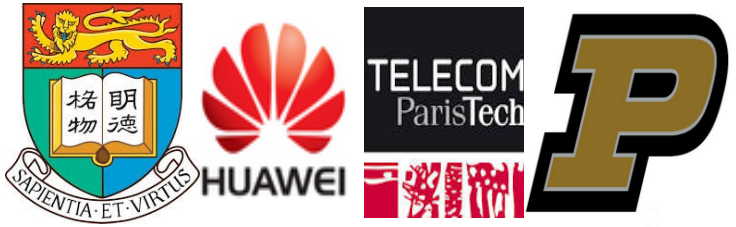
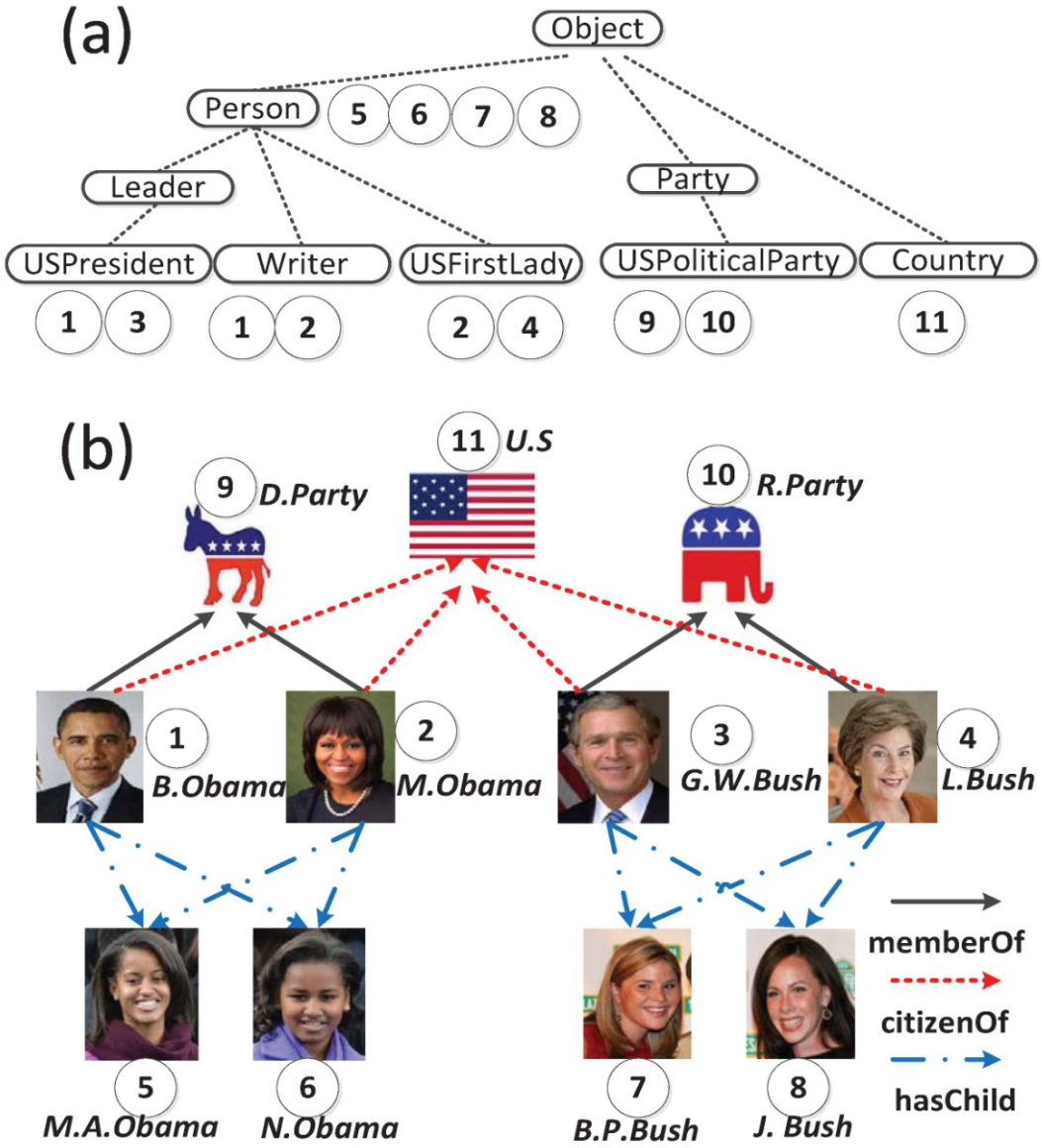


Discovering Meta-Paths in Large Heterogeneous Information Network

Changping Meng (Purdue University)
Reynold Cheng (University of Hong Kong)
Silviu Maniu (Noah's Ark Lab, Huawei Technologies)
Pierre Senellart (Télécom ParisTech)
Wangda Zhang (University of Hong Kong)

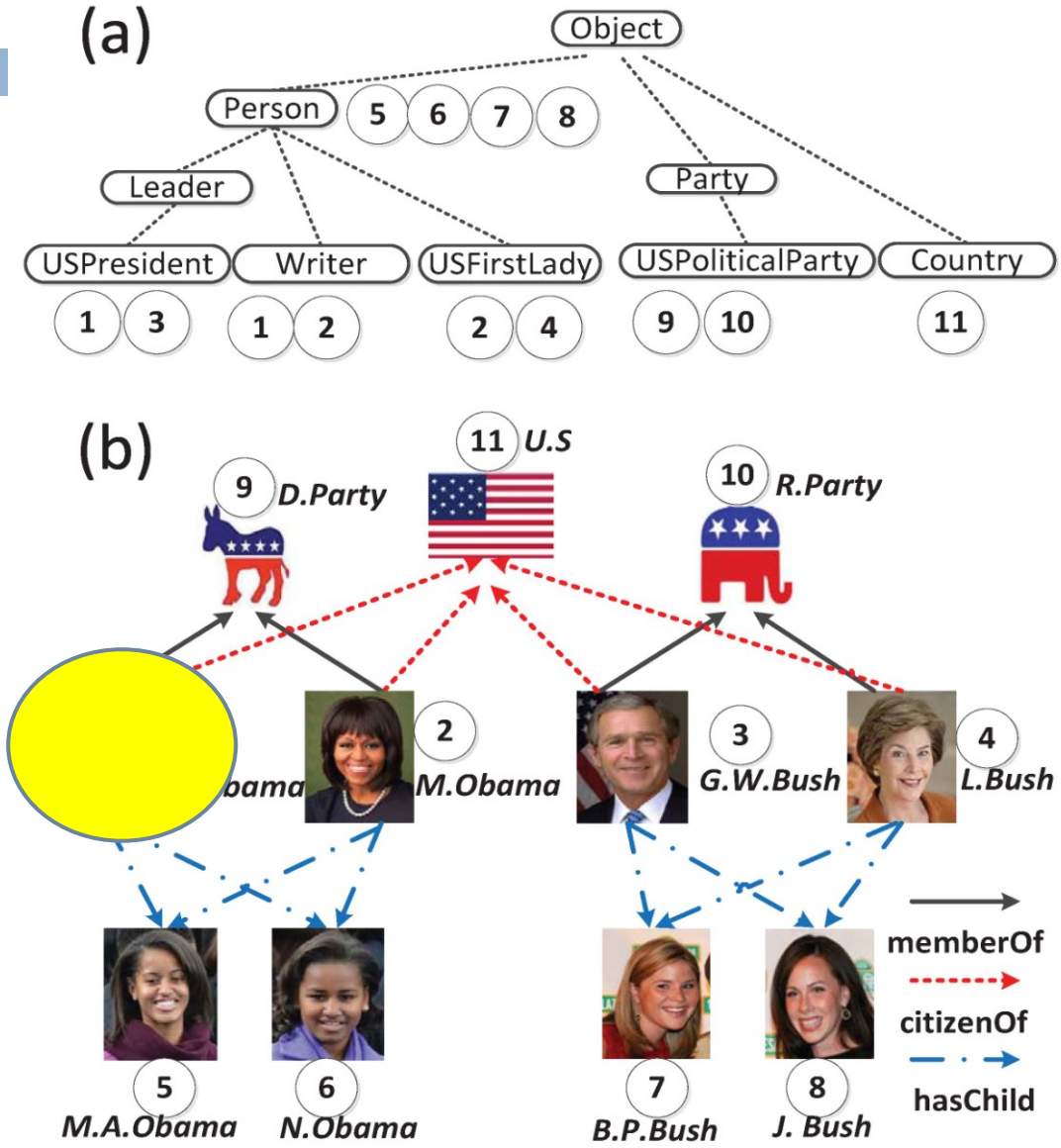
Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



