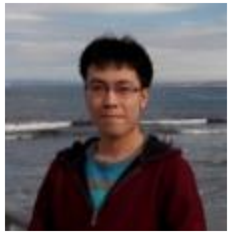
Meta Paths and Meta Structures: Analysing Large Heterogeneous Information Networks

Reynold Cheng



Database Group: Zhipeng Huang, Yudian Zheng, Jing Yan, Ka Yu Wong, Eddie Ng

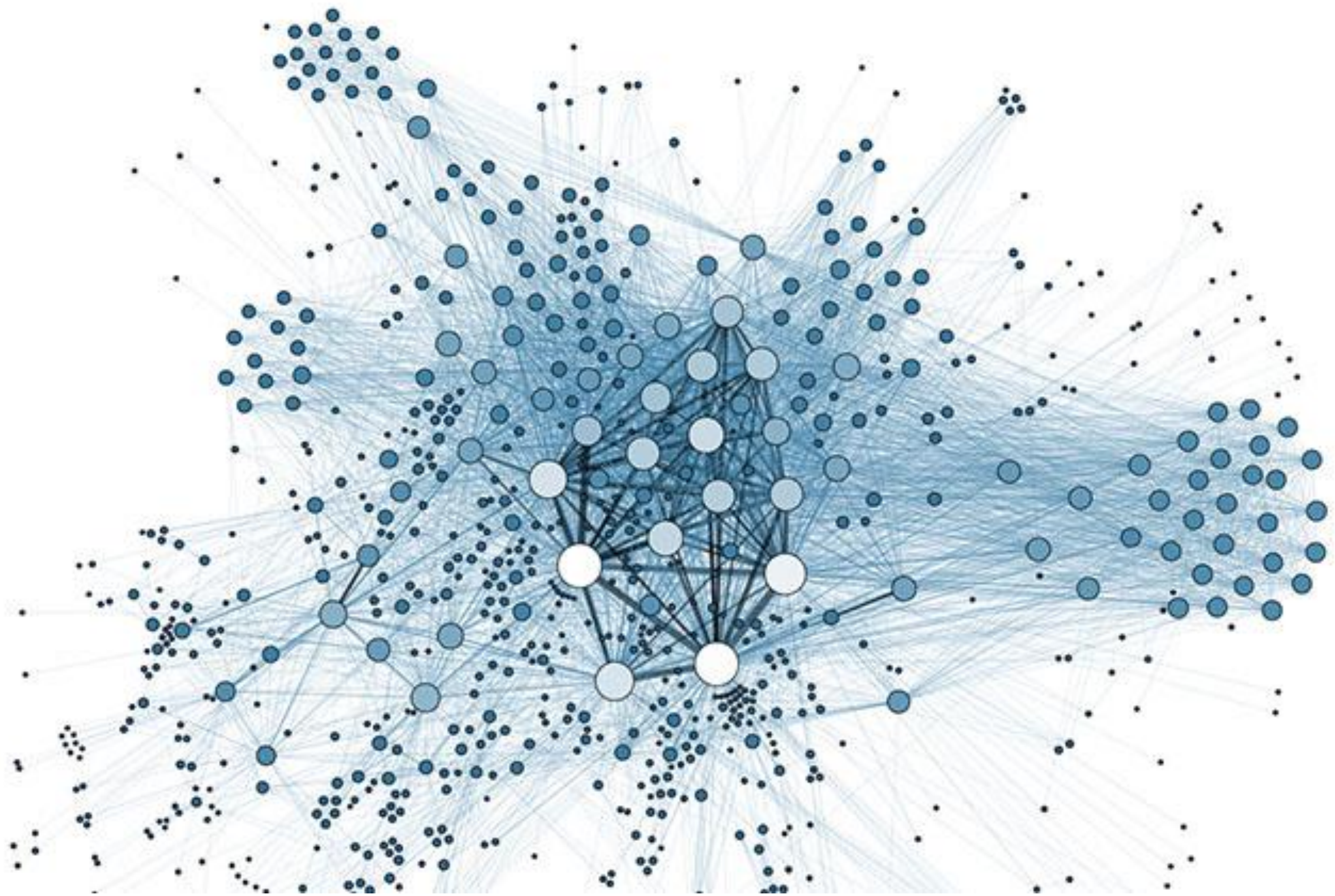
Information is Everywhere !

- Social Networking Websites

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

Information is Everywhere !

- Biological Network



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

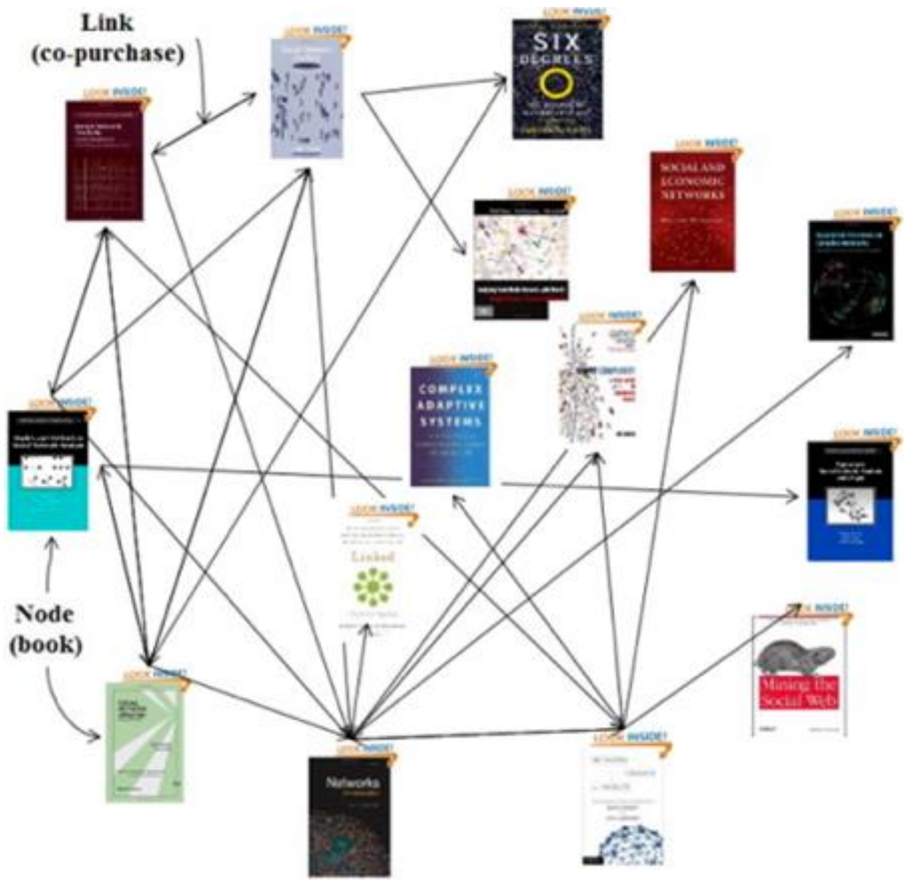
Information is Everywhere !

- **Research Collaboration Network**

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

Information is Everywhere !

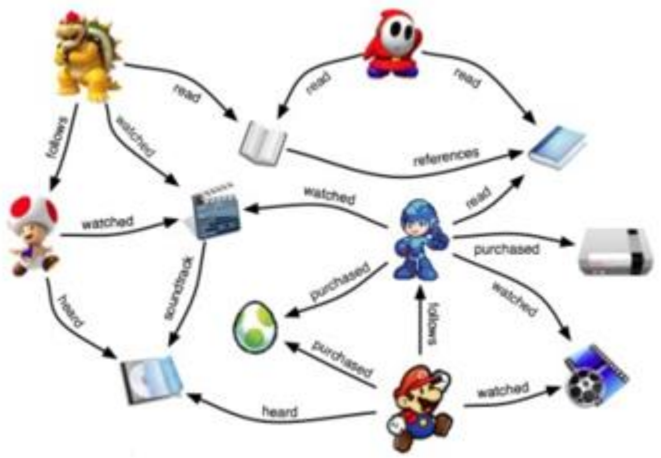
- **Product Recommendation Network**

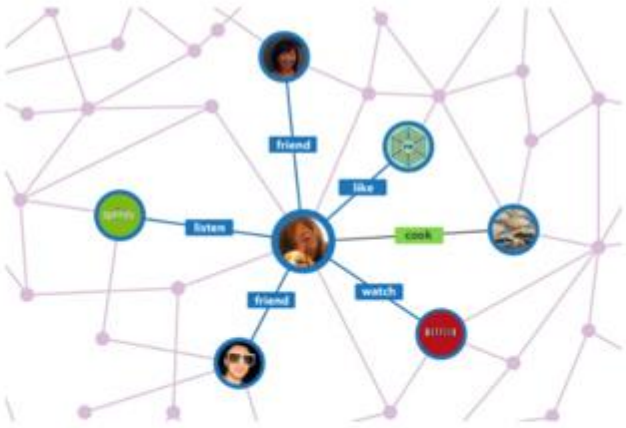


http://www.sciencedirect.com/science/article/pii/S0957417413006921
Byunghak Leem. Heuiju Chun. An impact of online recommendation network on demand

The Real World

- Heterogeneous Information Network(s), i.e. HIN(s).



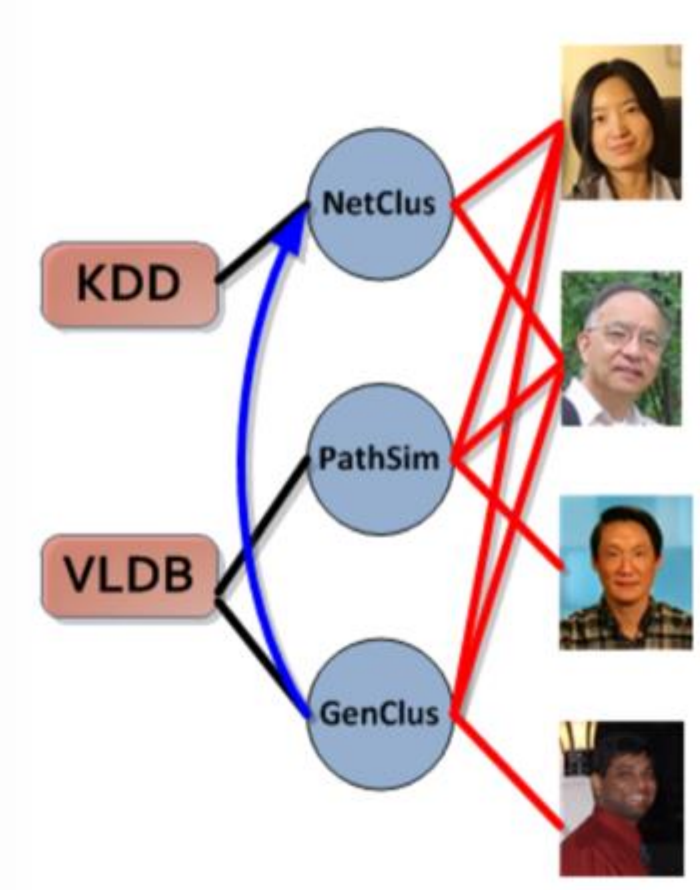


- Networks: Nodes & Links
  - Nodes: Various Types
  - Links: Various Types

Example HINs

- DBLP Bibliographic Network

- Networks: Nodes & Links
  - Node (Type):
    - KDD (Venue)
    - Jiawei Han (Author)
  - Link (Type):
    - Write (Author → Paper)
    - Publish (Paper → Venue)

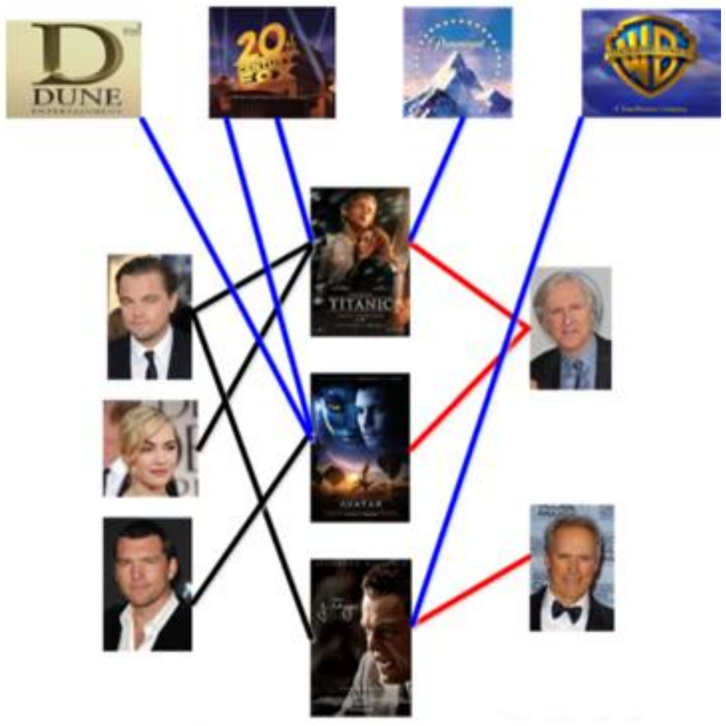


Jiawei Han. A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks.

Example HINs

- The IMDB Movie Network

- Networks: Nodes & Links
  - 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**

- **Networks:**
  - **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.

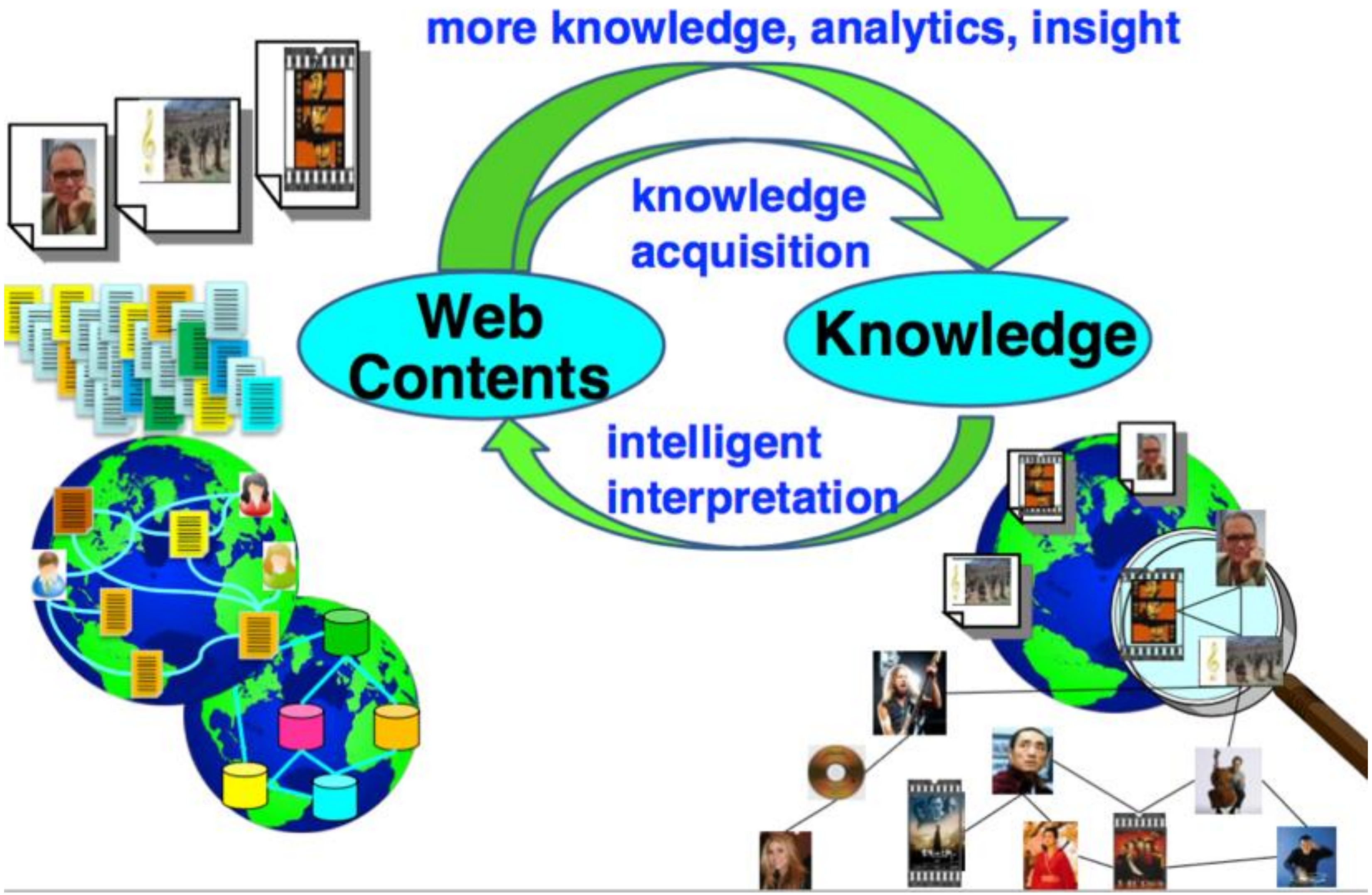
HINs are Ubiquitous !

- 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.

