | PCRW | PathSim | SC | SCSE | BSCSE*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

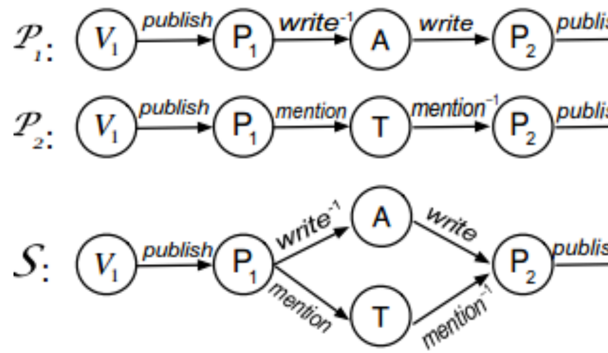
# Relevance Ranking

- We label the relevance of venues in DBLP_4_Area.
- 0 = not relevant; 1 = relevant; 2 = strongly relevant.
  - E.g., <SIGMOD, VLDB>: 2; <SIGMOD, CIKM>: 1
- Normalized Discounted Cumulative Gain (nDCG)

| | $P_1$ | | | $P_2$ | | |
|---|---|---|---|---|---|---|
| sure | PathCount | PCRW | PathSim | PathCount | PCRW | PathSim |
| CG | 0.9004 | 0.9047 | 0.9083 | 0.8224 | 0.8901 | 0.8834 |
| | Linear Combination(optimal ) | | | Meta Structure $S$ | | |
| sure | PathCount | PCRW | PathSim | SC | SCSE | BSCSE* |
| CG | 0.9004 | 0.9100 | 0.9083 | 0.9056 | 0.9104 | **0.9130** |



# Clustering




- Clustering on venues in YAGO
- Normalized Mutual Information (NMI) and Purity

| | $P_1$ | | | $P_2$ | | |
|---|---|---|---|---|---|---|
| Measure | PathCount | PCRW | PathSim | PathCount | PCRW | Pat |
| NMI | 0.4932 | 0.6866 | 0.6780 | 0.3595 | 0.6866 | 0. |
| | Linear Combination(optimal ) | | | Meta Structure S | | |
| Measure | PathCount | PCRW | PathSim | SC | SCSE | BS |
| NMI | 0.4932 | 0.6866 | 0.6780 | 0.3202 | 0.8065 | **0.** |

| | $P_1$ | | | $P_2$ | | |
|---|---|---|---|---|---|---|
| Measure | PathCount | PCRW | PathSim | PathCount | PCRW | Pat |
| Purity | 2.75 | 3.50 | 3.00 | 2.50 | 3.50 | 2 |
| | Linear Combination(optimal ) | | | Meta Structure S | | |
| Measure | PathCount | PCRW | PathSim | SC | SCSE | BS |
| Purity | 2.75 | 3.50 | 3.50 | 2.25 | 3.50 | 3 |

# Outline

- Introduction
  - Motivation
  - Heterogeneous Information Network (HIN)
  - Applications
- Meta-Path
  - Definition
  - Relevance Search
  - Meta-Path Discovery
  - Query Recommendation
- Meta-Structure
  - Definition
  - Relevance Search
- Conclusions & Future Work

# Conclusions

- Relevance of HIN objects can be defined based on meta-paths.
- Query-by-Example can be used to discover meta-paths.
- Meta-structure captures more complex relationships among HIN objects.

# Future Work 1: Efficient Queries on HIN

- Given the complexity of relevance measures, how can we perform graph-based queries on HIN in an efficient and scalable manner?
  - Shortest paths, Top-k, centrality,…
  - Single-disk or cloud-based?

# Future Work 2: Meta-Path/Structure Discovery & Mining

- Design effective and efficient techniques to discover meta structures
- Use meta structures to perform data mining tasks on HINs, e.g., recommendation, classification and clustering.

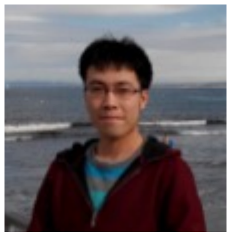
# Future Work 3: HIN and crowdsourcing

- Q1: Can we employ crowdsourcing solutions to discover meta- paths and structures?
- Q2: Can crowdsourcing be used to manage HIN?
- Q3: Can HIN be used to facilitate crowdsourcing? (See our VLDB'17 paper on DOCS)

# Many Thanks!

Dr. Reynold Cheng
URL: http://www.cs.hku.hk/~ckcheng/
email: ckcheng@cs.hku.hk

**Our HKUCS Database Group**



**Zhipeng Huang**

**Yudian Zheng**

**Jing Yan**

**Ka Yu Wong**

**Eddie Ng**

**Jason Meng (Purdue)**