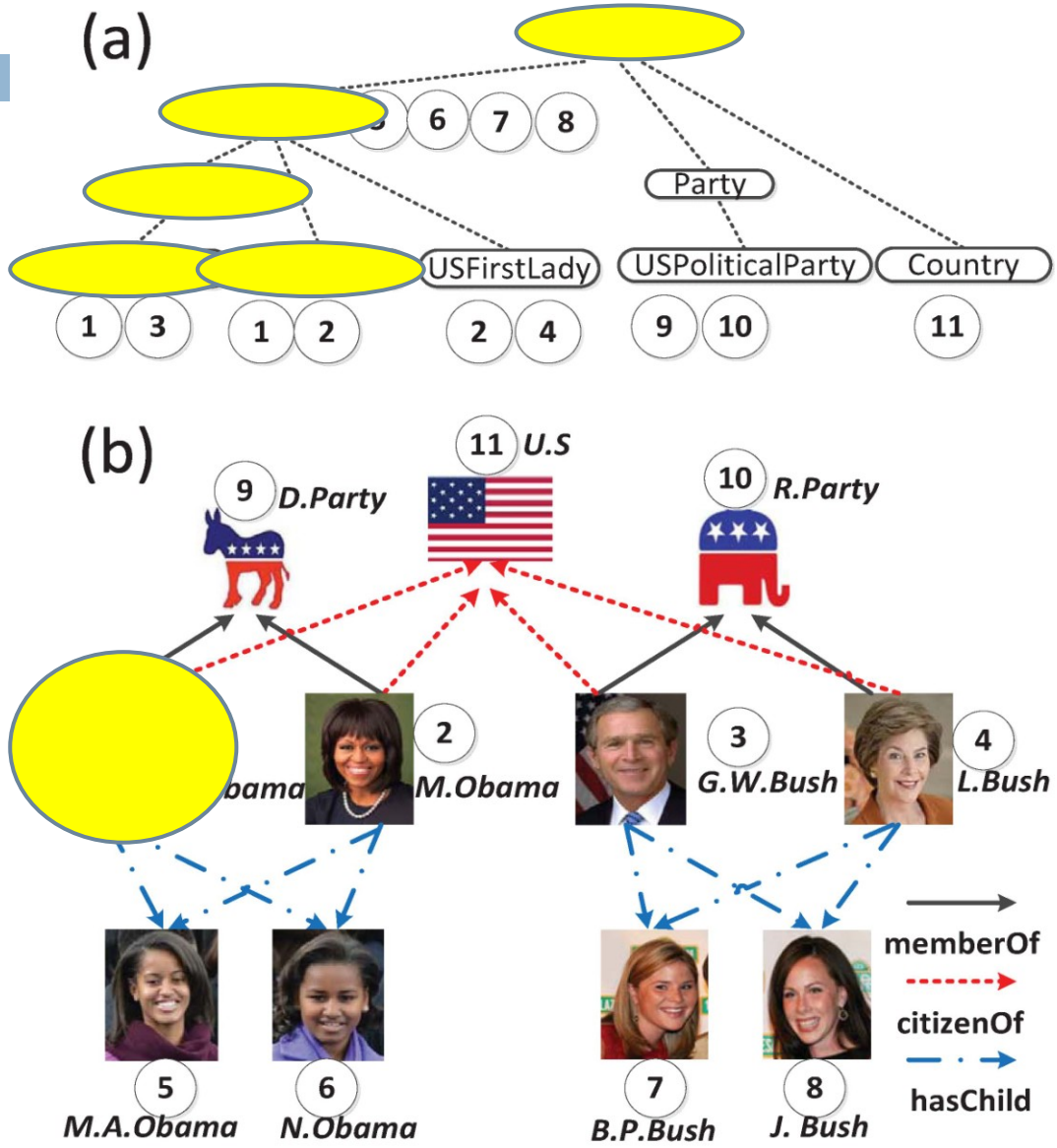
Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



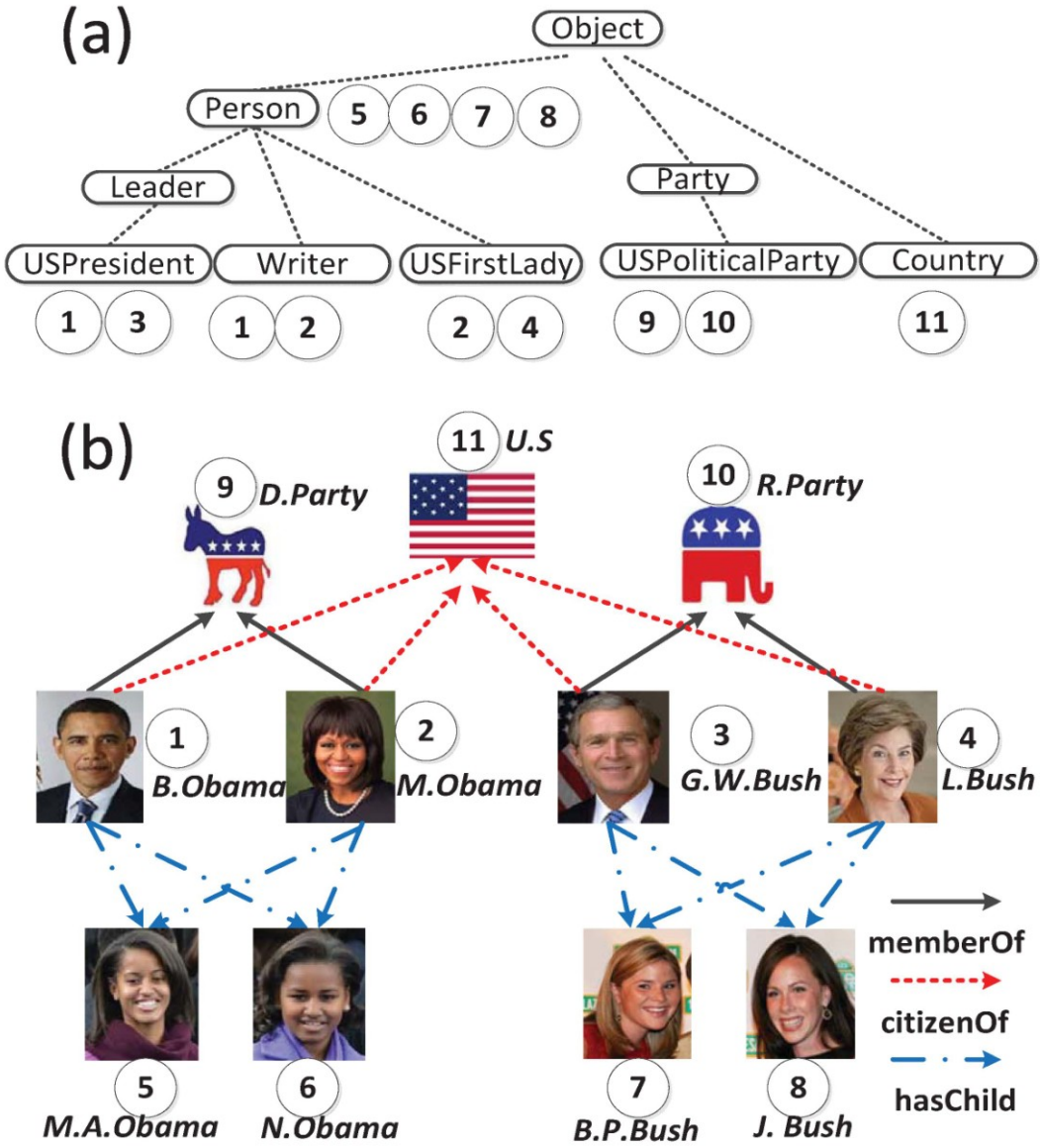
Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