Knowledge Graph (KG)

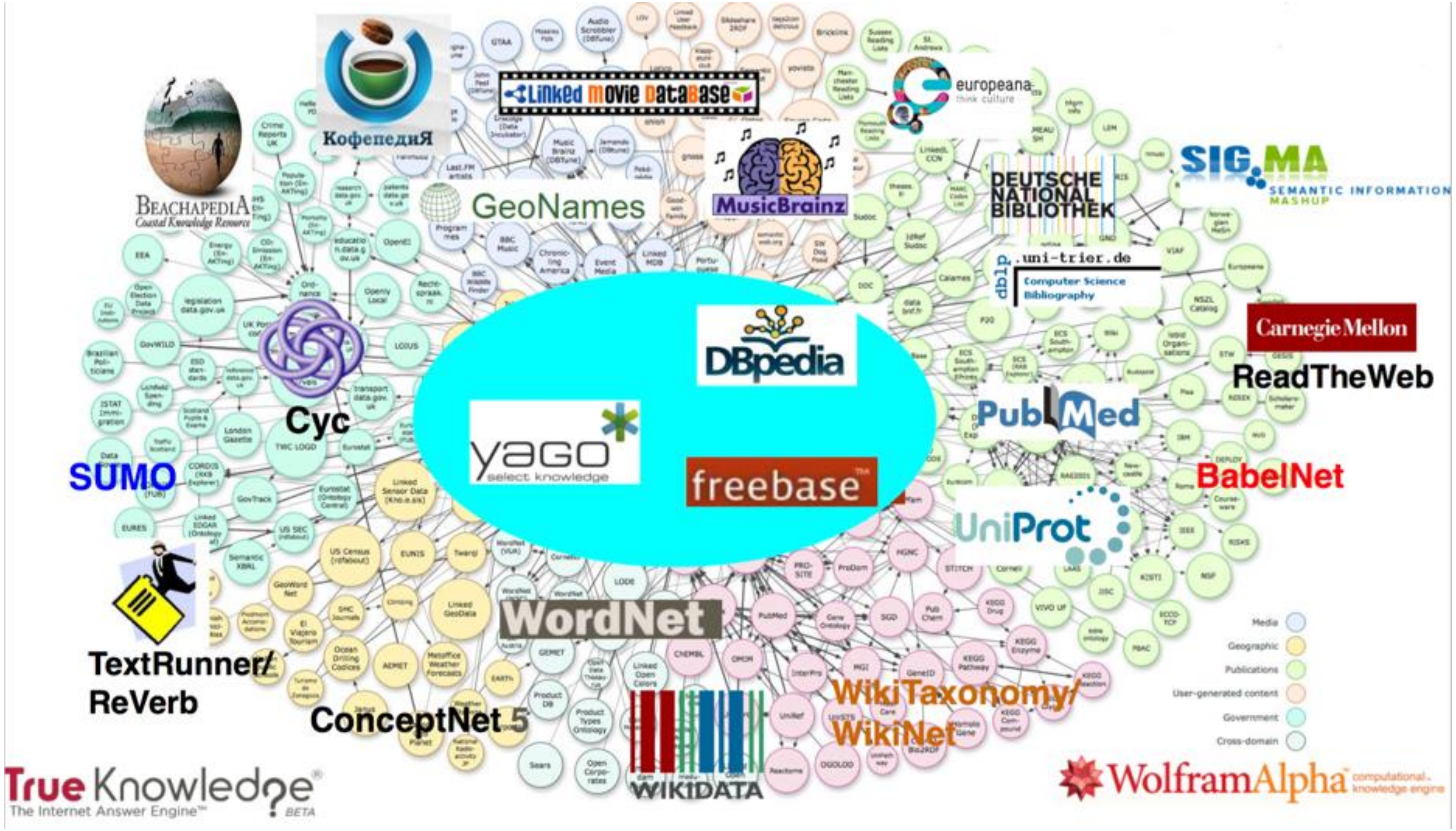
- Turn Web Knowledge into KG



Gerhard Weikum. Knowledge Graphs: from a Fistful of Triples to Deep Data and Deep Text.

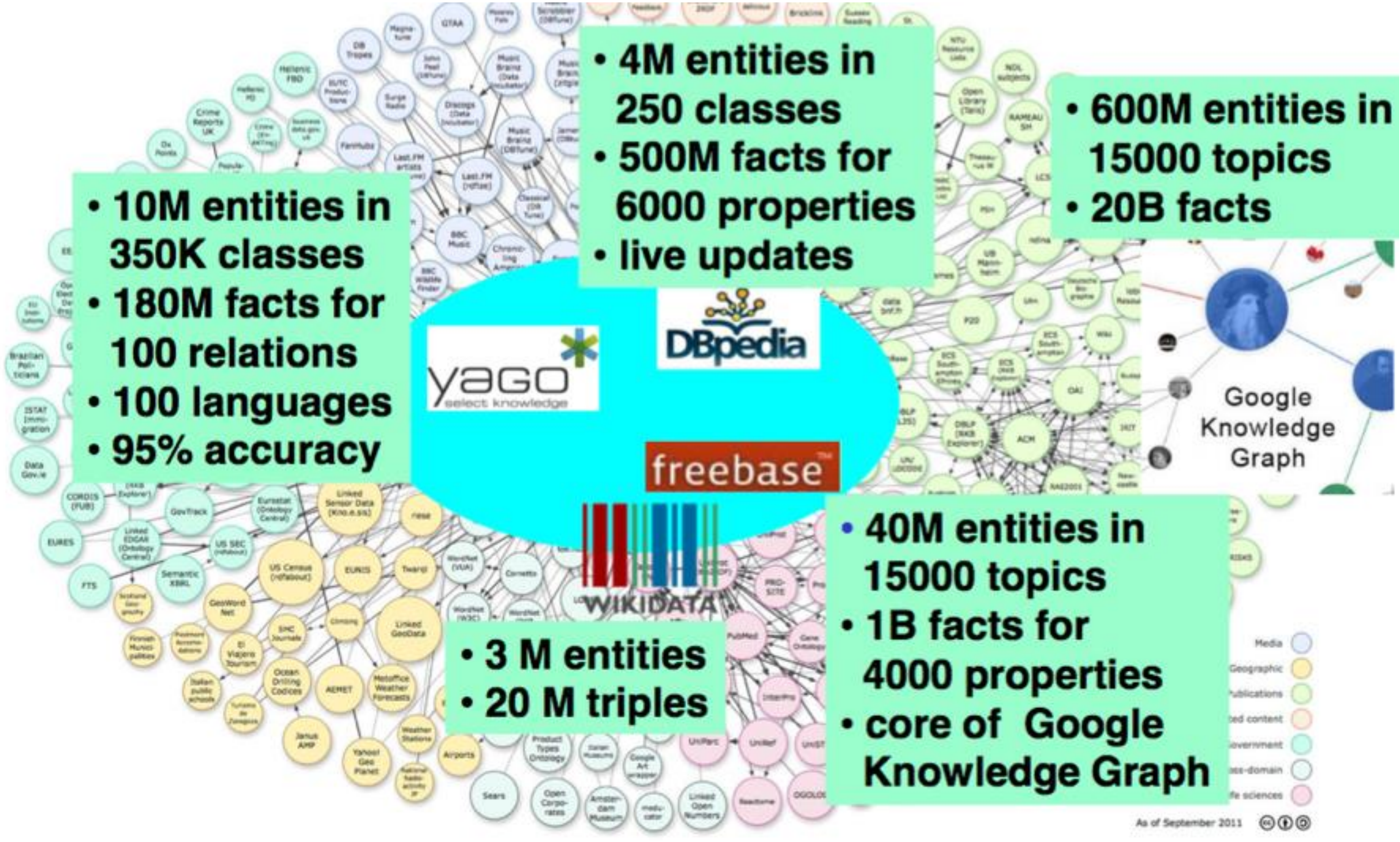
Knowledge Graph (KG)

- **Example KGs**



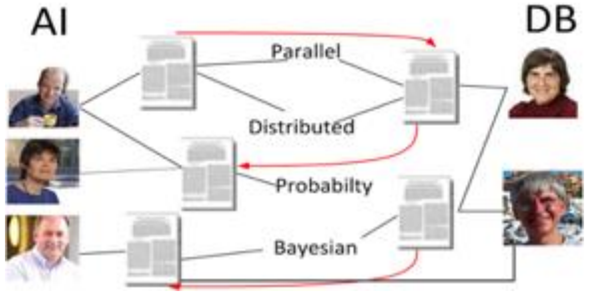
Knowledge Graph (KG)

- **Statistics in Existing KGs**



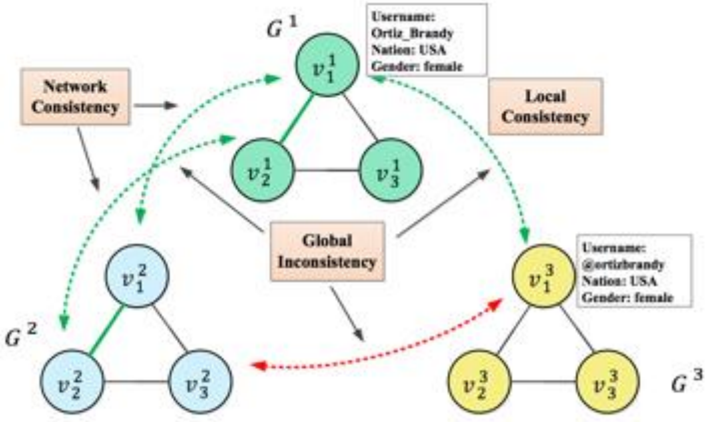
Problems in HIN

- **Link Prediction**



- **Entity Profiling**

- **Data Integration**

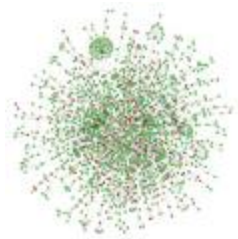


Yangqiu Song. Recent Development of Heterogeneous Information Networks: From Meta-paths to Meta-graphs
Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S. Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency, KDD 2015.

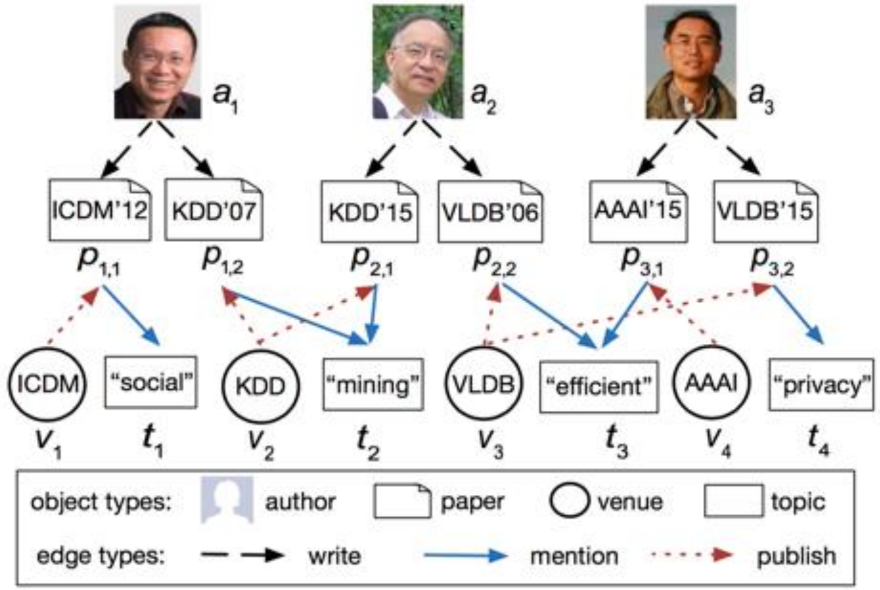
Overview of the Tutorial

- **Relevance Search**

  Find **Similar/Relevant** Objects in Networks



- **Examples**



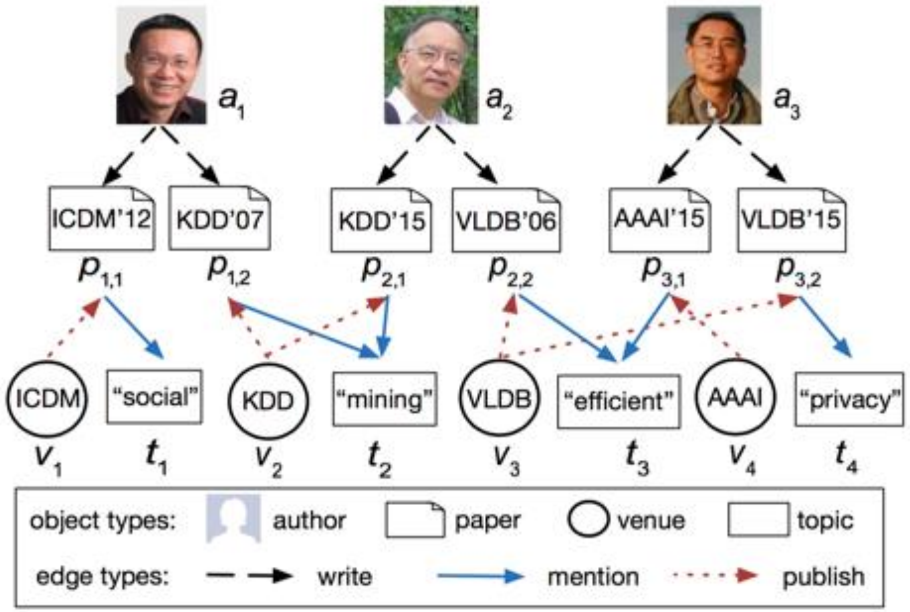
DBLP[1]

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

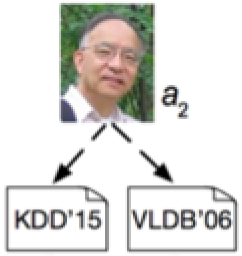
[1] http://dblp.uni-trier.de/

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

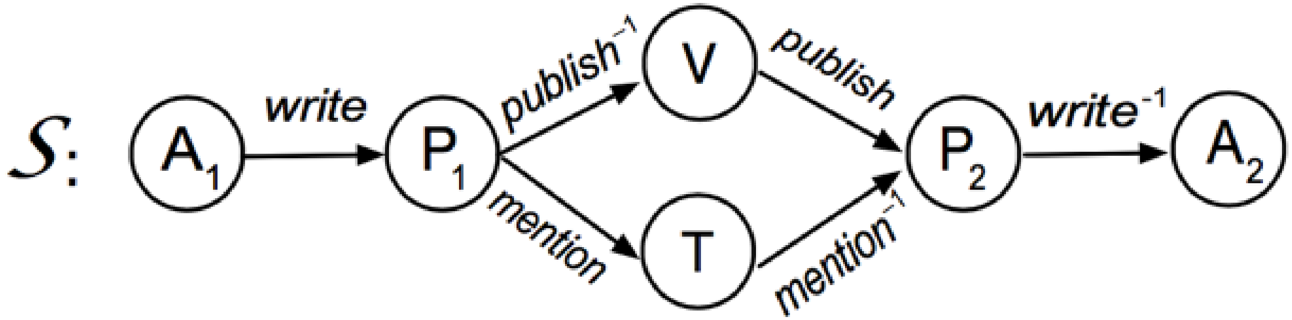
- Query Recommendation: to suggest alternate relevant queries to a search engine user

- 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.

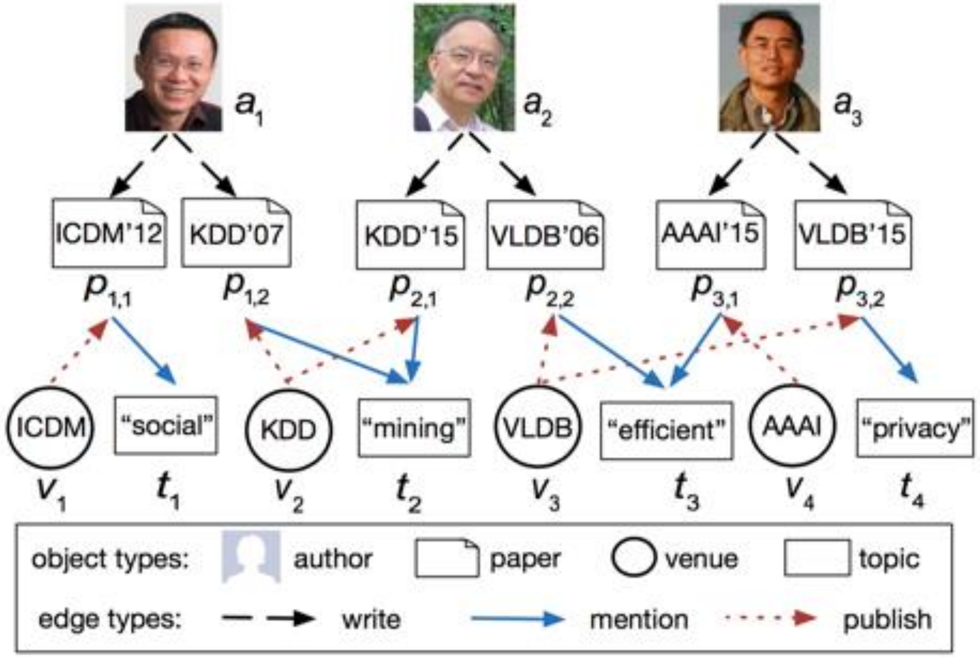
Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, Xiang Li. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. KDD 2016.

Outline

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

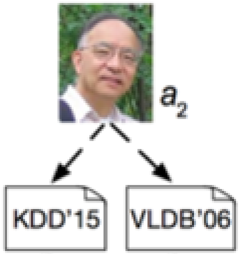
Definition of Meta-Path

- Definition [Sun et al. VLDB 2011]



- Example

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



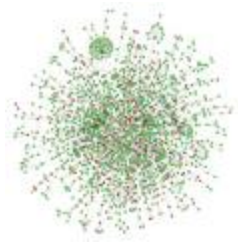
Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. VLDB 2011.

Outline

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

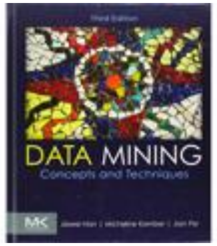
Relevance Search

○ Motivation



Find **Similar/Relevant** Objects in Networks

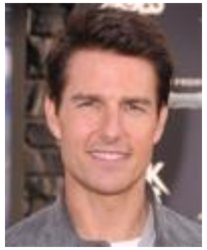
○ Examples



DBLP[1]

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

IMDb[2]



- **Who** are most similar to *Tom Cruise* ?
- **Which movie** is most relevant to *Tom Cruise*?

[1] http://dblp.uni-trier.de/
[2] http://www.imdb.com/

Relevance Search

- **Target**

  To answer these questions systematically

- **Solutions**

  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

SimRank

○ Model

Idea: Two objects are similar if they are referenced by similar objects

○ Definition

- ***S(a,b)*** = **Average similarity** between in-neighbors of object *a* ***I(a)*** and in-neighbors of object *b* ***I(b).*** Between [0, 1].

- ***S(a,b)*** = 1, if a=b

$$= s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) \quad \text{, if } a \neq b$$

where c is the constant and 0<c<1

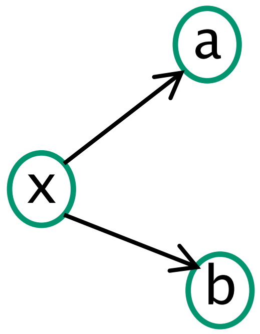
[Jeh, Glen, and Jennifer Widom. KDD'02] Jeh, Glen "SimRank: a measure of structural-context similarity."

SimRank

- **Example**

$$S(a,a) = 1$$

$$S(a,b) = \frac{c}{1 \times 1} \times 1 = c$$



- S(a,b) **ideally** should be 1.
- But, in reality the graph does **not describe everything** about them, so by using the C to make s(a,b)<1. Adding C is to expresses limited confidence or decay with distance.