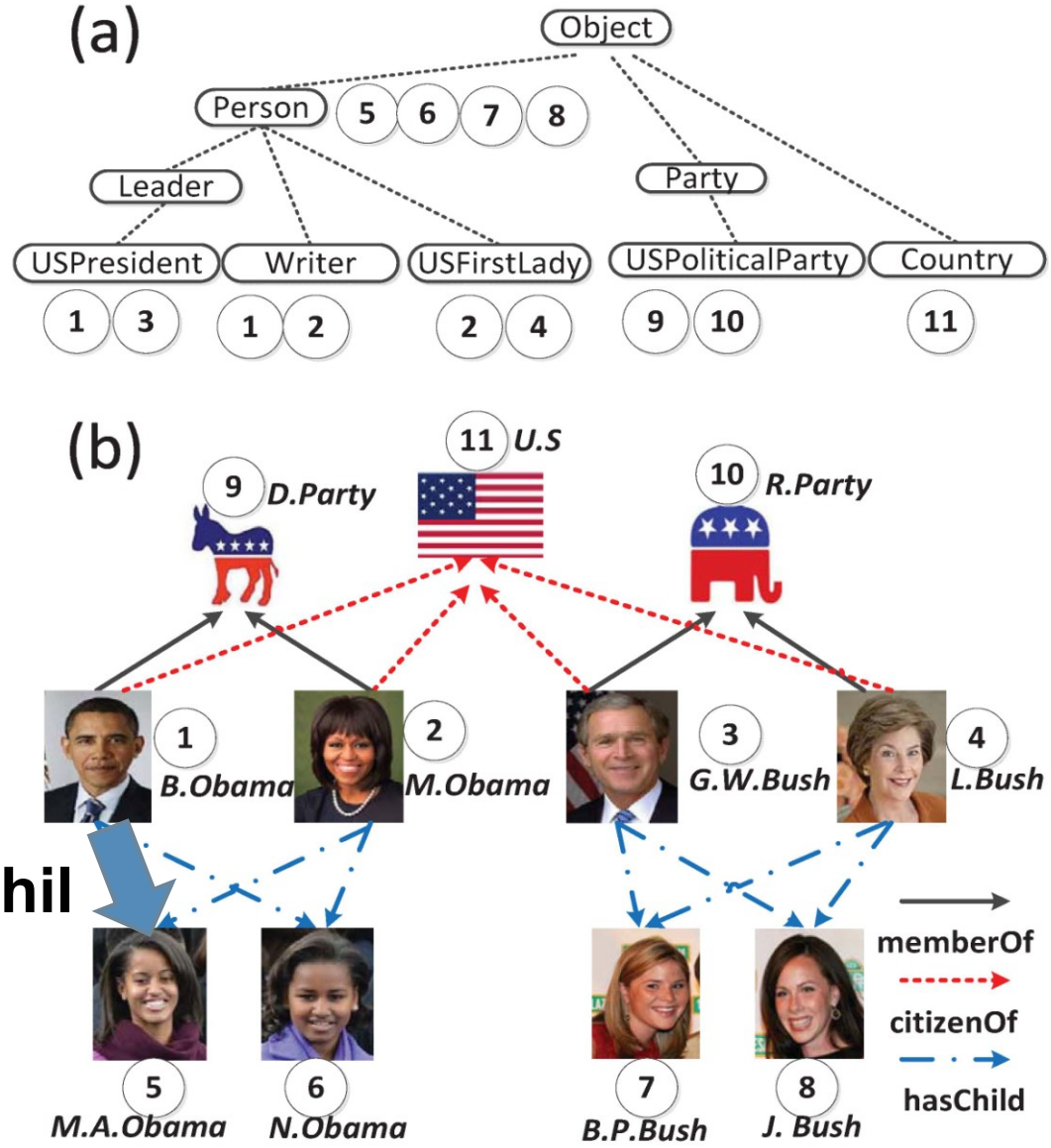
Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



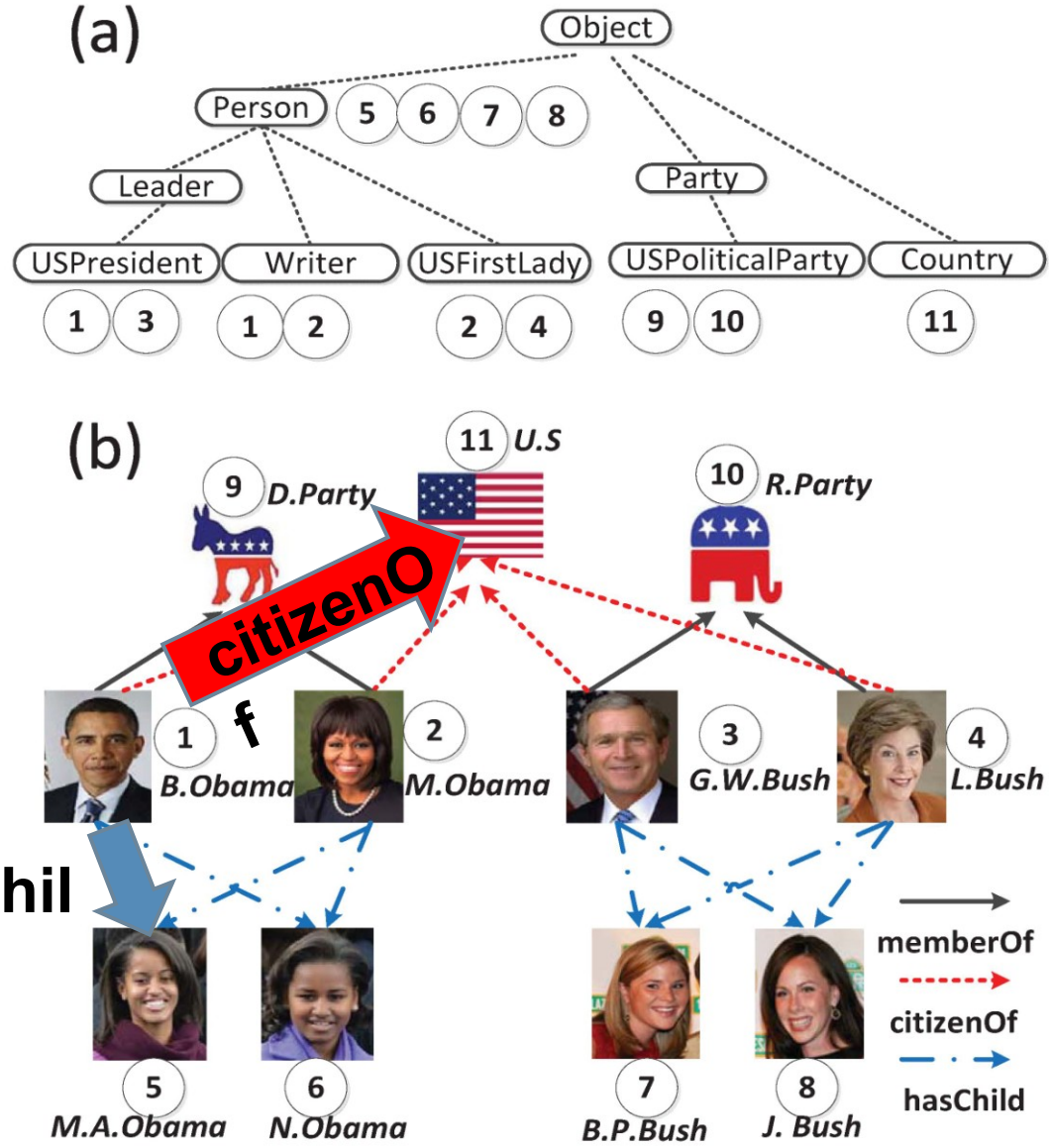
Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



Introduction

- Heterogeneous Information Network (HIN) modeled in a directed graph



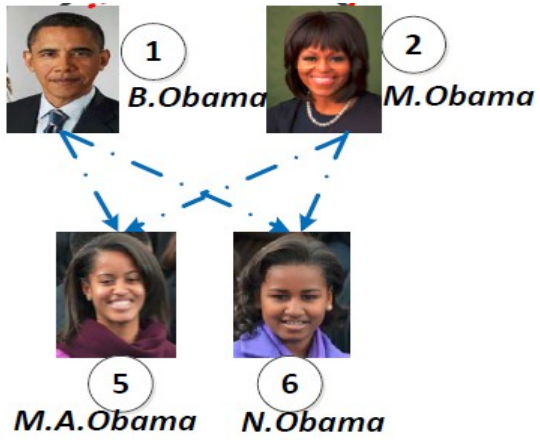
Introduction

Meta path [Han VLDB'11]

- A sequence of node class sets connected by edge types

$$\Pi^{1\ldots n} = C_1 \xrightarrow{e_1} \ldots C_i \xrightarrow{e_i} \ldots C_n$$

- Benefits of Meta Paths
  - Multi-hop relationships instead of direct links
  - Combine multiple relationships



Introduction

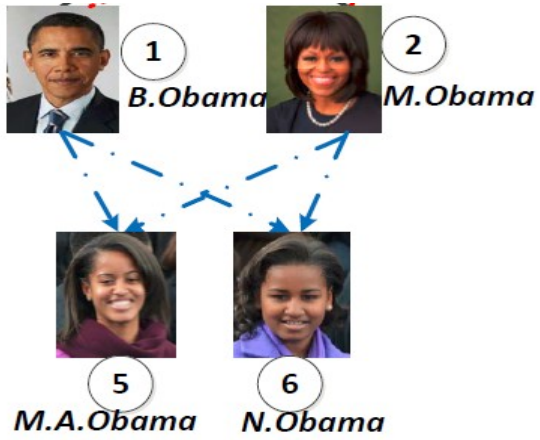
Meta path [Han VLDB'11]

- A sequence of node class sets connected by edge types

$$\Pi^{1\ldots n} = C_1 \xrightarrow{e_1} \ldots C_i \xrightarrow{e_i} \ldots C_n$$

- Benefits of Meta Paths
  - Multi-hop relationships instead of direct links
  - Combine multiple relationships

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



Introduction

Meta path [Han VLDB'11]

- A sequence of node class sets connected by edge types

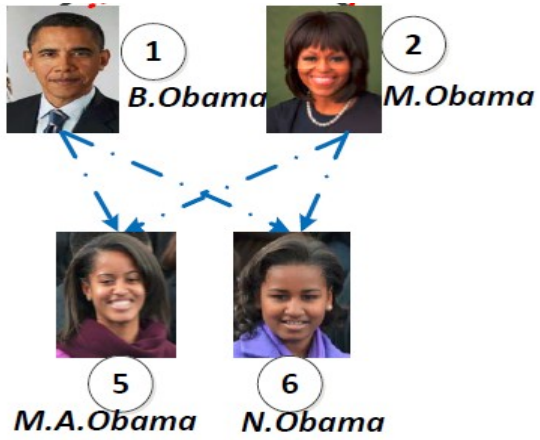
$$\Pi^{1\ldots n} = C_1 \xrightarrow{e_1} \ldots C_i \xrightarrow{e_i} \ldots C_n$$

- Benefits of Meta Paths
  - Multi-hop relationships instead of direct links
  - Combine multiple relationships

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

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

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



Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path

Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path
  - We propose a general form --Biased Path Constrained Random Walk(BPCRW)

Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path
  - We propose a general form --Biased Path Constrained Random Walk(BPCRW)
- Similarity score for a node pair following a combination of multiple meta-paths

  Aggregate Function ***F*** to combine the similarity scores for each single meta path

Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path
  - We propose a general form --Biased Path Constrained Random Walk(BPCRW)
- Similarity score for a node pair following a combination of multiple meta-paths

  Aggregate Function ***F*** to combine the similarity scores for each single meta path

Given meta - paths m_1, m_2, m_3

$\sigma(s,t|m_1) = 1$

$\sigma(s,t|m_2) = 0.2$

$\sigma(s,t|m_3) = 0.3$

Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path
  - We propose a general form --Biased Path Constrained Random Walk(BPCRW)
- Similarity score for a node pair following a combination of multiple meta-paths

  Aggregate Function ***F*** to combine the similarity scores for each single meta path

$$\text{Given} \quad \text{meta} \quad \text{- paths} \quad m_1, m_2, m_3$$

$$\sigma(s,t|m_1) = 1$$

$$\sigma(s,t|m_2) = 0.2$$

$$\sigma(s,t|m_3) = 0.3$$

$$F = 3 \times \sigma(s,t|m_1) + 2 \times \sigma(s,t|m_2) + \sigma(s,t|m_3)$$

Introduction

- Similarity score for a node pair following a single meta-path
  - Path Count(PC) [Han ASONAM'11]
    - Number of the paths following a given meta-path
  - Path Constrained Random Walk(PCRW) [Cohen KDD'11]
    - Transition probability of a random walk following a given meta-path
  - We propose a general form --Biased Path Constrained Random Walk(BPCRW)
- Similarity score for a node pair following a combination of multiple meta-paths

  Aggregate Function ***F*** to combine the similarity scores for each single meta path

Given meta - paths m_1, m_2, m_3

$\sigma(s,t|m_1) = 1$

$\sigma(s,t|m_2) = 0.2$

$\sigma(s,t|m_3) = 0.3$

$$F = 3 \times \sigma(s,t|m_1) + 2 \times \sigma(s,t|m_2) + \sigma(s,t|m_3)$$

$$\sigma(s,t|\theta) = 3.7$$

Introduction

Applications

- 1. Query by example
  - When user inputs example pairs of similar objects, we could model the user's preference and find more pairs.

Input:< Barack Obama, Michelle Obama>

NELL

yago
select knowledge

ProBase

Introduction

Applications

- 1. Query by example
  - When user inputs example pairs of similar objects, we could model the user's preference and find more pairs.

NELL yago select knowledge

ProBase

Input:< Barack Obama, Michelle Obama>

$m1$: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.

Introduction

Applications

- 1. Query by example
  - When user inputs example pairs of similar objects, we could model the user's preference and find more pairs.