[Jeh, Glen, and Jennifer Widom. KDD'03] Jeh, Glen "SimRank: a measure of structural-context similarity."

Personalized PageRank (PPR)

○ Model

Idea: Originally defined by Google as
a measure of importance for web-pages.

○ Definition

- Given a graph G, a **starting source** node **s**, a **target** node t, and a teleport probability α. Perform random walk from s. At each step **stop with the probability** α, otherwise continue **performing random walk**.

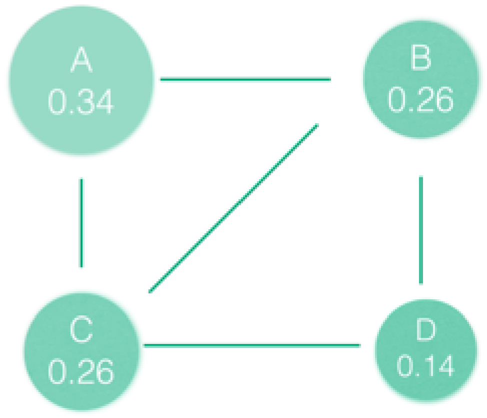
- Then the Personalized PageRank from s to t is

$$\mathrm{PPR}_{s\sim t} = \mathrm{P}(\boldsymbol{s} \ \rightarrow \boldsymbol{t})$$

[Jeh, Glen, and Jennifer Widom. WWW'02] Jeh, Glen "Scaling personalized web search."

Personalized PageRank (PPR)

○ Example



Starting from A, and $\alpha = 0.2$

For each target A, B, C, D

○ Calculation

Iterative computation (Power Method);

Monte-carlo simulation (Approximation);

Bookmark Coloring Algorithm, and etc…

Path Constrained Random Walk

- **Model**

  Random walk on given paths.

- **Definition**
  - Performing random walks on given meta-paths with the fixed starting point and target point.
  - **PCRW**: Transition probability of the random walk following **a given meta-path**.

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

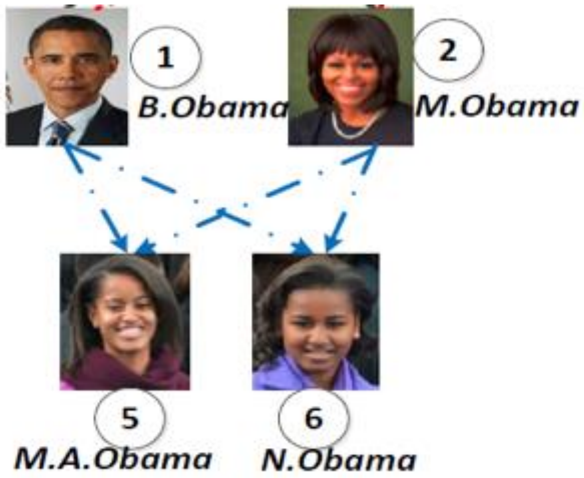
  - Between [0, 1].

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

Path Constrained Random Walk

- **Example**

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



m_1 = P1 -> P2 -> P3

1. Pro(B.Obama | P1)=1
2. Pro(M.A. Obama | P2) = Pro(B.Obama | P1) / 2 = 0.5

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

   Pro(B.Obama | P3) = Pro(M.A. Obama | P2) 2 + Pro(N.Obama | P2) /2 = 0.5

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

PathSim

- **Model**

  Path Counts (PC):

  #paths following a given meta-path

- **Definition**

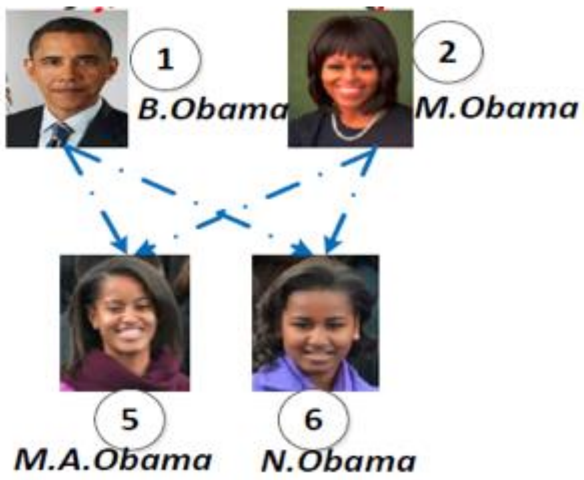
- Can only be applied on **symmetric** meta paths (consider the node type and link type)
- **Normalized** version of PC. Between [0, 1].
- $\mathrm{PathSim}(s, t \mid m) = \frac{2 \times \mathrm{PC}(s,t|m)}{\mathrm{PC}(s,s)+\mathrm{PC}(t,t)}$

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

PathSim

- **Example**

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



PC(B.Obama, M.Obama)=2

PC(B.Obama, B.Obama)=2
PC(M.Obama, M.Obama)=2

PS(B.Obama,M.Obama)=2*2/(2+2) =1

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

HeteSim

○ **Model**

Improvement of SimRank for Heterogeneous Information Network

○ **Definition**

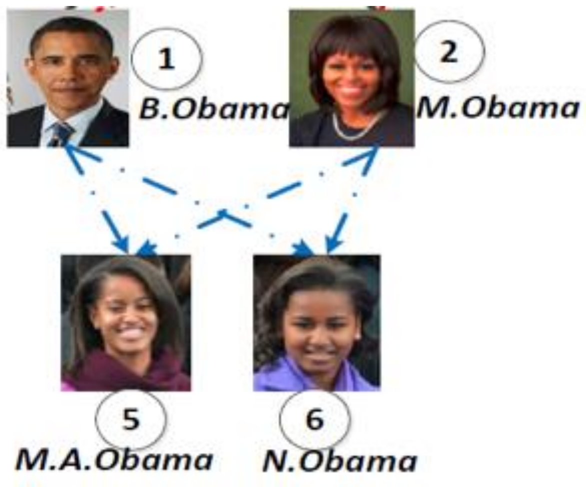
- Any **arbitrary** meta paths.
- Given relations $P = R_1 \circ R_2 \circ \cdots \circ R_l,$

$$HeteSim(s, t|R_1 \circ R_2 \circ \cdots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j(t|R_l)|R_2 \circ \cdots \circ R_{l-1})$$

[Shi, Kong, Huang TKDE'2014] Hetesim: A general framework for relevance measure in heterogeneous networks.

HeteSim

- **Example**



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

$\boldsymbol{m_1}$= P1 -> P2 -> P3

HeteSim (B.Obama, M.Obama$|m_1$)=

$$\frac{1}{|O_{B.Obama}|+|I_{M.Obama}|}\left(HeteSim\,\textbf{\textit{(M.A.Obama, M.A.Obama)+Hetesim(N.Obama, N.Obama)}}\right)$$

$$\boldsymbol{=\frac{1}{(2\times 2)}(1+1)=0.5}$$

[Shi, Kong, Huang TKDE’2014] Hetesim: A general framework for relevance measure in heterogeneous networks.

Comparison

- For PathSim, HeteSim and PCRW, even for the same example they have different values.

- These metrics are designed for different applications or measurement scenarios.

- No dominating similarity measurements so far.

Other Measurements

- **KnowSim (APWeb’14)**

  Measure similarity between nodes by RWs on given meta-path and the reverse meta-path respectively.

- **AvgSim (ICDM’16)**

  Measure the similarity of Documents by modeling them into heterogeneous information networks.

- **RelSim (SDM’16)**

  Measure the similarity relations in heterogeneous information network.

…

Summary

	Structure-based	Semantic-based	Symmetric?
SimRank	√		Yes
PPR	√		Yes
PCRW		√	No
PathSim		√	Yes
HeteSim		√	Yes
…			

Outline

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

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 ?

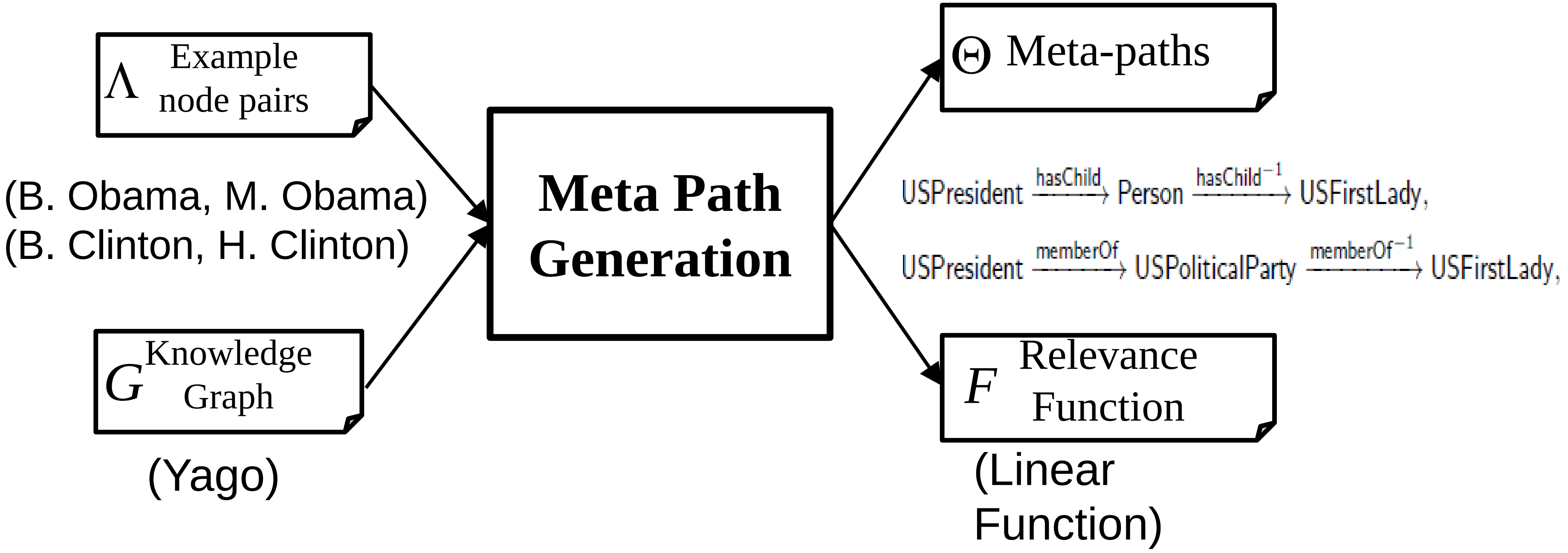
Our Contributions (WWW'15)

- **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

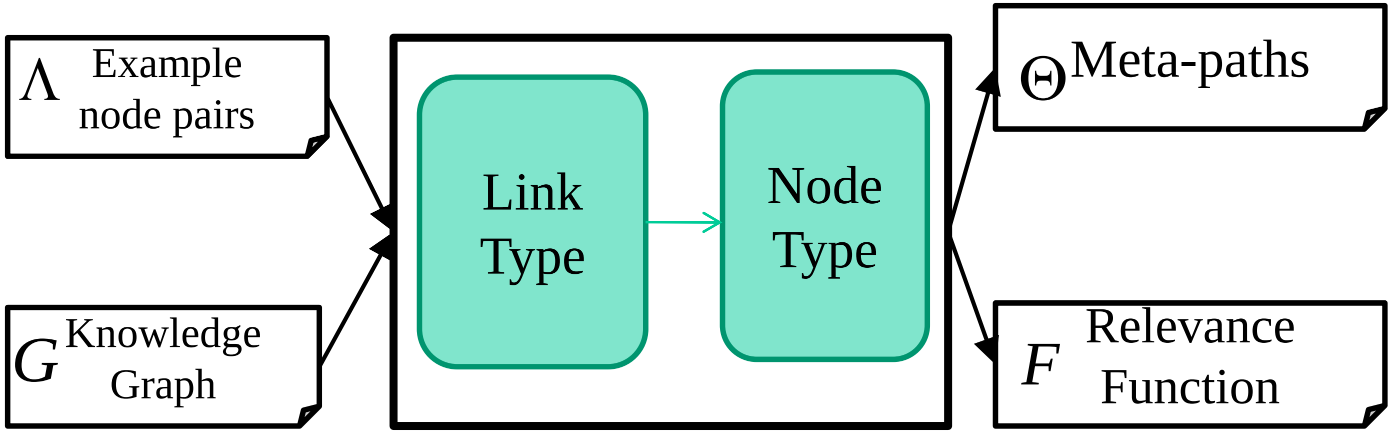
- **Framework**



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

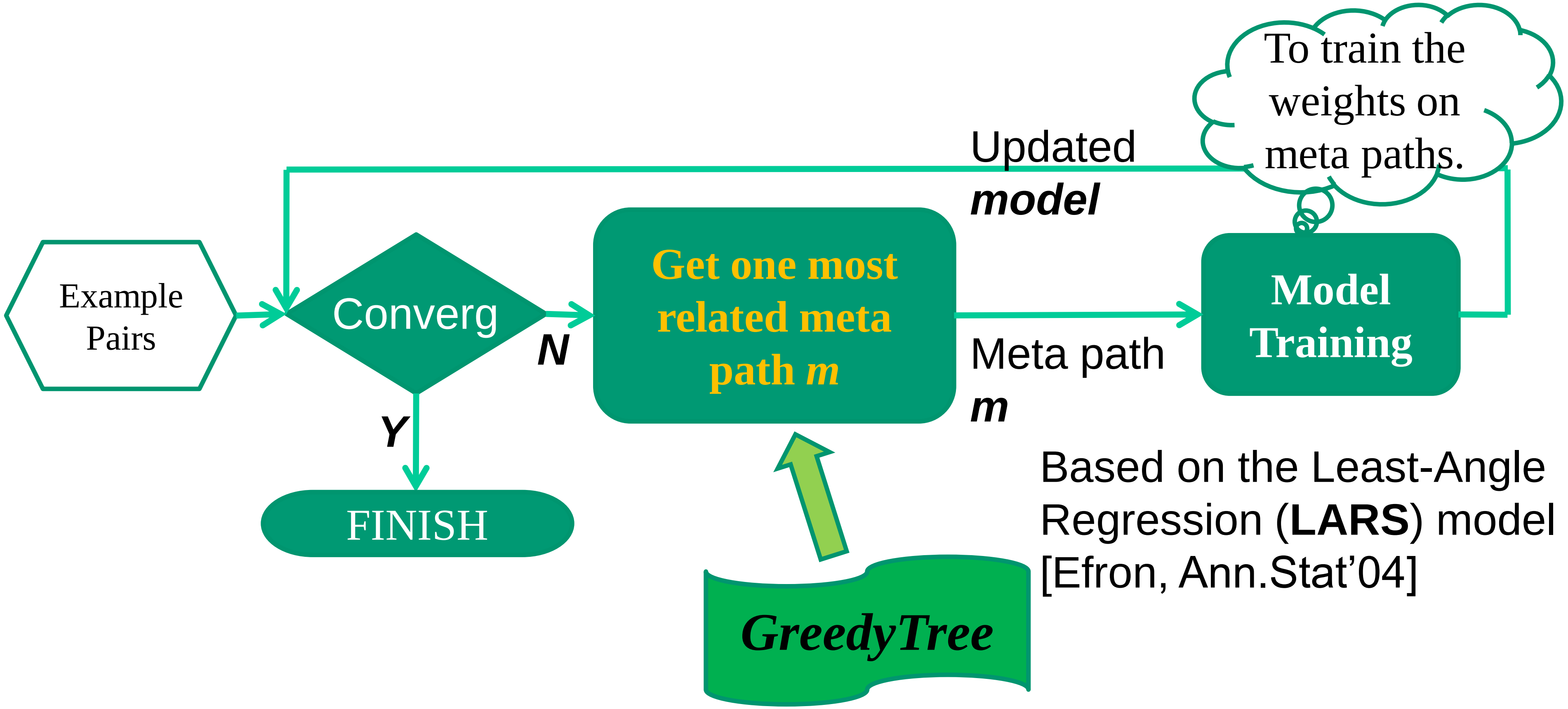
Generating Meta-Paths

- In Two Phases



Phase 1: Link-Only Path Generation

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



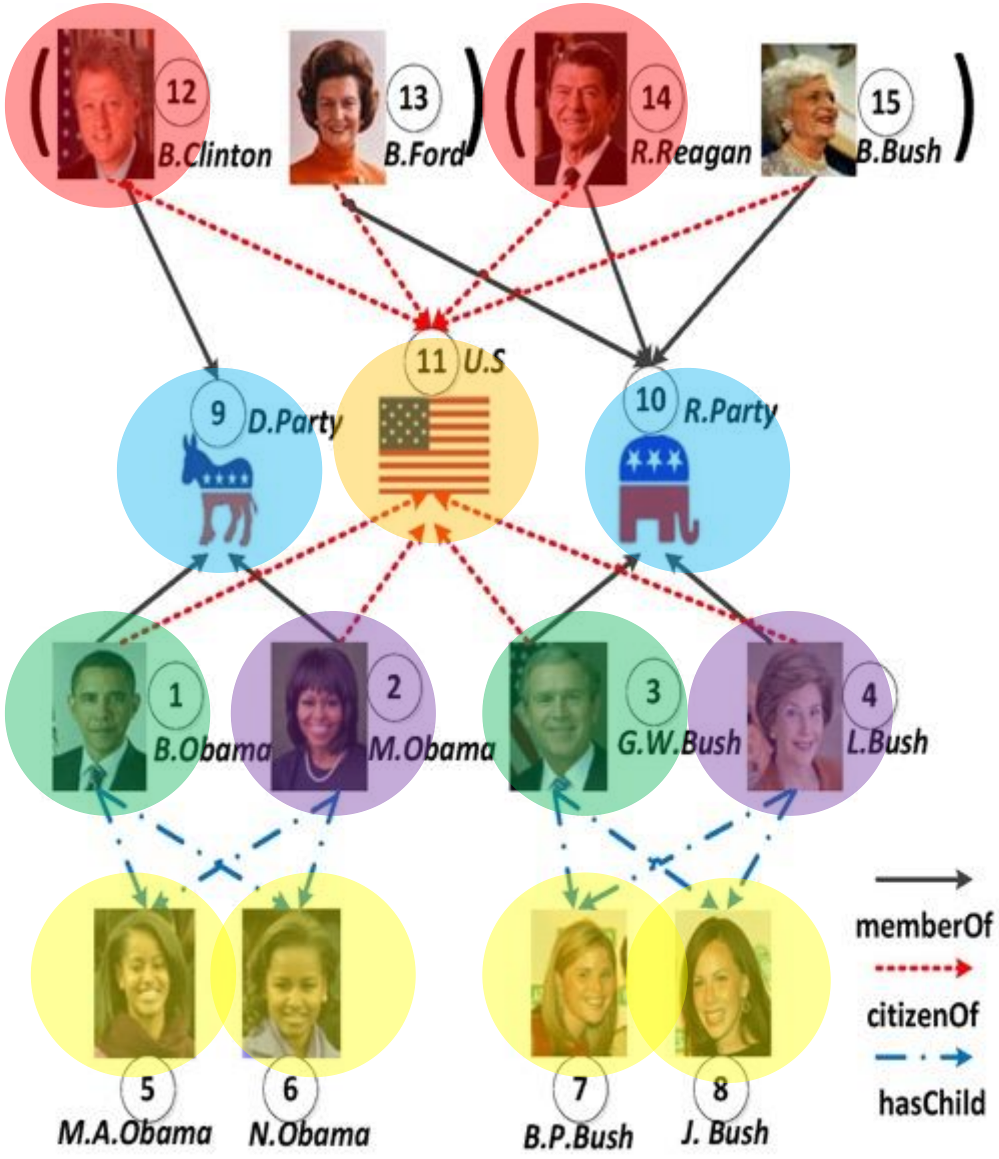
Phase 1: Link-Only Path Generation

- **GreedyTree**
  - **A tree that greedily expands the node which has the largest priority score**