NELL

yago
select knowledge

ProBase

Input:< Barack Obama, Michelle Obama>

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

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

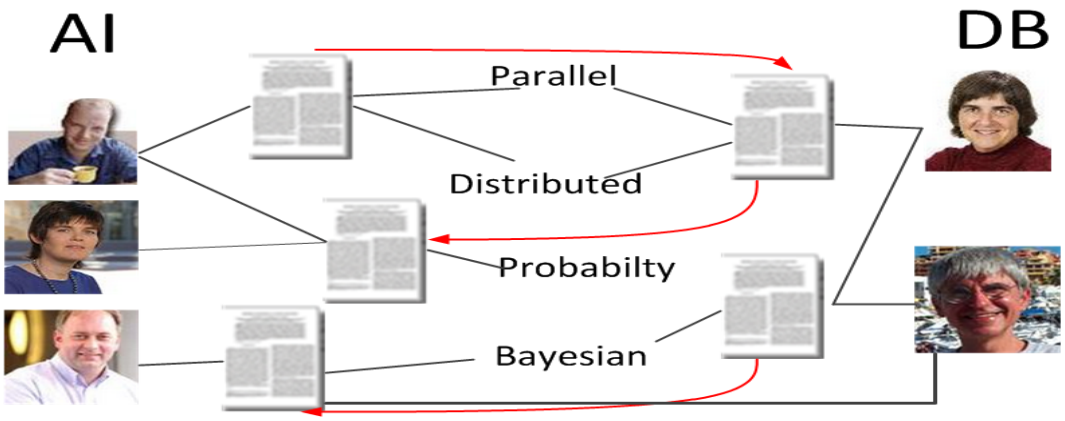
$m3$:USPresident $\xrightarrow{\text{citizenOf}}$ Country $\xrightarrow{\text{citizenOf}^{-1}}$ USFirstLady.

Output:
< George Bush, Laura Bush>
< Bill Clinton, Hillary Clinton>

Introduction

Applications

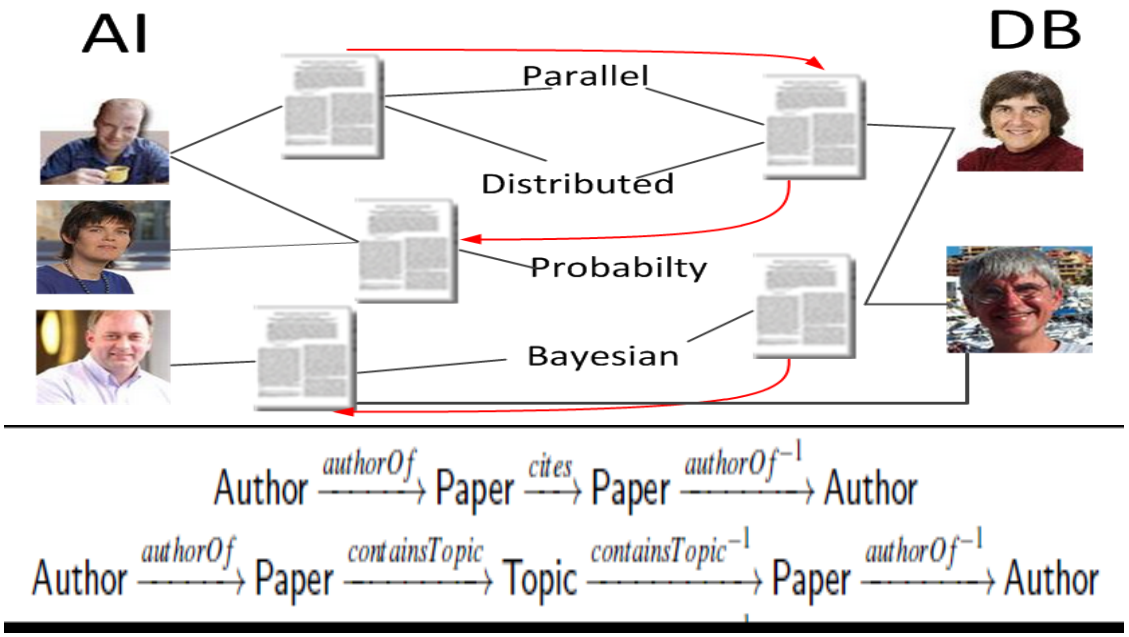
- 2. Link prediction
  - Coauthor prediction (Authors from DB and AI to collaborate)
  - Friendship prediction (Online Social Network)



Introduction

Applications

- 2. Link prediction
  - Coauthor prediction (Authors from DB and AI to collaborate)
  - Friendship prediction (Online Social Network)



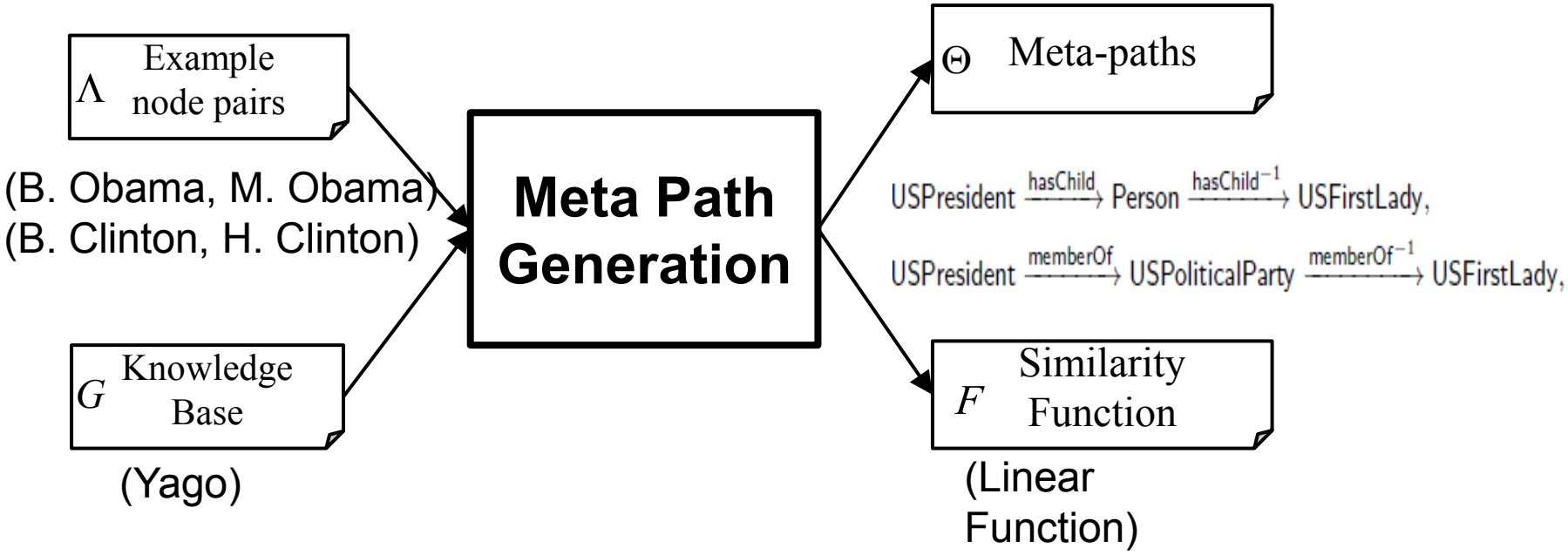
Existing work

- Designed by experts [Han VLDB'11]
  - Complex to analyze big data
  - Do not consider user-preference

- Enumeration within a given length [Cohen ECML'11]
  - Max length L is large, redundant (Curse of dimension)
    - In Yago, L=3, 135 meta-paths . L=4, 2000 meta-paths
  - L is small, miss some important ones