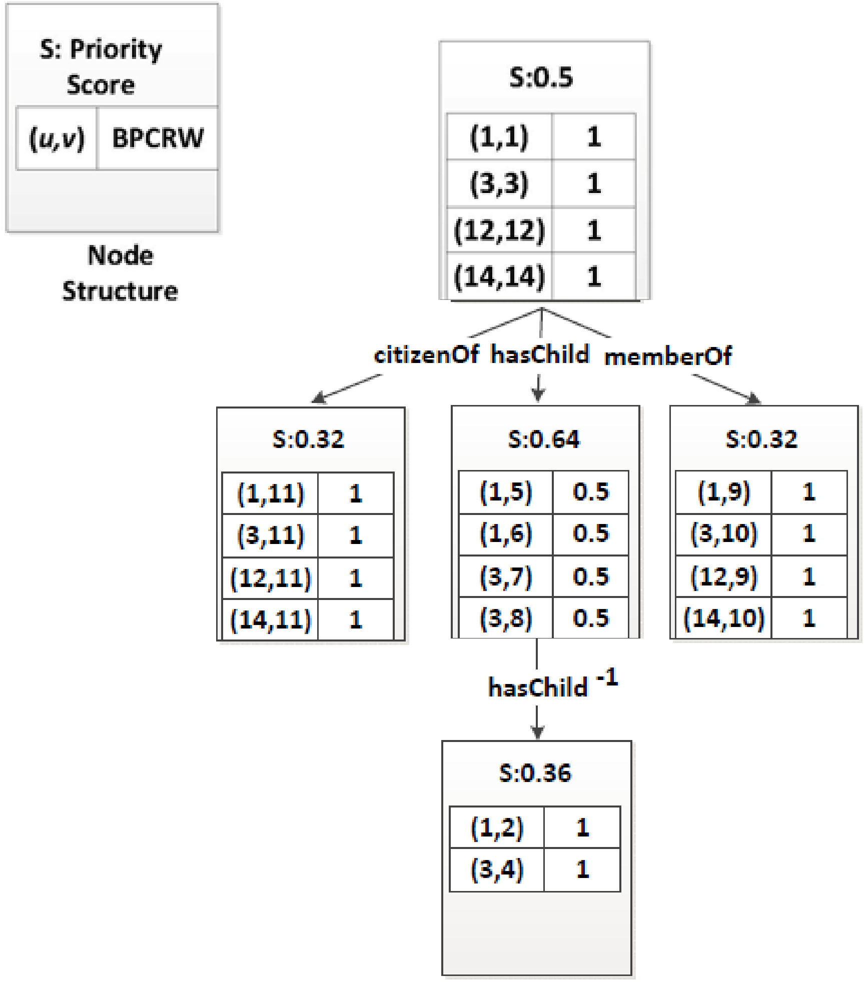
  - **Priority Score : related to the correlation between *m* and *r***
    - ***m* is the vector expression of a meta path, *r* is the residual vector which evaluates the gap between the truth and current model**

$$\cos(\mathbf{m}, \mathbf{r}) = \frac{\mathbf{m} \cdot \mathbf{r}}{\|\mathbf{m}\| \times \|\mathbf{r}\|} \qquad S = \frac{\sum_{u+} \sigma(u, v \mid \Pi) \cdot \mathbf{r}(u, *)}{\sqrt{\sum_{u} \sigma(u, v \mid \Pi)^2} \times |\mathbf{r}|} \cdot \beta^L$$

Phase 1: Link-Only Path Generation



GreedyTree

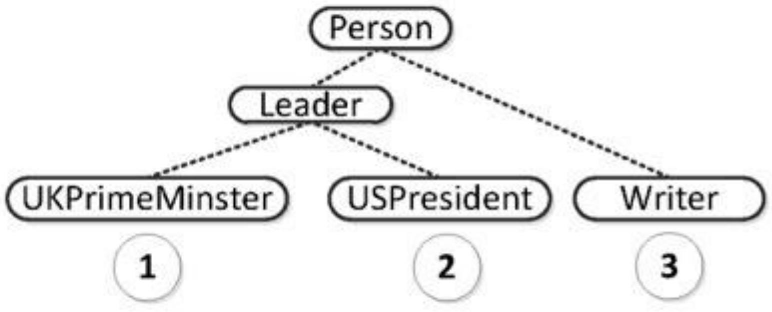


Phase 2: Node Class Generation

- **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**

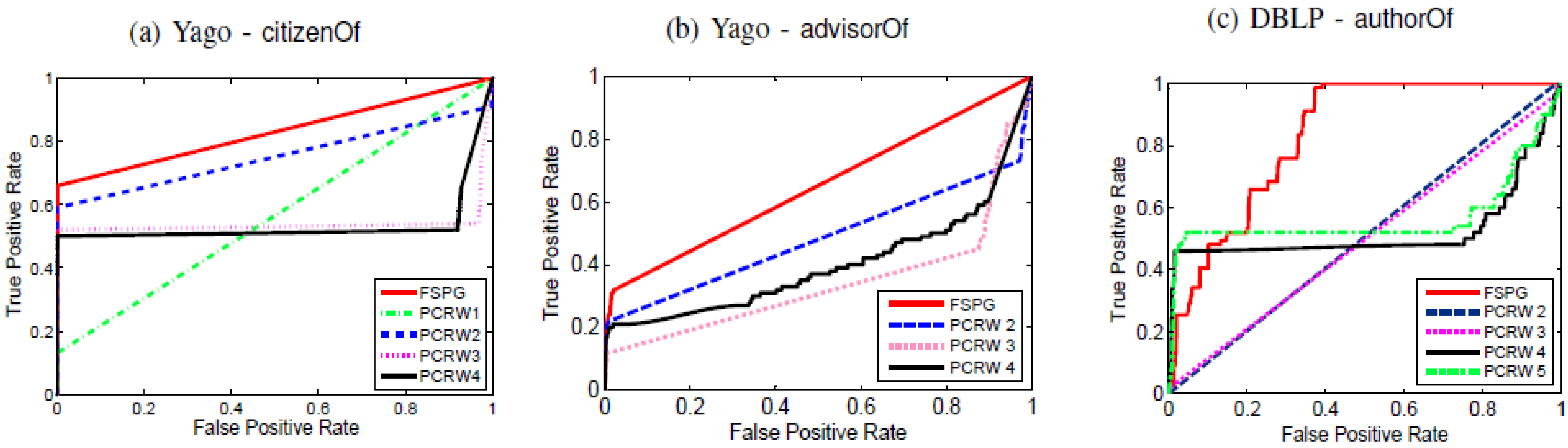


Experiments

- **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.**

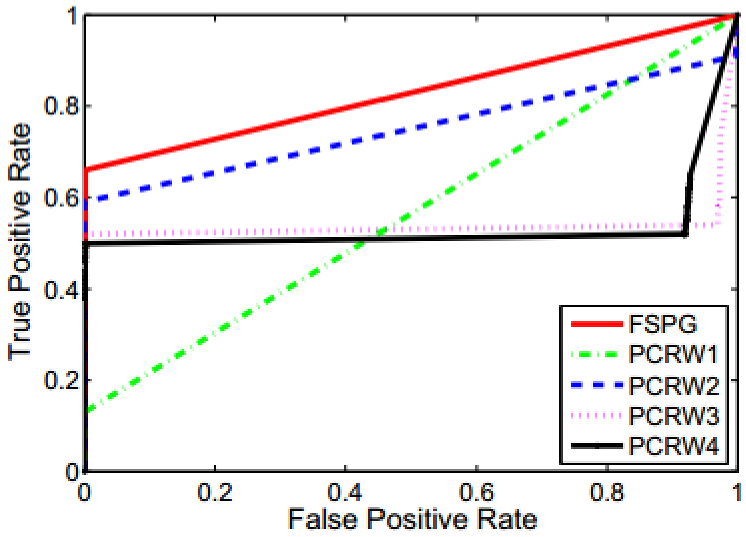


ROC for link prediction

Experiment 2

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

(a) Yago - citizenOf

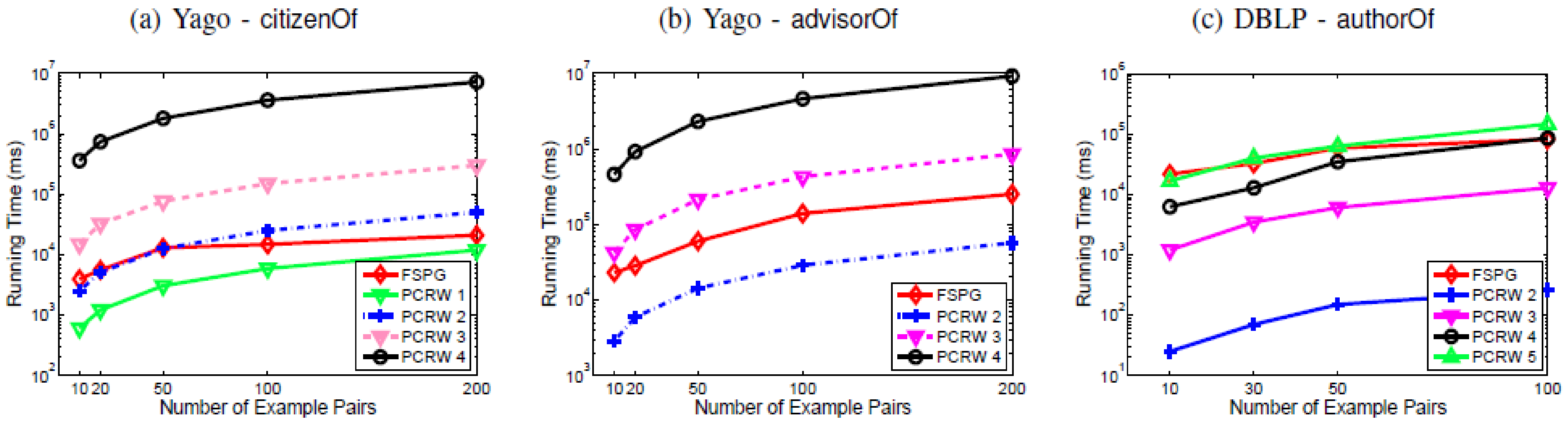


meta-path	w
Person $\xrightarrow{\text{bornIn}}$ City $\xrightarrow{\text{locatedIn}}$ Country	5.477
Person $\xrightarrow{\text{livesIn}}$ Country	0.361
Person $\xrightarrow{\text{graduateOf}}$ University $\xrightarrow{\text{locatedIn}}$ Country	0.023
Person $\xrightarrow{\text{diedIn}}$ City $\xrightarrow{\text{locatedIn}}$ 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

Outline

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

One Application

- Query Recommendation: to suggest alternate 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.

Long Tail Distribution

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

Motivation

- **Ubiquitous:**
  - **84% of 10M queries appear no more than 3 times.**
- **Necessary:**
  - **Existing works that only rely on query log alone can no longer handle well these queries.**

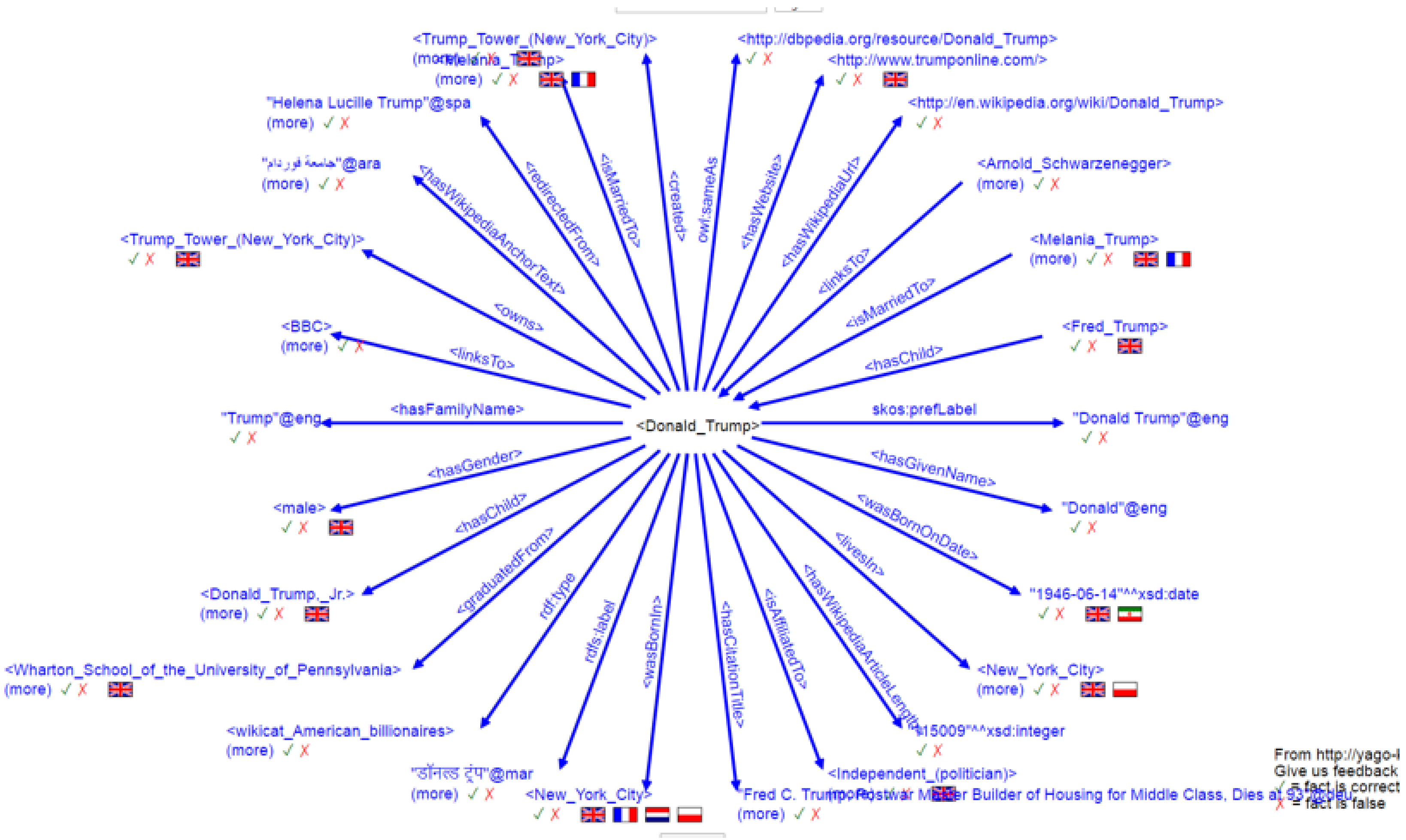
Query Log

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

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.

Knowledge Graph



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

Relationship in the KG

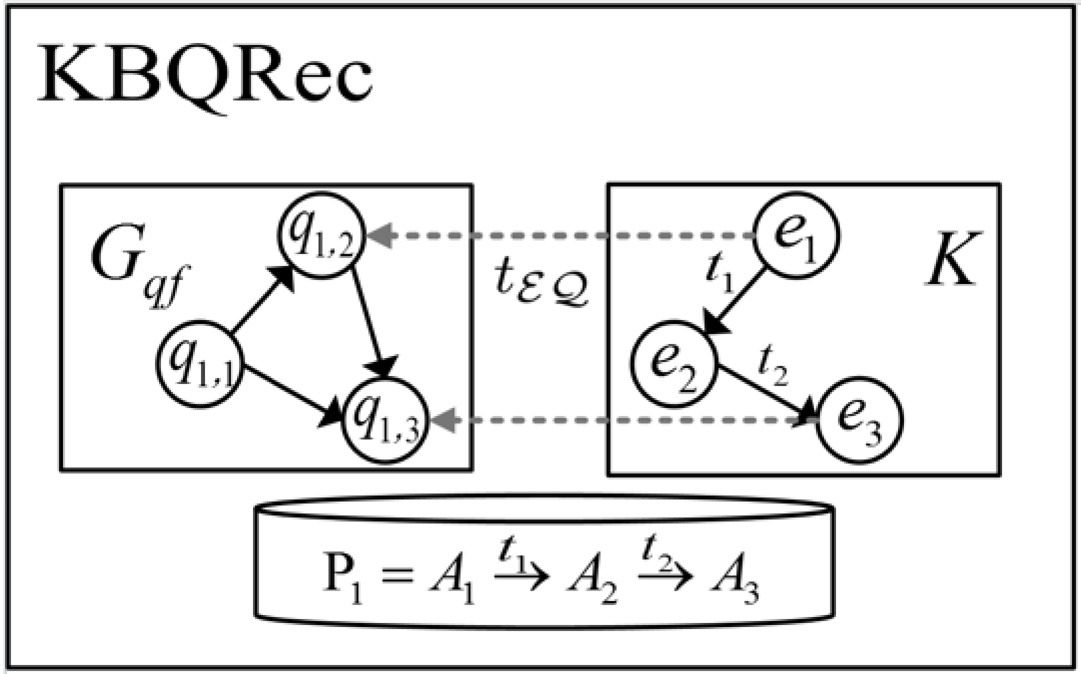
- **Meta path representation:**
  - **P: city $\xrightarrow{\text{nextTo}}$ city**

- **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

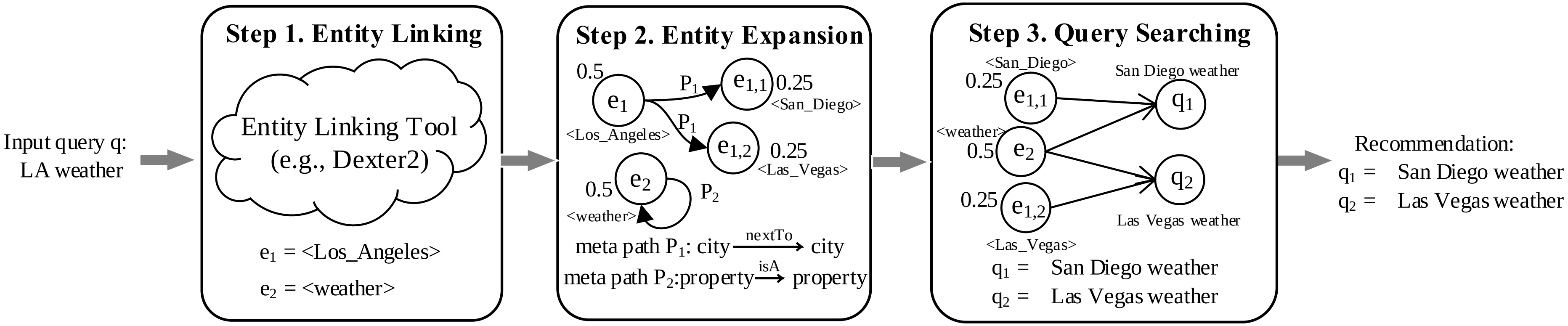


Offline

- G_{qf} is built as described in [1].
- t_{eq} is built from entity linking and normalizing the weights.
- P:
  - Get the set of entity pairs within the same session: $\left\{\left(e_i, e_j\right) \middle| e_i, e_j \in s_k\right\}$
  - Get the meta path between e_i and e_j (we use the shortest path for simplicity)
  - Stored by the type of e_i

Online

- Three Steps:
  – Entity Linking (use existing tool)
  – Entity Expansion (use P)
  – Query Searching (PPR)



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 NextTo city
- Return
  - e_2 = Las_Vegas
  - e_3 = San_Diego

Step 3. Query Searching

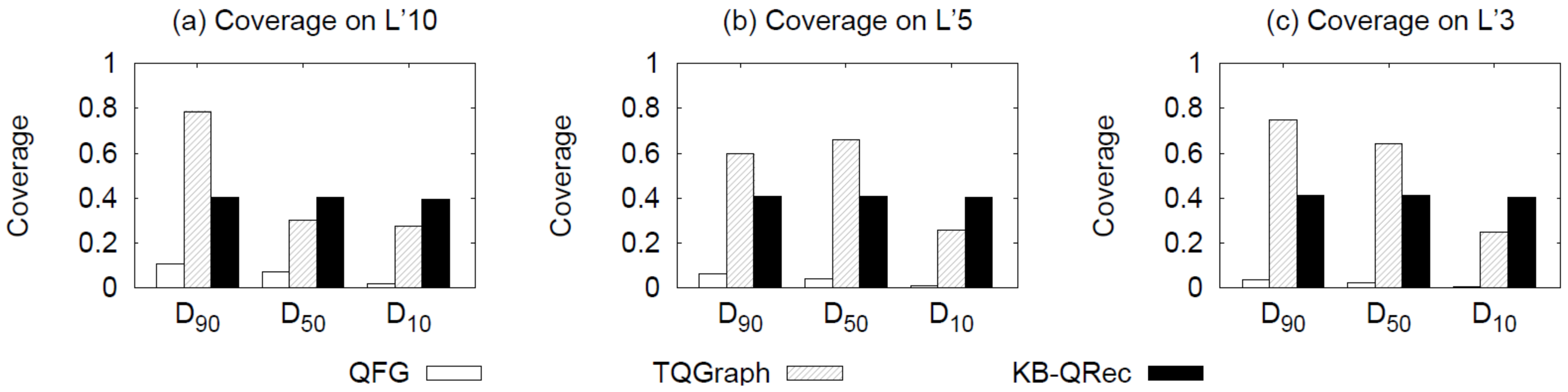
- **Given:**
  - **e_2 = Las_Vegas**
  - **e_3 = San_Diego**

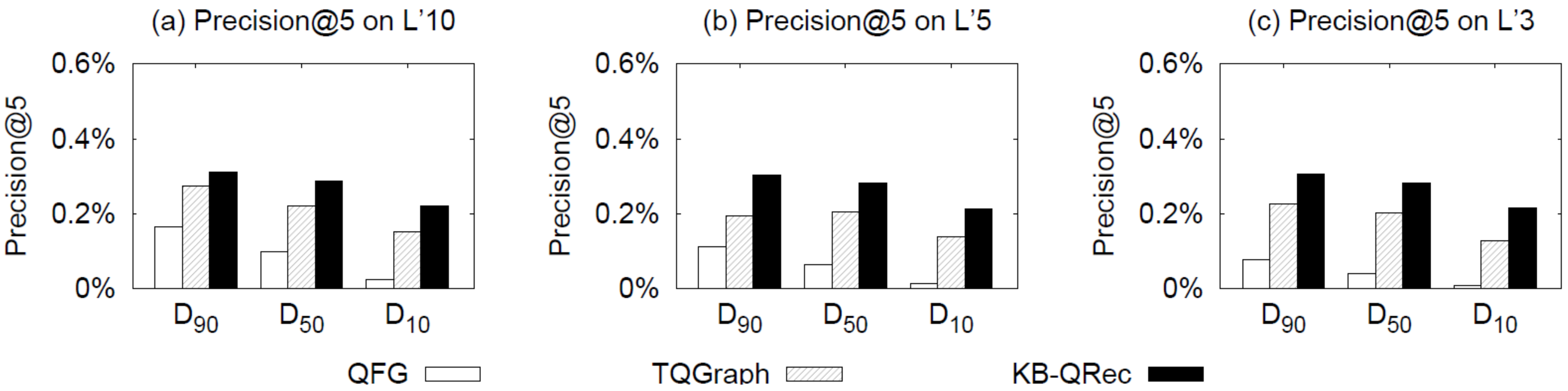
- **Return:**
  - **q_1 = “weather las vegas”**
  - **q_2 = “weather san diego”**

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, and can reduce to 0.4ms if we use FEL method.**

Table 5: Efficiency (in ms)

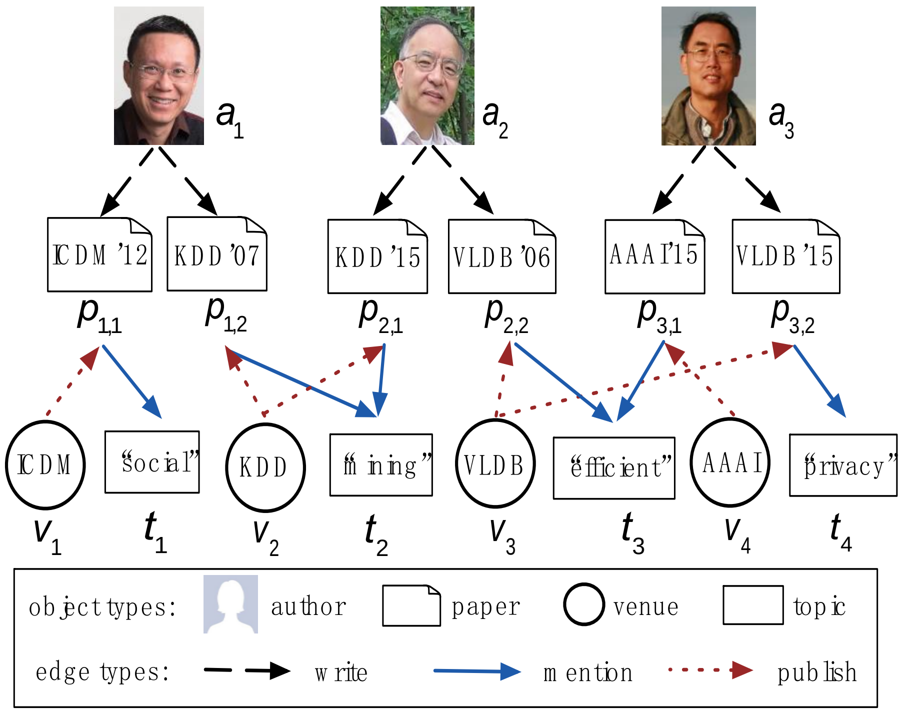
	entity expansion	PPR (no cache)	PPR (cache)	KB-QREC (no cache)	KB-QREC (cache)
D_{90}	34 ms	91 ms	9 ms	143 ms	60 ms
D_{50}	34 ms	55 ms	5 ms	100 ms	47 ms
D_{10}	33 ms	13 ms	1 ms	59 ms	37 ms

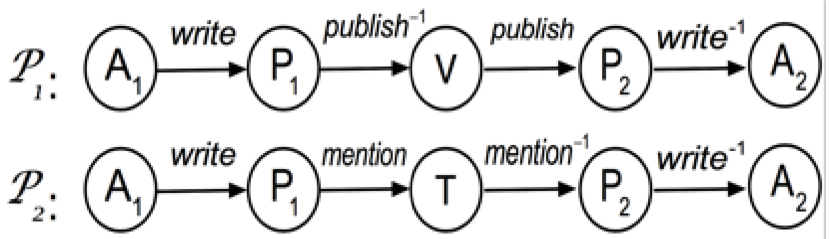
Outline

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

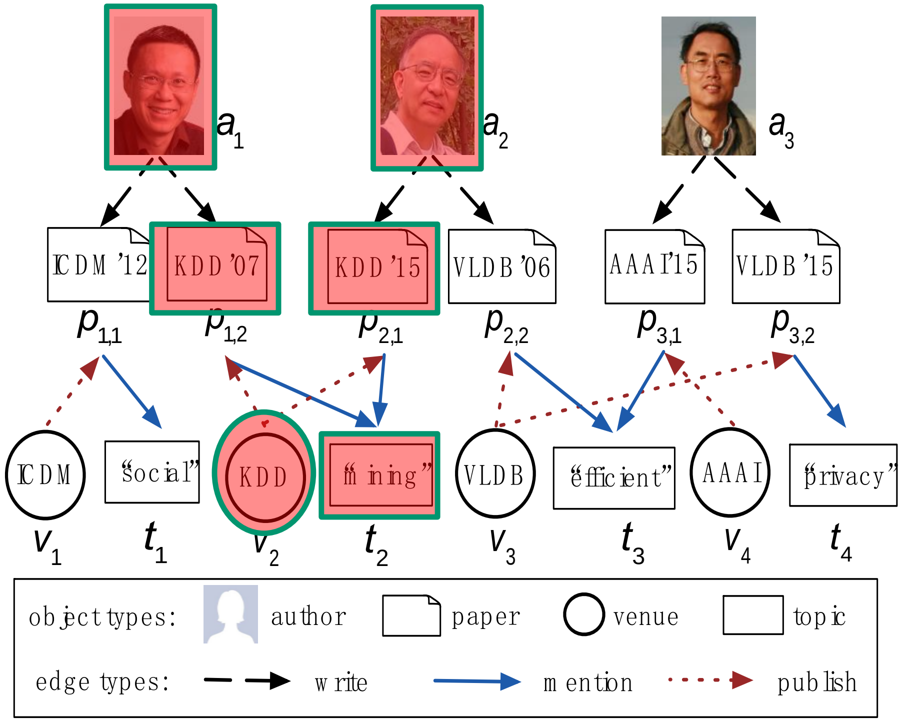


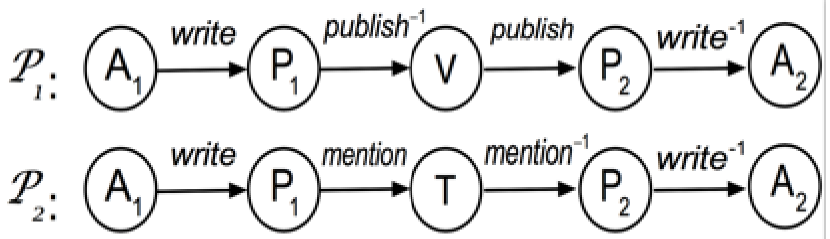


Pair	Meta Path Measures		
	PathCount	PathSim	PCRW
a_2, a_1	2	0.5	0.25
a_2, a_3	2	0.5	0.25

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

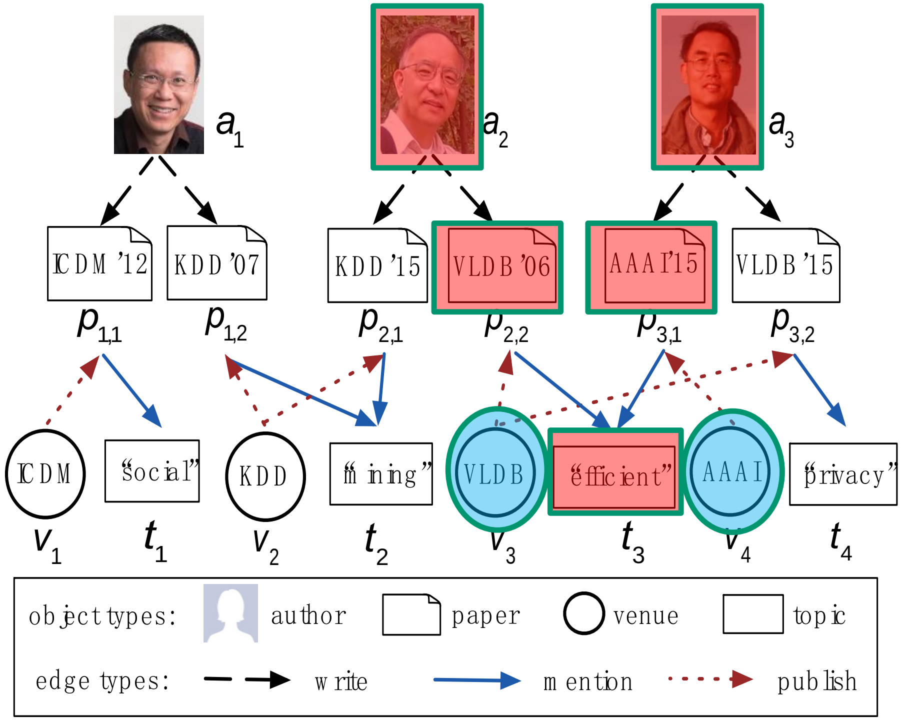


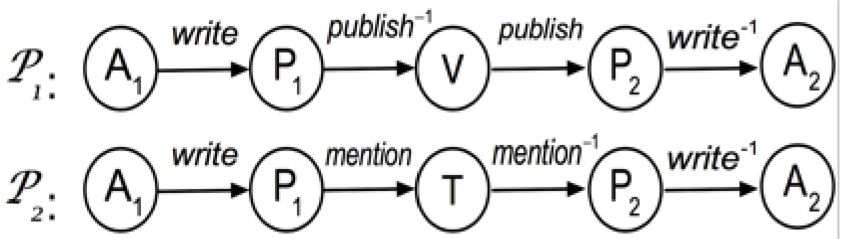


Pair	Meta Path Measures		
	PathCount	PathSim	PCRW
a_2, a_1	2	0.5	0.25
a_2, a_3	2	0.5	0.25

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

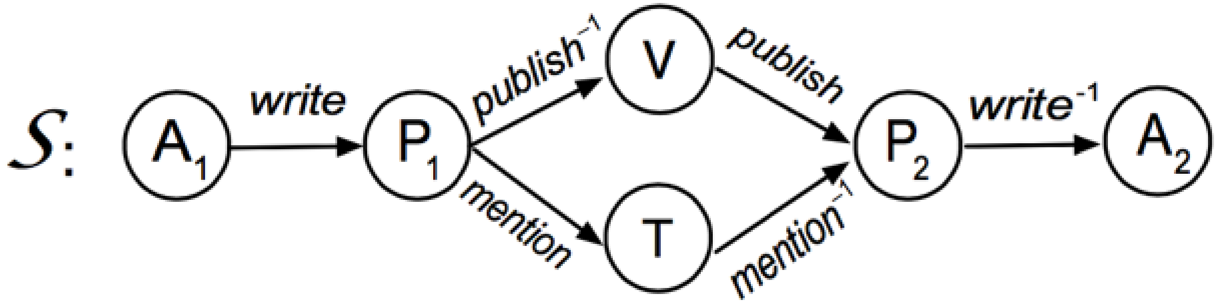




Pair	Meta Path Measures		
	PathCount	PathSim	PCRW
a_2, a_1	2	0.5	0.25
a_2, a_3	2	0.5	0.25

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
- Demo
- Conclusions & Future Work

Relevance Measure 1: StructCount

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

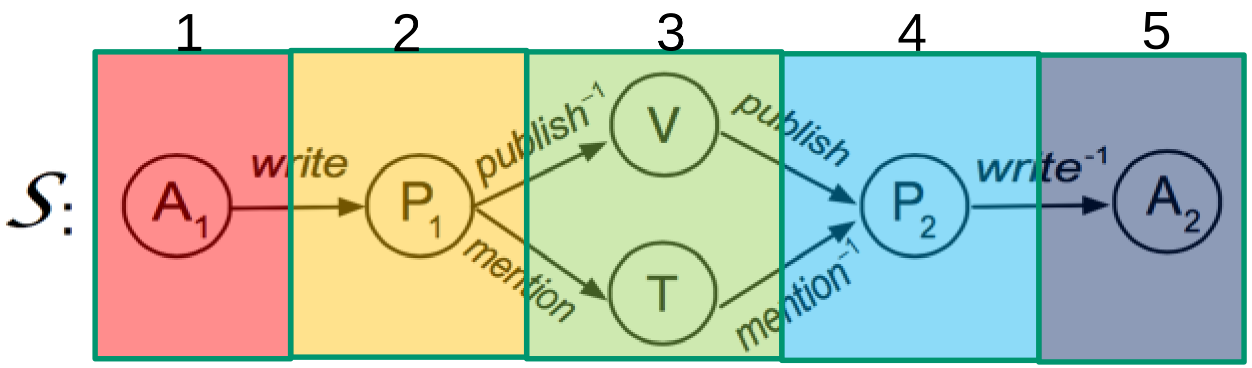
$$StructCount(x_0, y_0 \mid S) = \left| GraphIns(x_0, y_0 \mid S) \right|$$