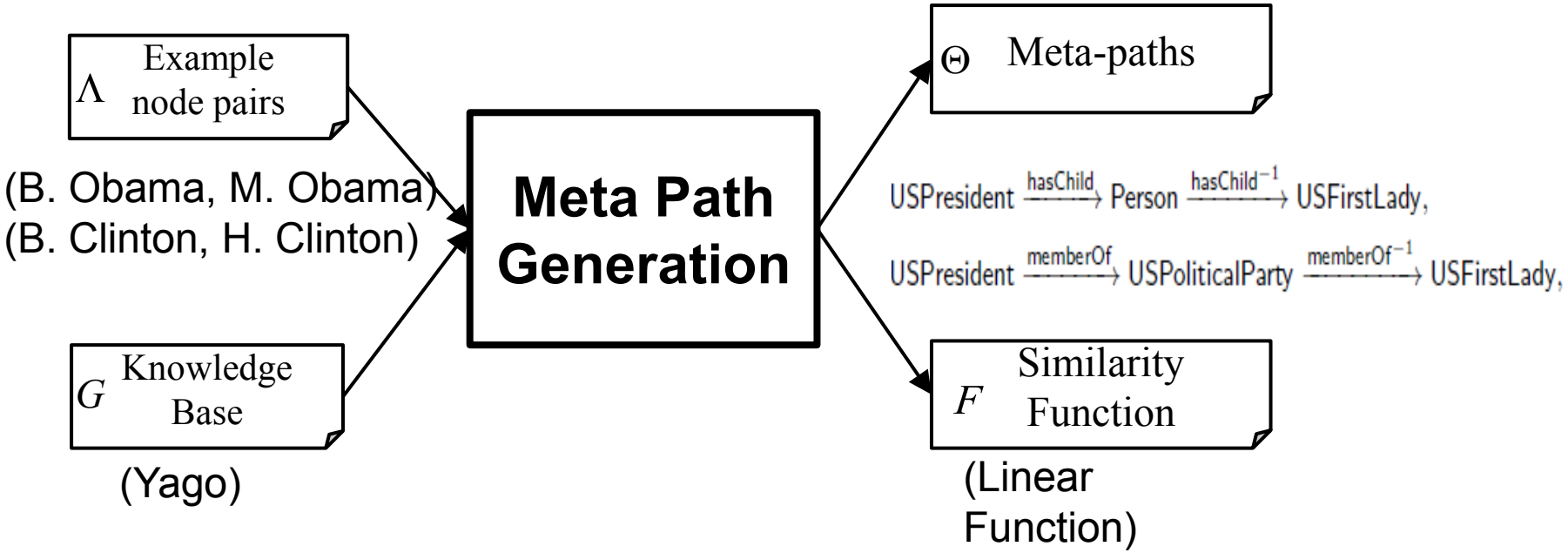
Problem Definition

Meta Paths Generation



Problem Definition

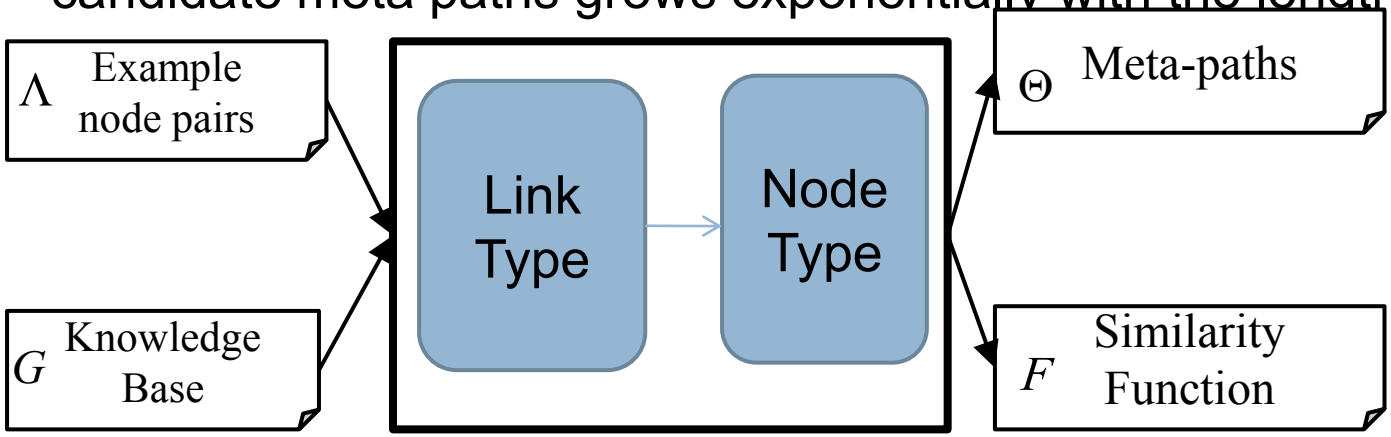
Meta Paths Generation



- Contributions
  - Consider the user's preference: let user input example pairs
  - Automatically generate meta paths without a max length: heuristic search instead of enumeration

Meta-path Generation

- **Two Phase Method**
  - **Challenge**: Each node has many class labels. The number of candidate meta paths grows exponentially with the length.



First Select Important Meta Path based on Links.

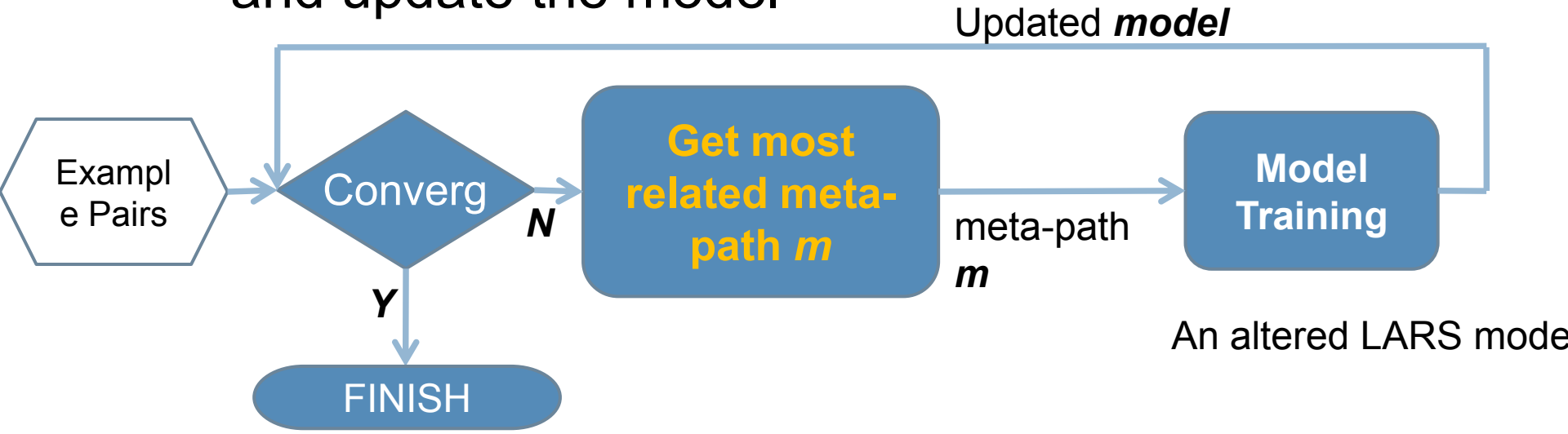
Next Refine the Node types

? liveIn ?

Scientist liveIn CapitalCity

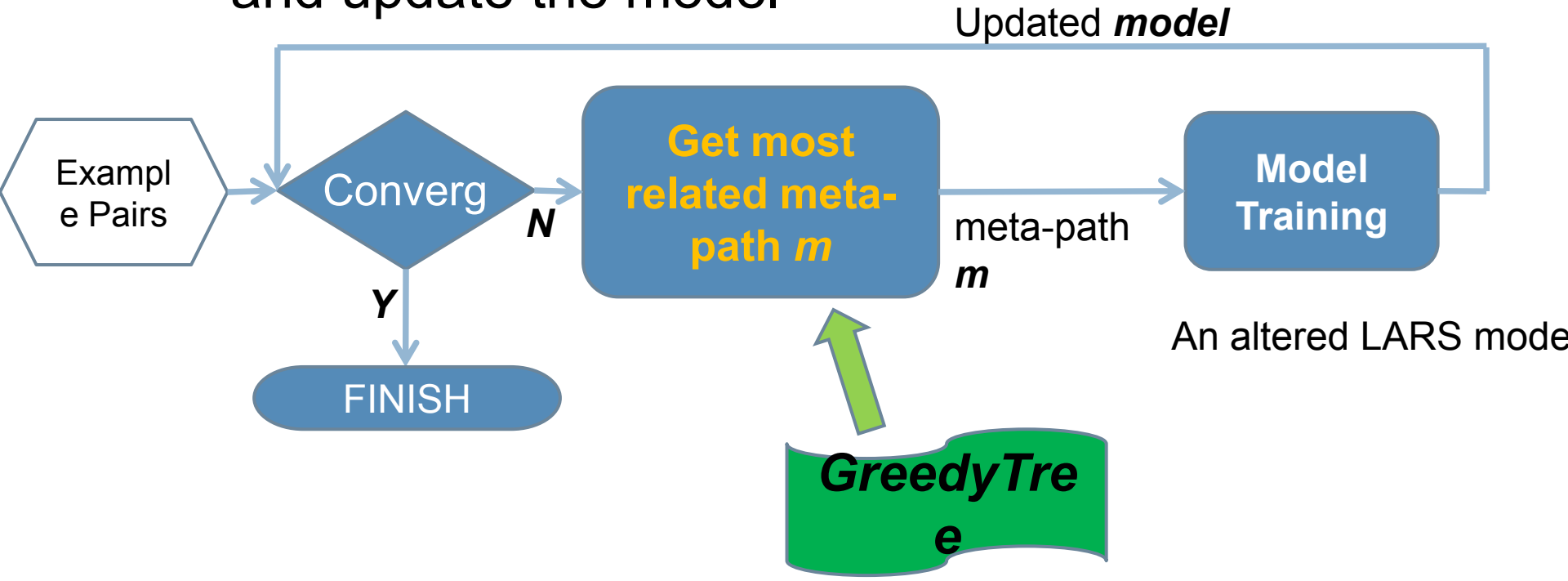
Generating Meta-Paths

- Phase 1: Link-Only Path Generation
  - ***Forward Stagewise Path Generation(FSPG)***: iteratively generate the most related meta-path and update the model



Generating Meta-Paths

- Phase 1: Link-Only Path Generation
  - ***Forward Stagewise Path Generation(FSPG)***: iteratively generate the most related meta-path and update the model



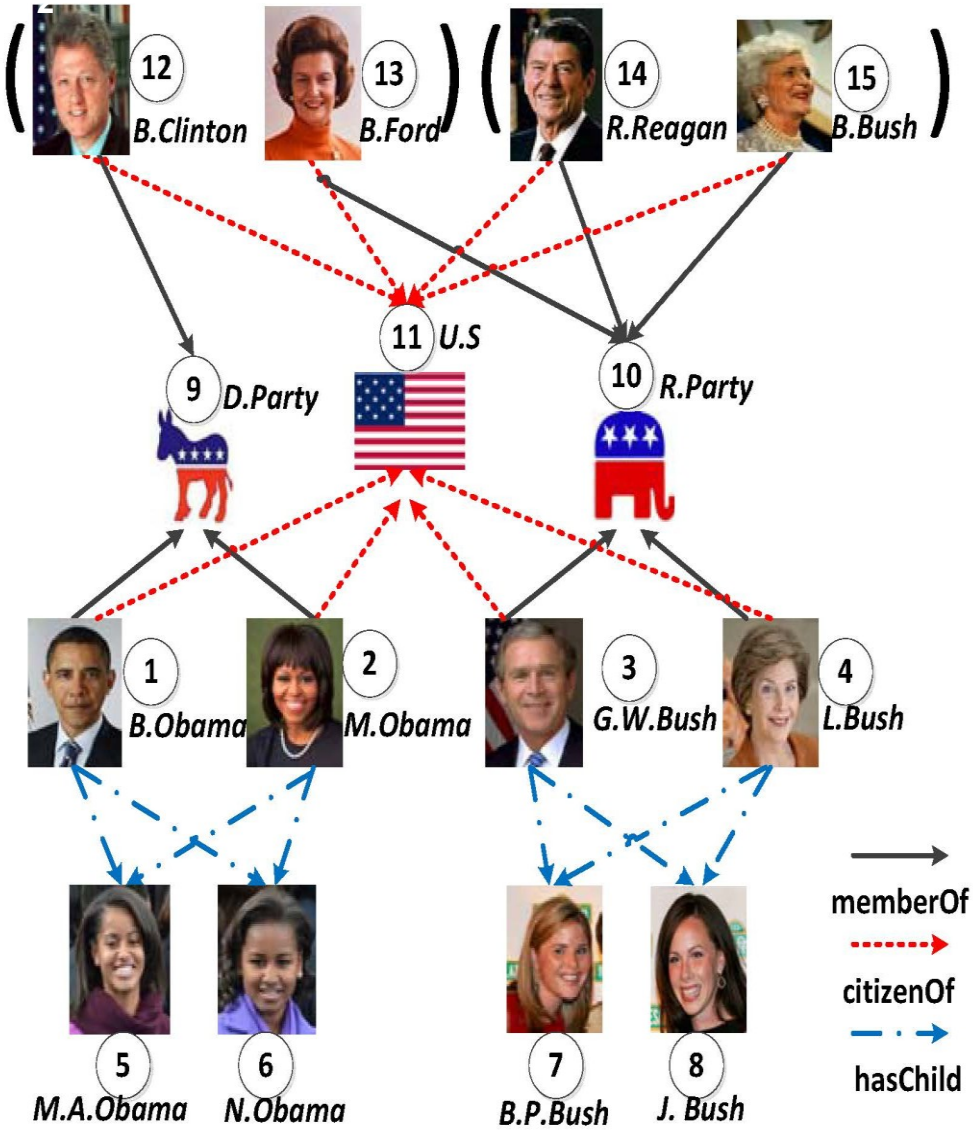
Generating Meta-Paths

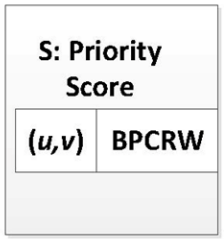
- GreedyTree
  - A tree that greedily expands the node which has the largest priority score
  - Priority Score : related to the correlation between ***m*** and ***r***
    - ***m*** is the meta path, ***r*** is the residual vector which evaluates the gap between the truth and current model

$$\cos(\mathbf{m},\mathbf{r}) = \frac{\mathbf{m}\cdot\mathbf{r}}{\|\mathbf{m}\| \times \|\mathbf{r}\|}$$