- **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

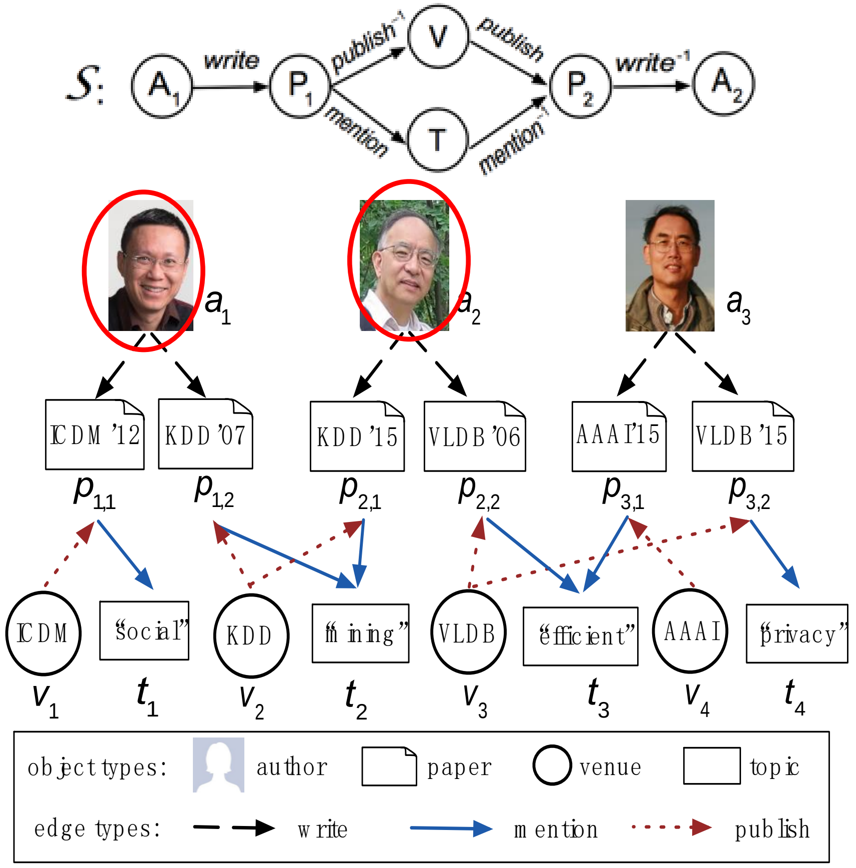
Layers of Meta Structure

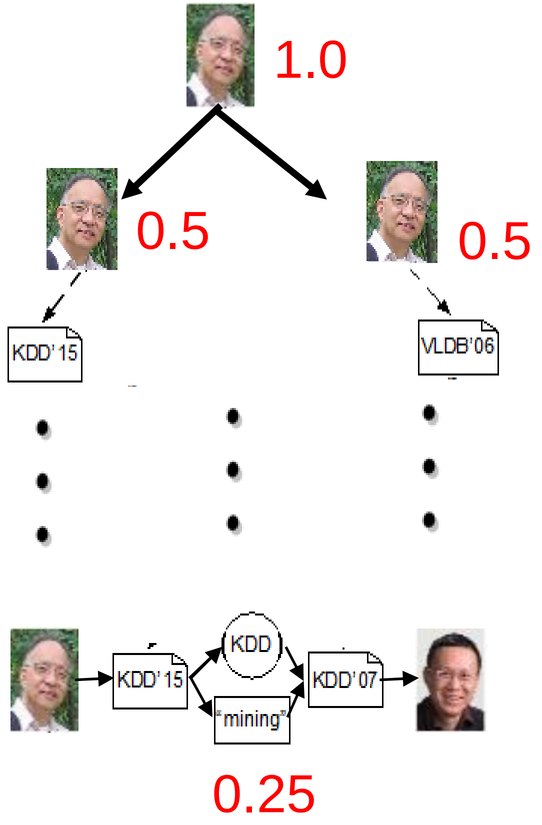
- **The layer of meta structure is a topological ordering of a DAG**



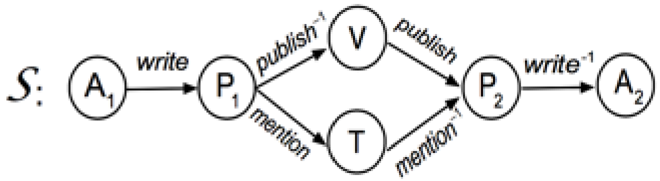
Relevance Measure 2: SCSE

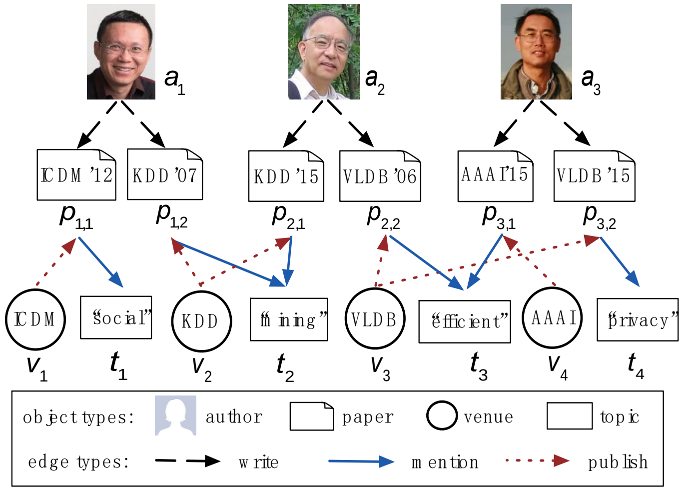
- **Structure Constrained Random Walk (SCSE): extension of PCRW.**



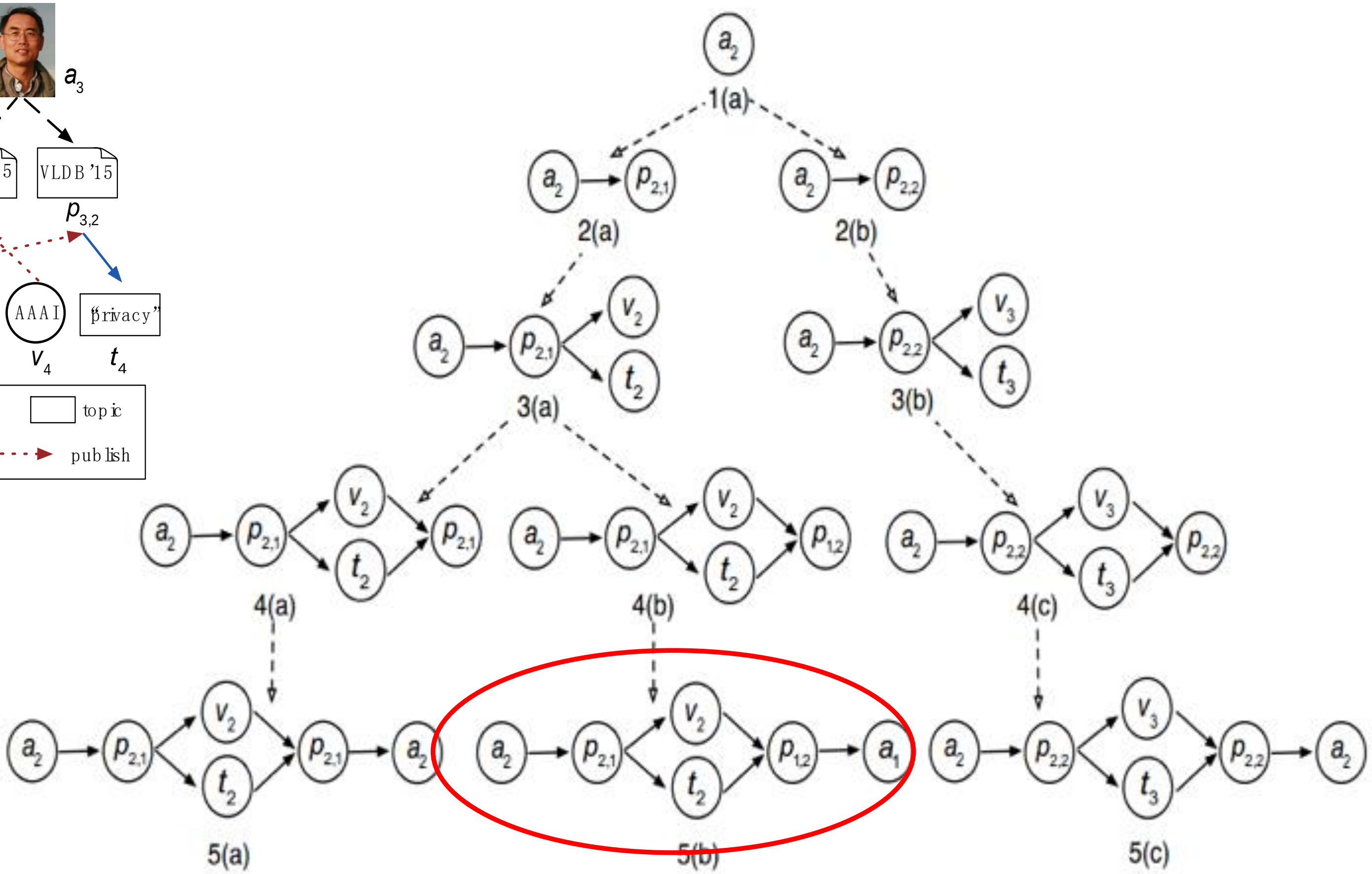


Relevance Measure 2: SCSE





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



Relevance Measure 3: BSCSE

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

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

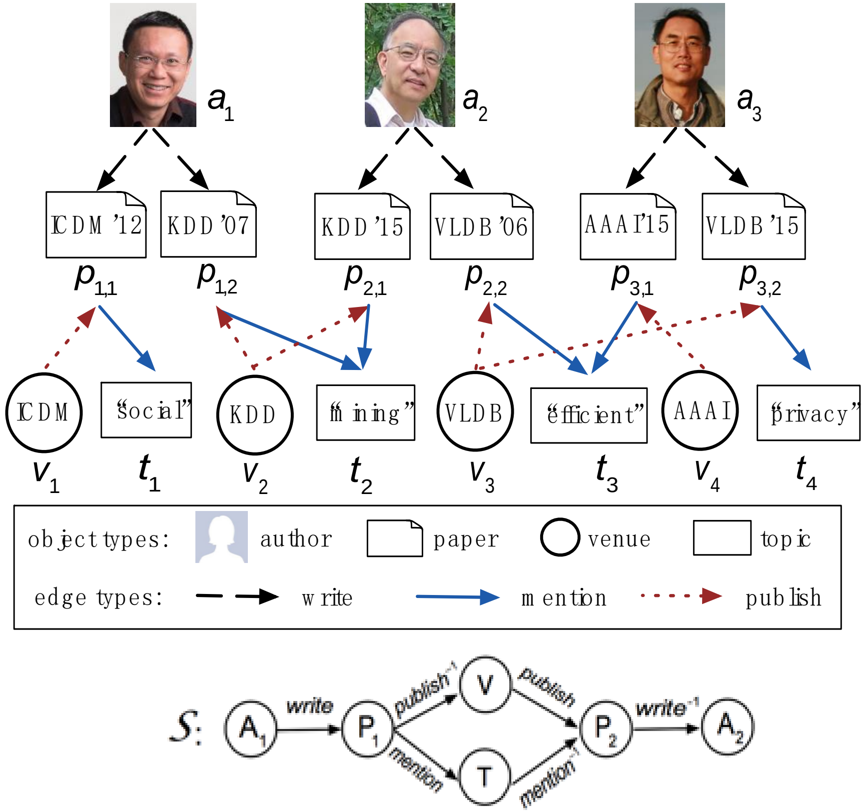
[Huang KDD’16] ZP. Huang “Meta Structure: Computing Relevance on Large Heterogeneous Information Networks” KDD 2016

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>

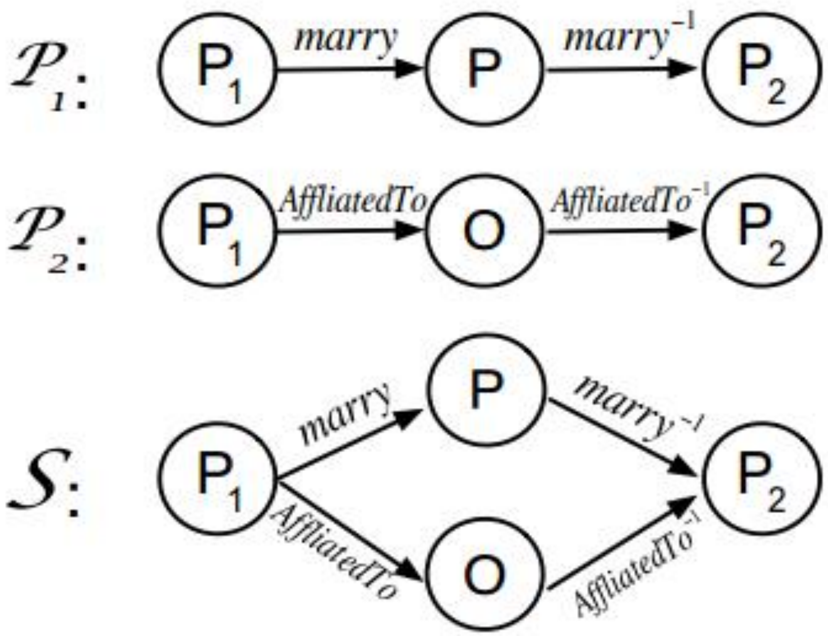
Experiment: Entity Resolution

- **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

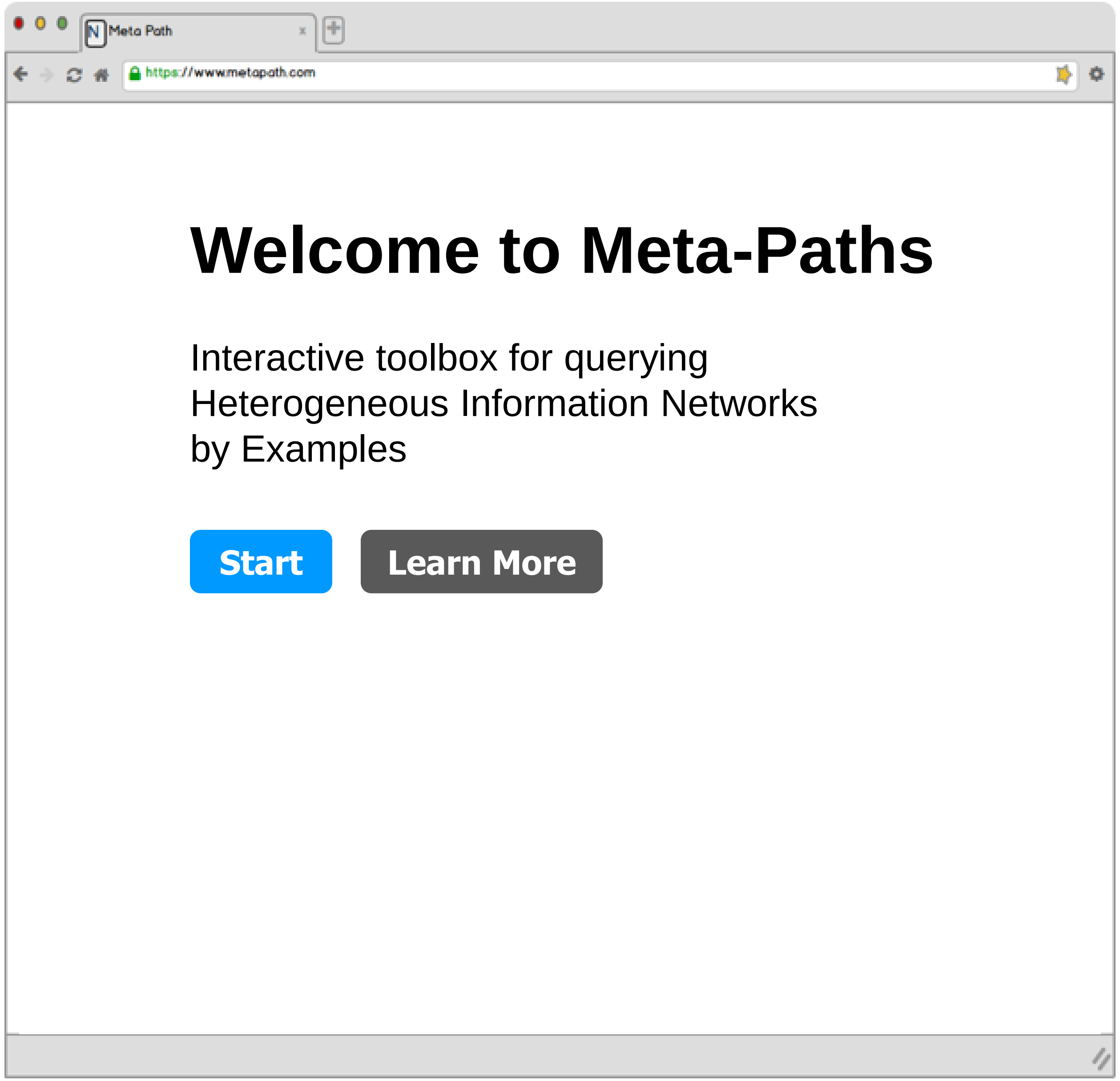


	P1			P2		
Measure	PathCount	PCRW	PathSim	PathCount	PCRW	PathSim
AUC	0.1324	0.0120	0.0097	0.0003	0.0014	0.0002
	Linear Combination(optimal)			Meta Structure S		
Measure	PathCount	PCRW	PathSim	SC	SCSE	BSCSE*
AUC	0.2898	0.2606	0.2920	0.5556	0.5640	**0.5640**

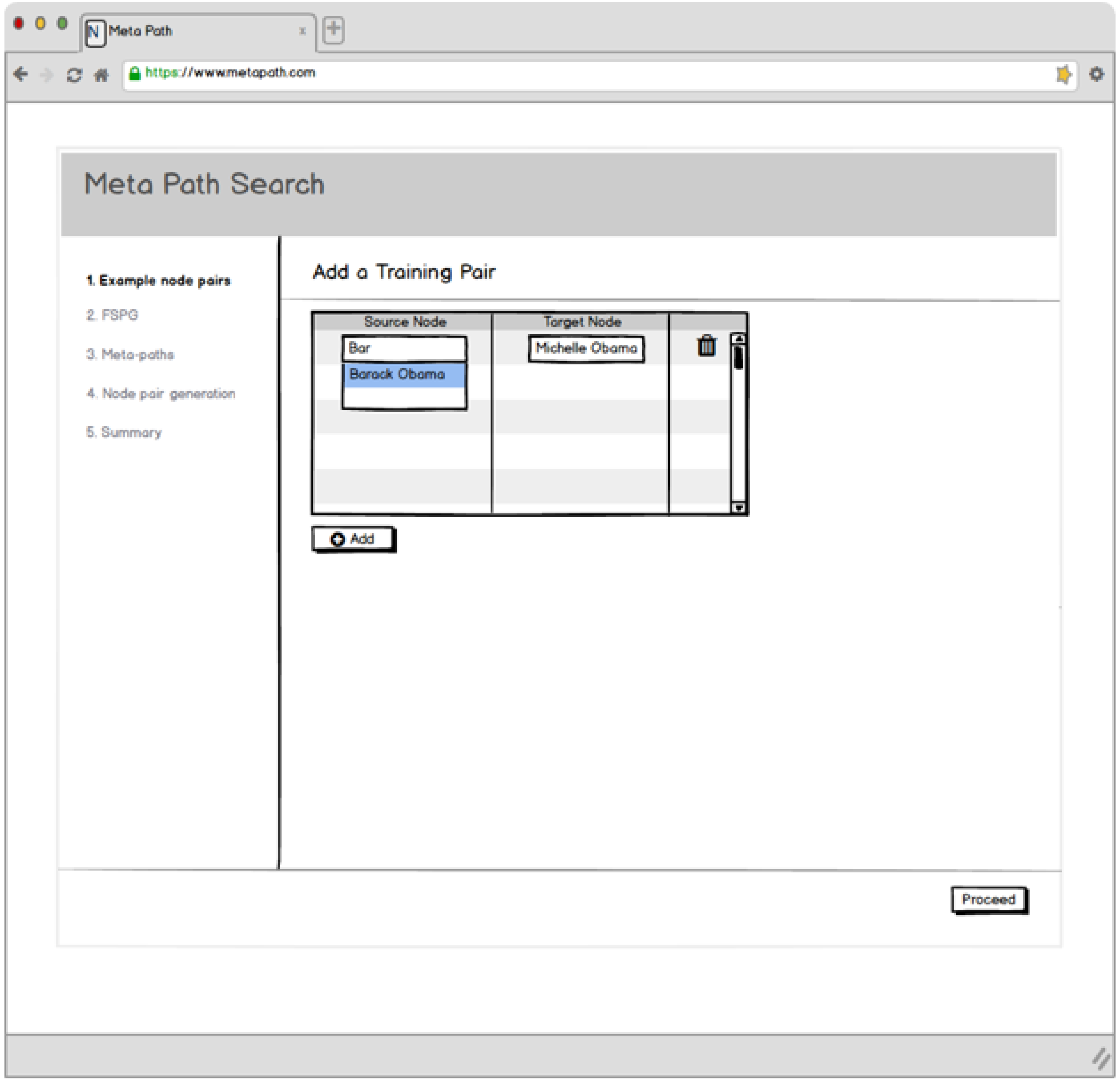
Outline

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

Meta-Paths Demo

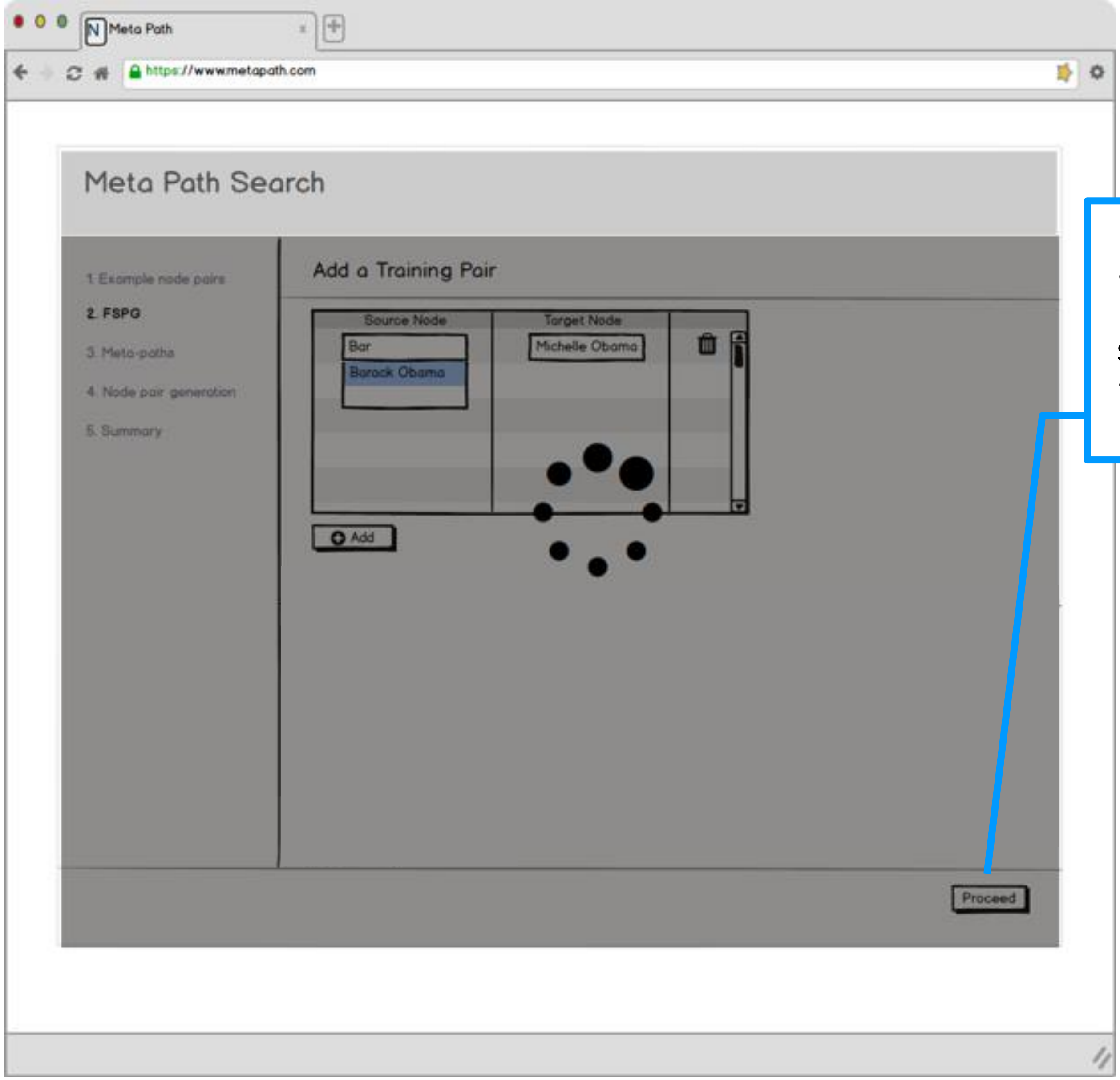


New Query



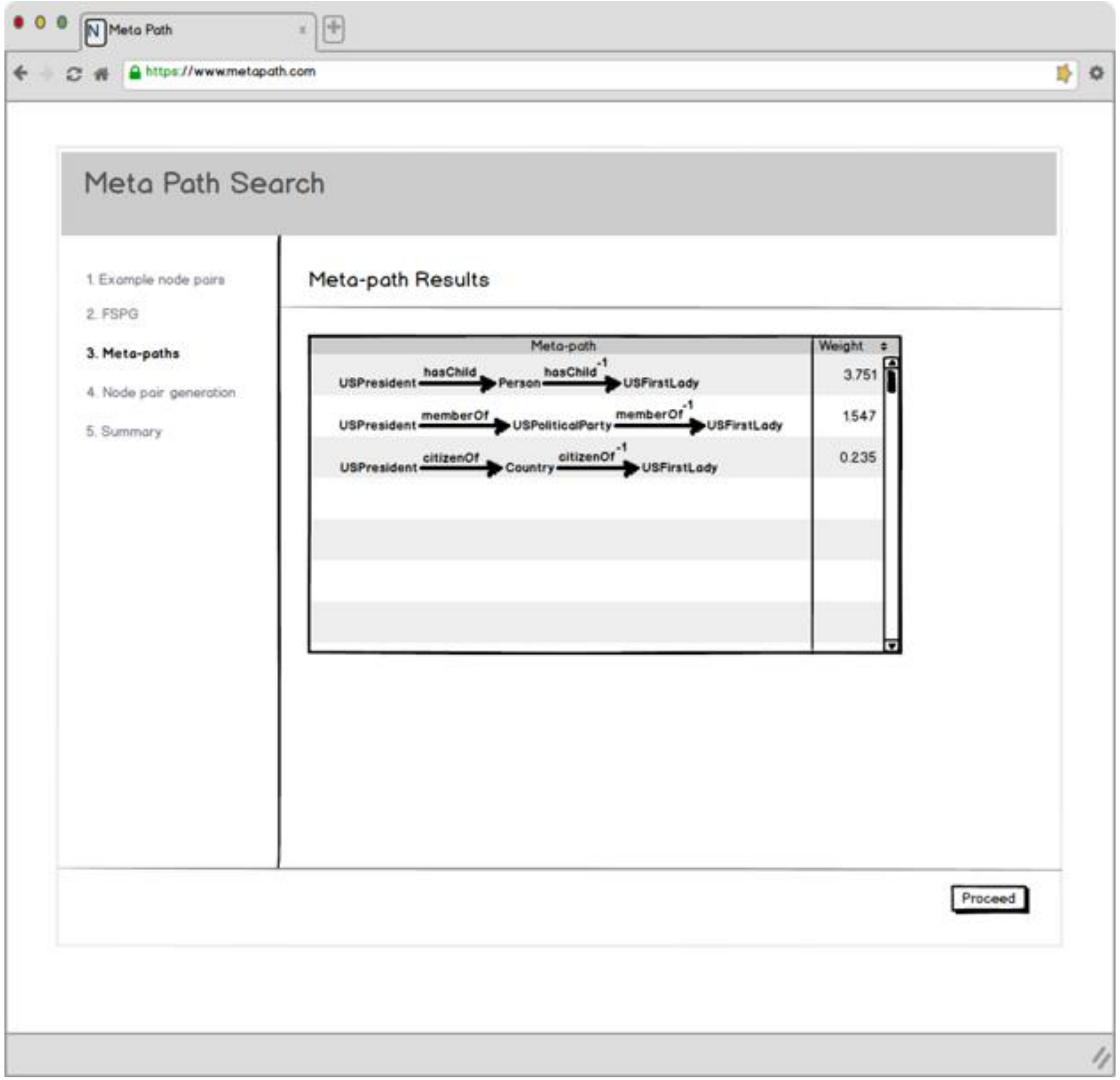
Meta Path
https://www.metapath.com
Meta Path Search
1. Example node pairs
2. FSPG
3. Meta-paths
4. Node pair generation
5. Summary
Add a Training Pair
Source Node
Target Node
Bar
Barack Obama
Michelle Obama
Add
Proceed

FSPG Execution

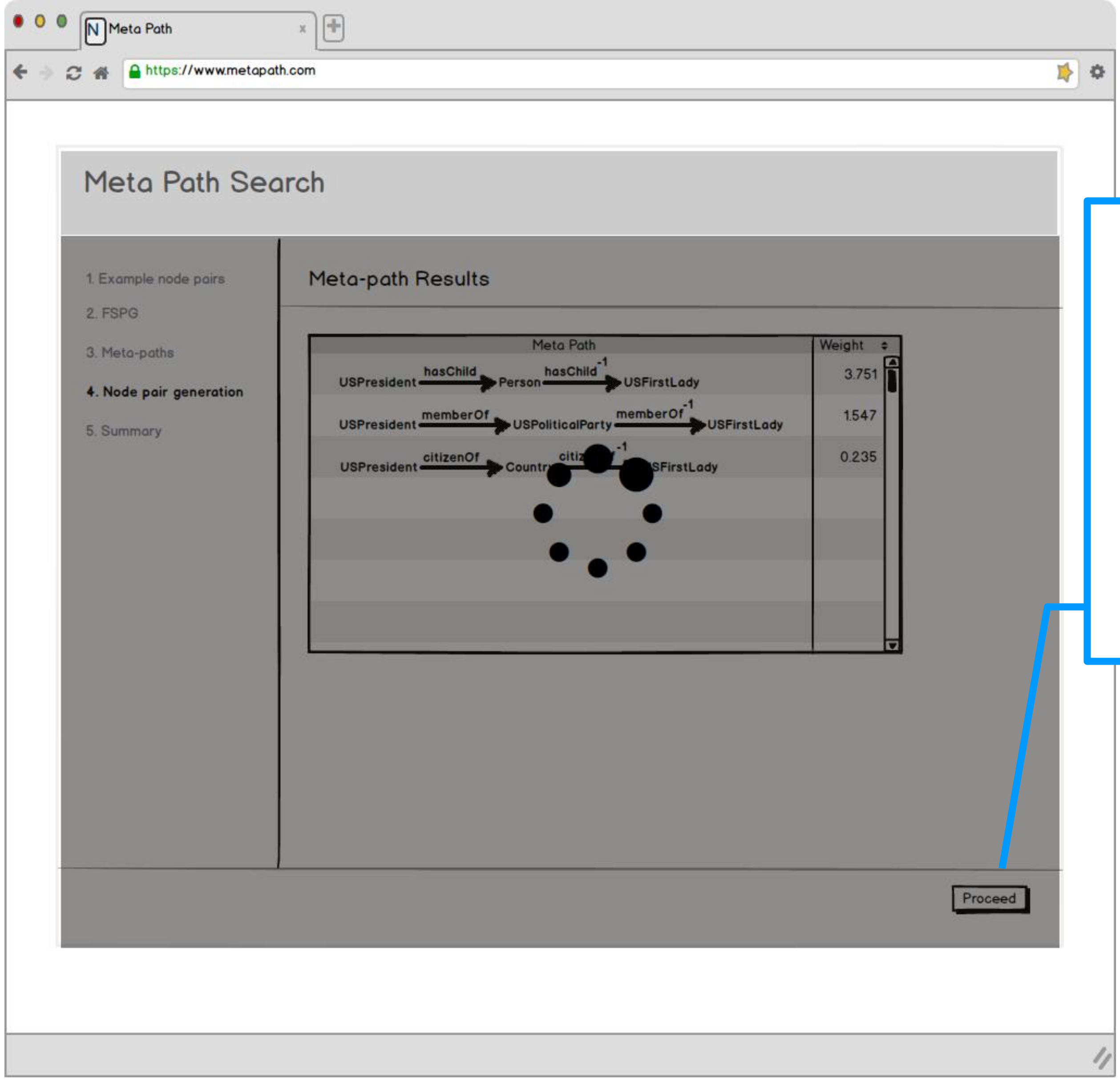


The FSPG algorithm will be triggered on the server, returning the results upon completion.

Generated Meta-Paths

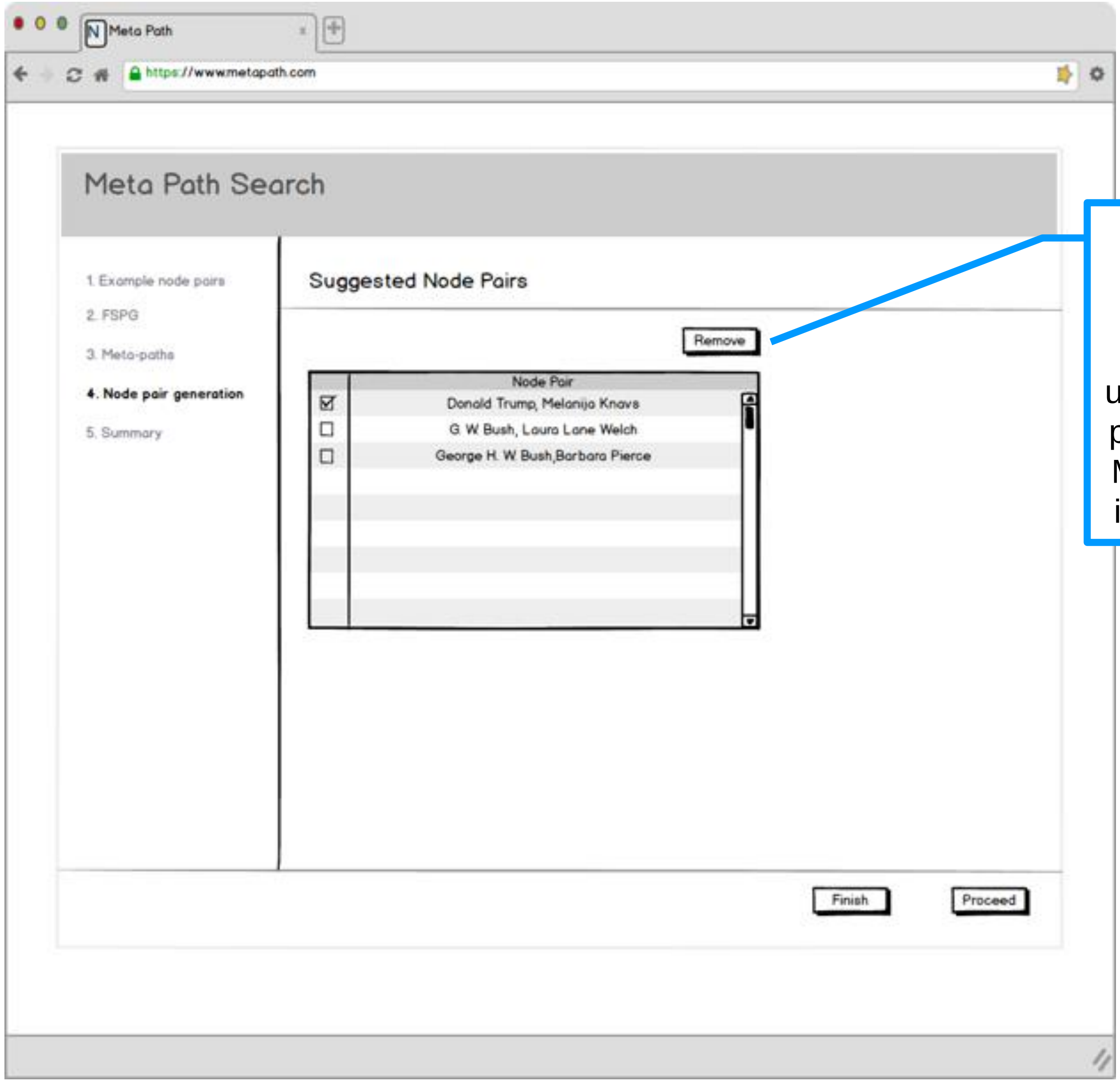


Node Pair Generation



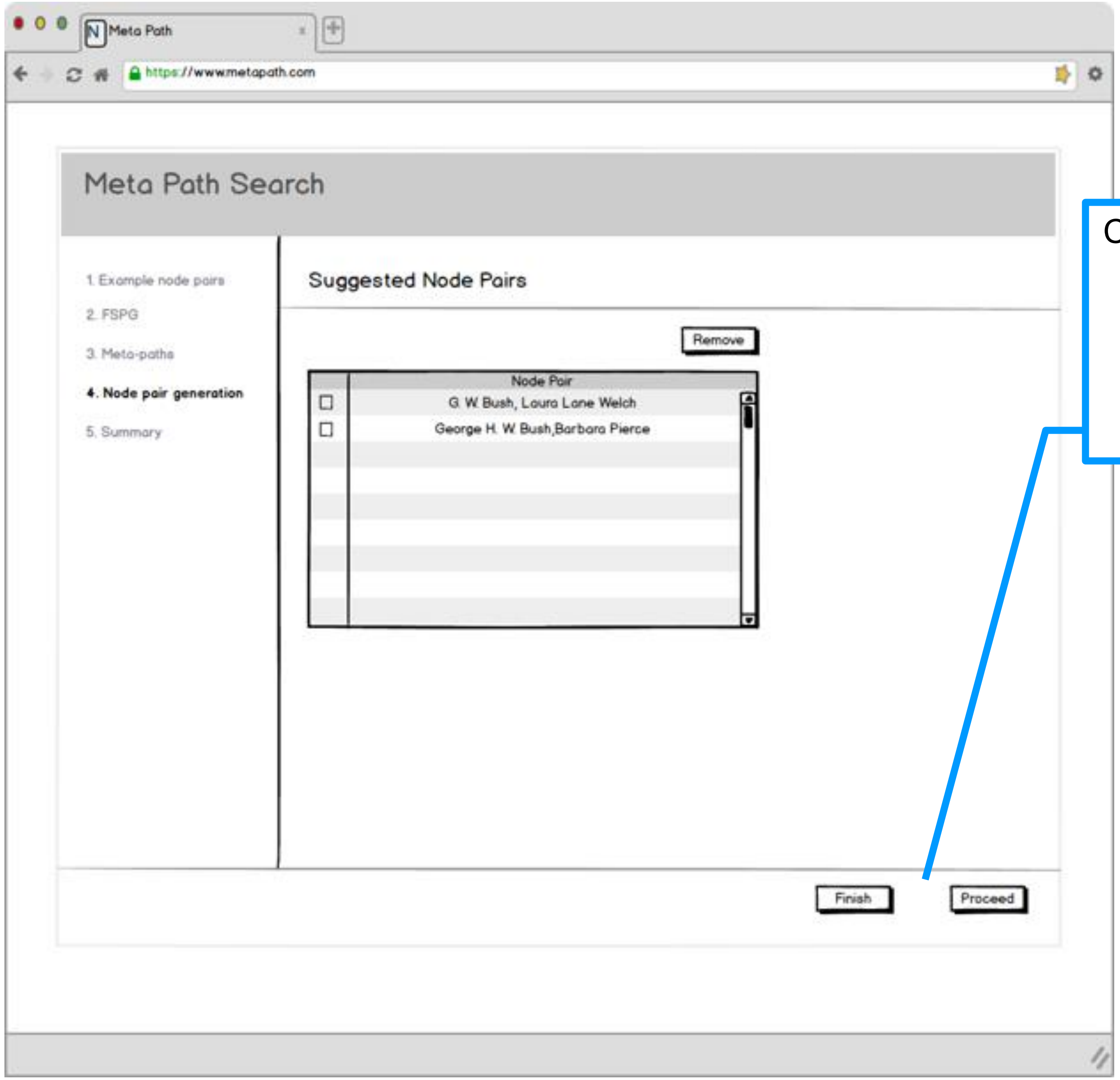
Results of the Meta-Paths algorithm are shown. Upon clicking “Proceed”, the node pair generation service will be triggered on the server.

Suggested Node Pairs



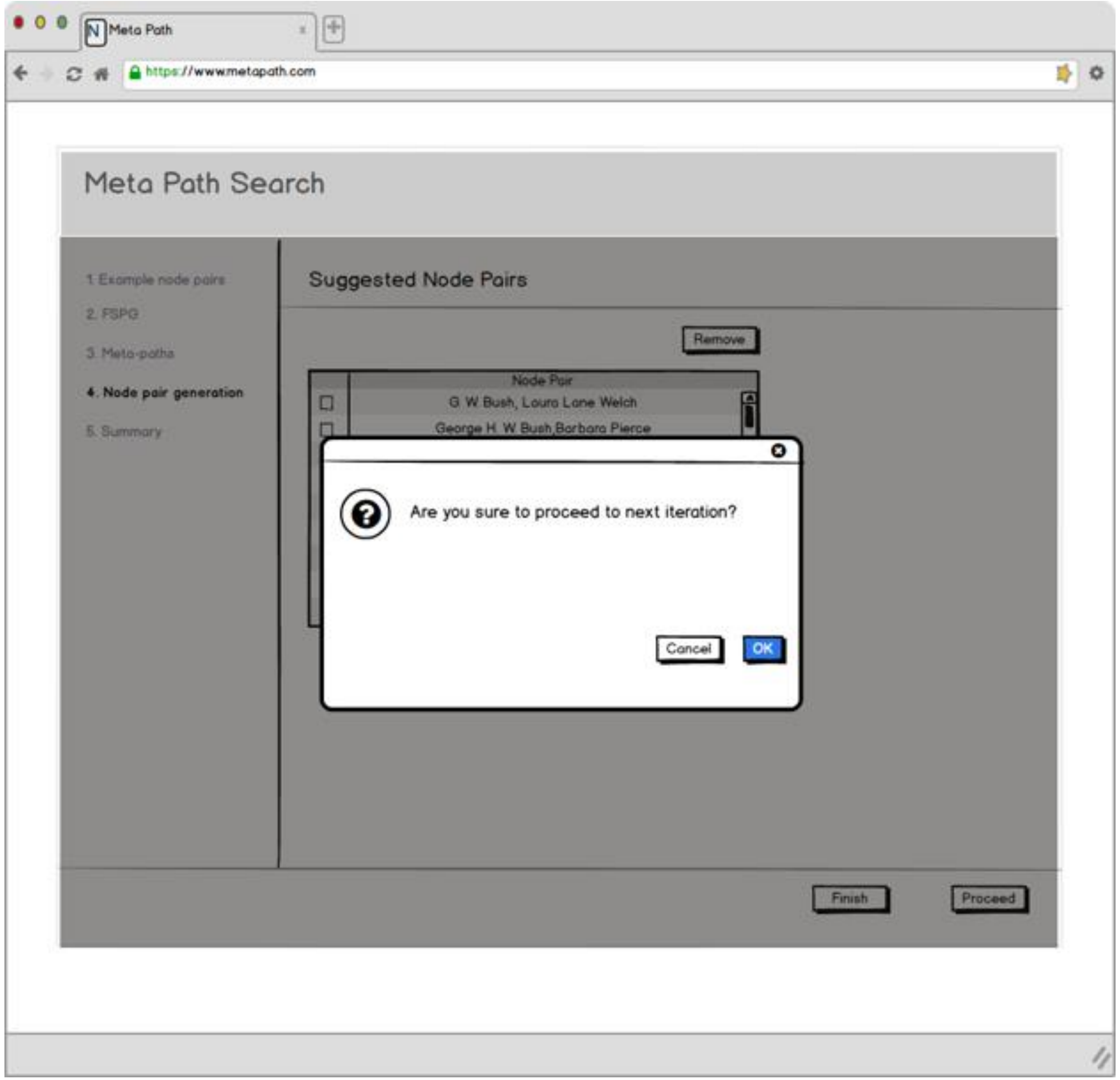
The user can remove some of the suggested node pairs, and use the remaining pairs to refine the Meta-Paths in an iterative manner.

Fine-tuning Node Pairs

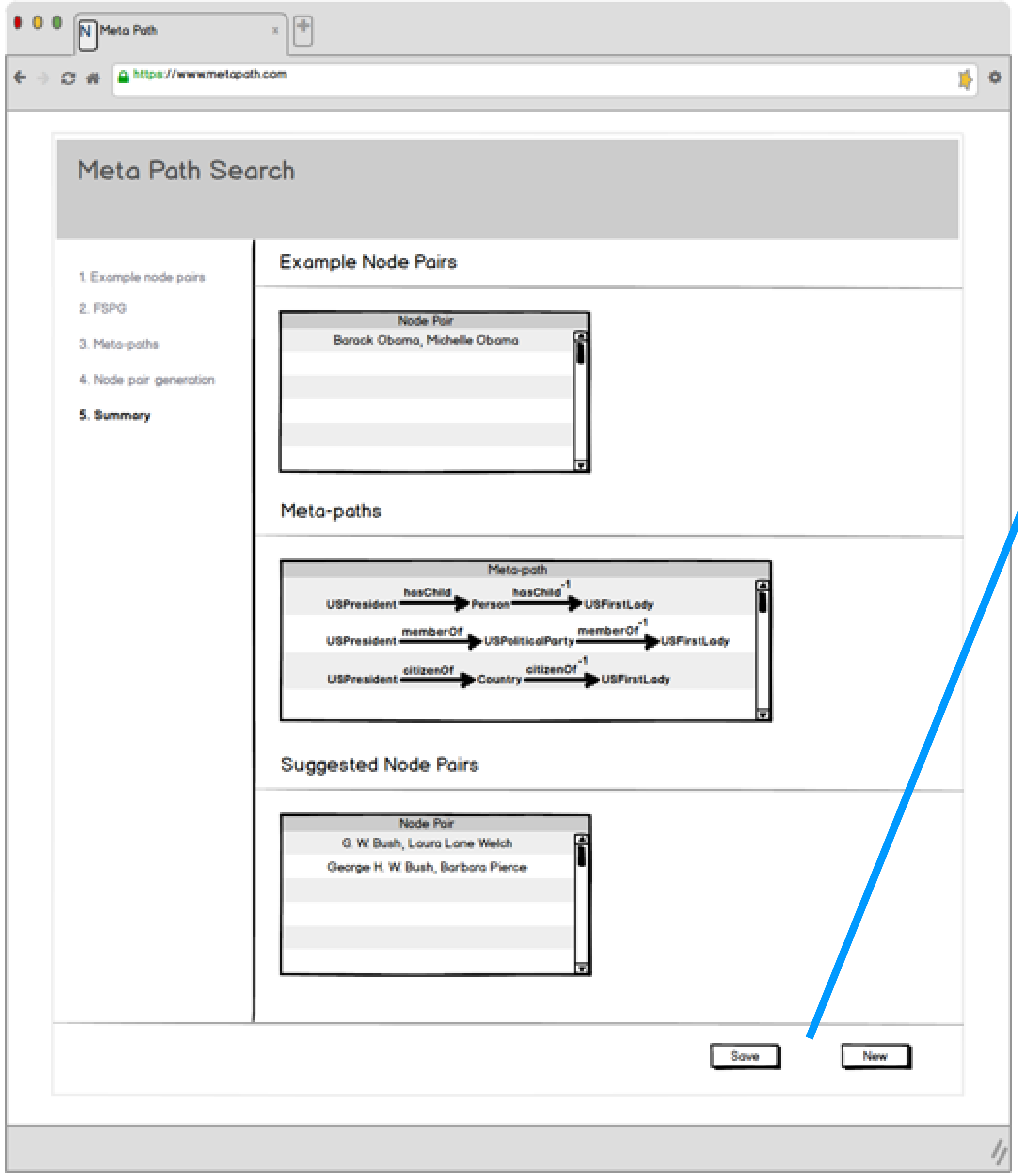


Click “Proceed” to start the next iteration, or “Finish” to view the final query results.

Next Iteration

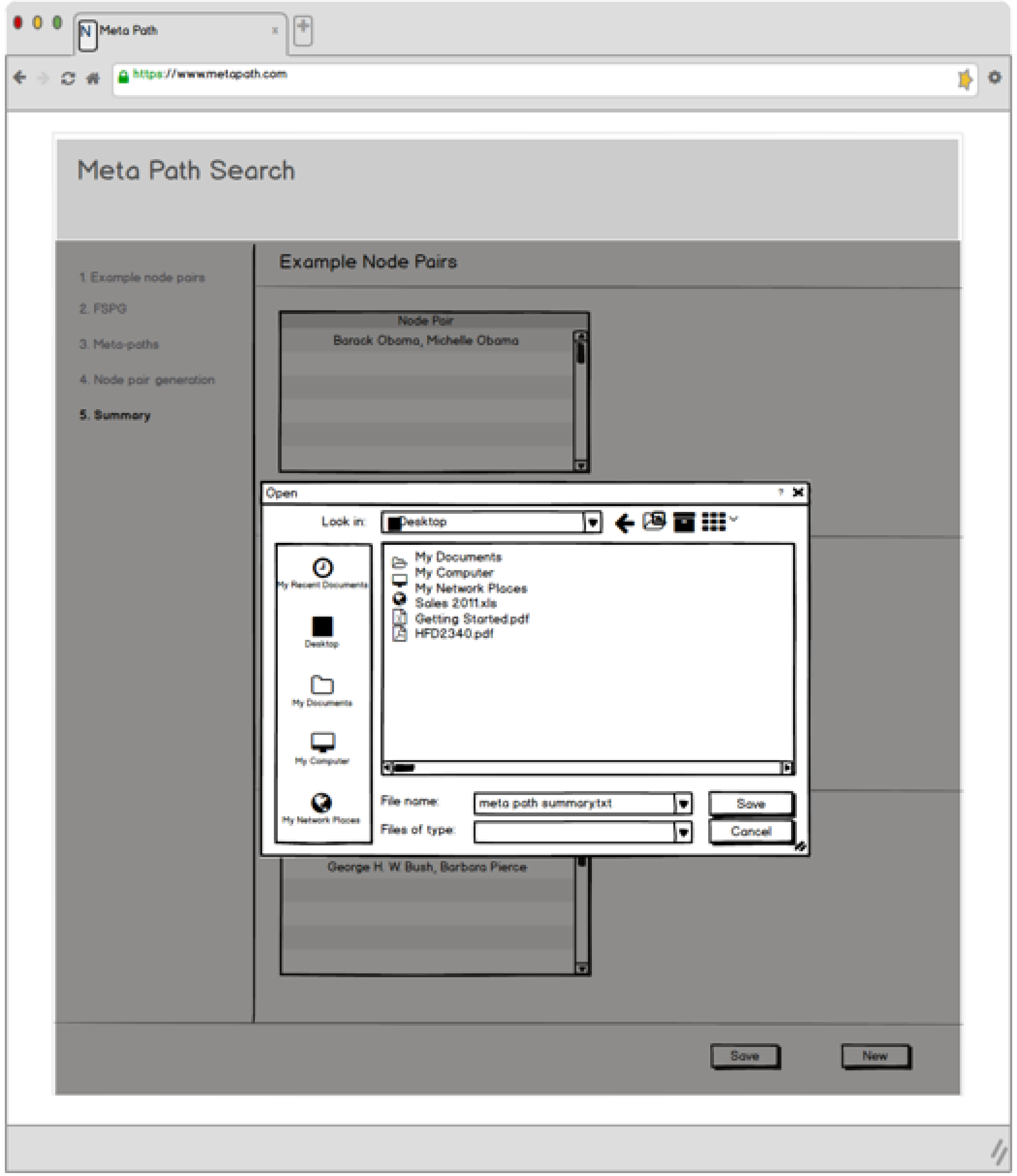


Final Results



Click “Save” to keep a copy of the query results. Alternatively, click “New” to start a new query.

Final Results



Outline

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

Conclusions

- **Heterogeneous Information Networks are more powerful than Homogeneous Information Networks**

- **Meta-path can capture the relevances (similarities) between two nodes**

- **Meta-structure captures more complex relationships in structures**

Future Work
Dynamic Similarity Search on Meta-Paths

- Sometimes the direct relevance search can not reveal the true relationship among entities.
- Solutions: Dynamic Network Search
- Problems: 1. No efficient top-k query algorithms. 2. No predicates or posterior knowledge of the network
- ML methods could help!

Future Work
Ming HINs with Meta Structure

- Use Meta Structure to perform various data mining tasks on HINs, e.g., recommendation, classification and clustering.
- Design effective and efficient techniques to discover meta structure to express the relationship between entities.

Future Work
Knowledge Graph exploration

- **Q1: Given an entity of interest in a KG, use different meta paths and meta structures to find related entities,and sort them according to relevance.**
- **Q2: Given some entity pairs, find some meta structures to account for their relationships (meta path version has been solved).**

Future Work
Personalized Knowledge Graph

- **Personalized Recommendation is popular and useful in recommendation.**
- **Rich information from query logs.**

- **Questions: How to build a Personalized KG for each user?**
- **Storage and efficiency**
- **Privacy issues**

Future Work
Knowledge Graph maintenance

- **Q1: Build a domain-specific KG from some given entity samples and a document corpus.**
- **Q2: Expand a KG by crawling info from internet.**
- **Q3: Error detection within a KG using meta path and meta structures.**
- **Q4: Error correction automatically.**

Future Work
Knowledge Graph cleaning

- Relations / Nodes in KG are inherently "dirty" (many are curated based on automatic tools / scripts, which lead to duplications or error data)

- How to clean the Knowledge Graph by removing dirty relations / nodes ?

Future Work
Machine Learning

- **Machine learning / deep learning is so hot nowadays !**

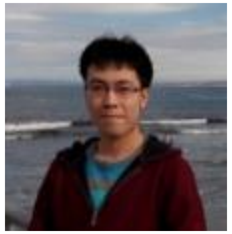
- **How to leverage the techniques in machine learning / deep learning to better enhance the heterogeneous information networks (or knowledge graphs) ?**

Future Work
Bioinformatics

- **The network is also very common in the biology. This can help interpret the network more accurately.**
- **Multi-discipline is very popular now.**
- **Can we find some typical examples in biological information networks and use meta-path or meta-structure to analyze them?**

Thanks ! Q & A

Reynold Cheng



Database Group: Zhipeng Huang, Yudian Zheng, Jing Yan, Ka Yu Wong, Eddie Ng