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

Generating Meta-Paths

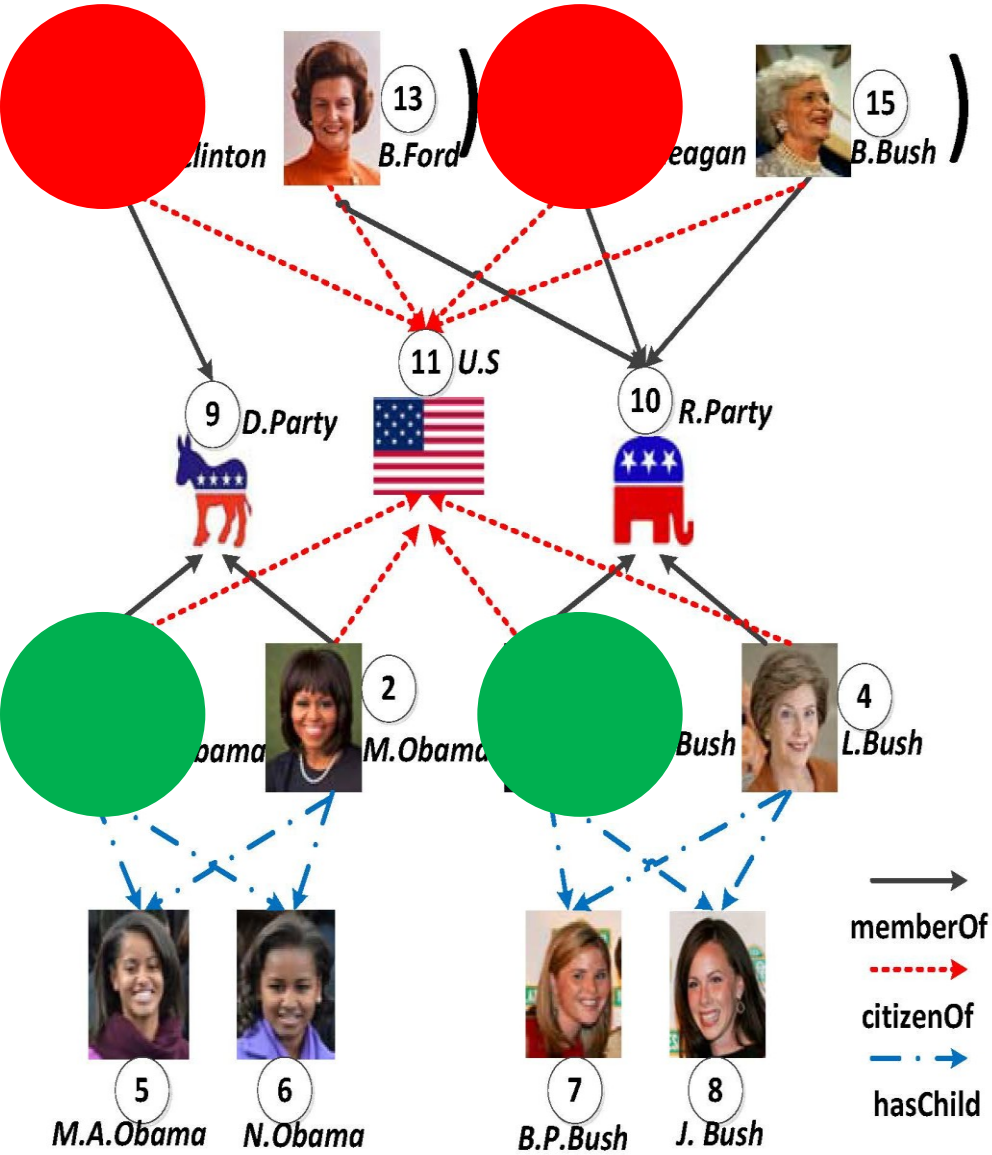




Node Structure

GreedyTree

Generating Meta-Paths



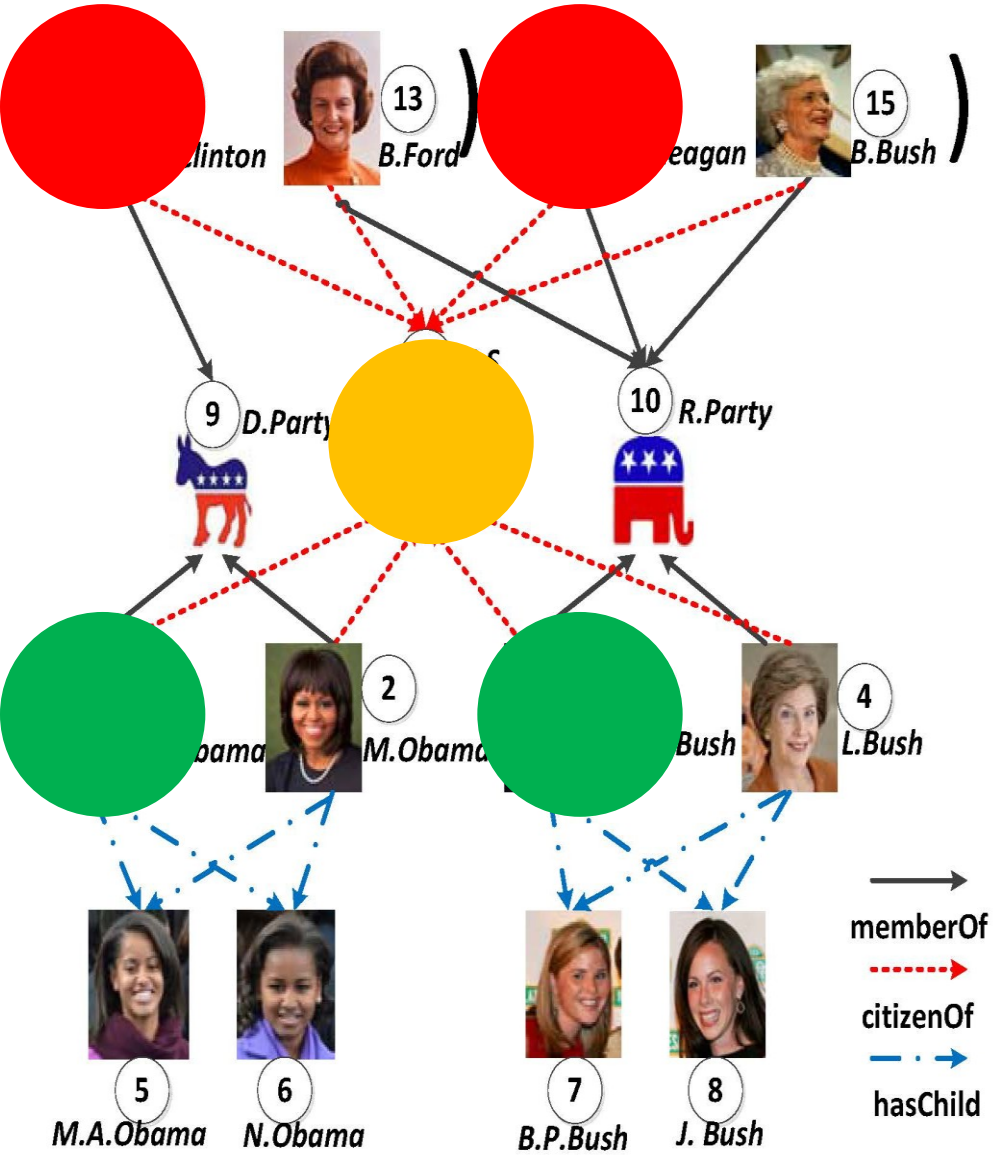
S: Priority Score	
(u,v)	BPCRW

Node Structure

S:0.5	
(1,1)	1
(3,3)	1
(12,12)	1
(14,14)	1

GreedyTree

Generating Meta-Paths



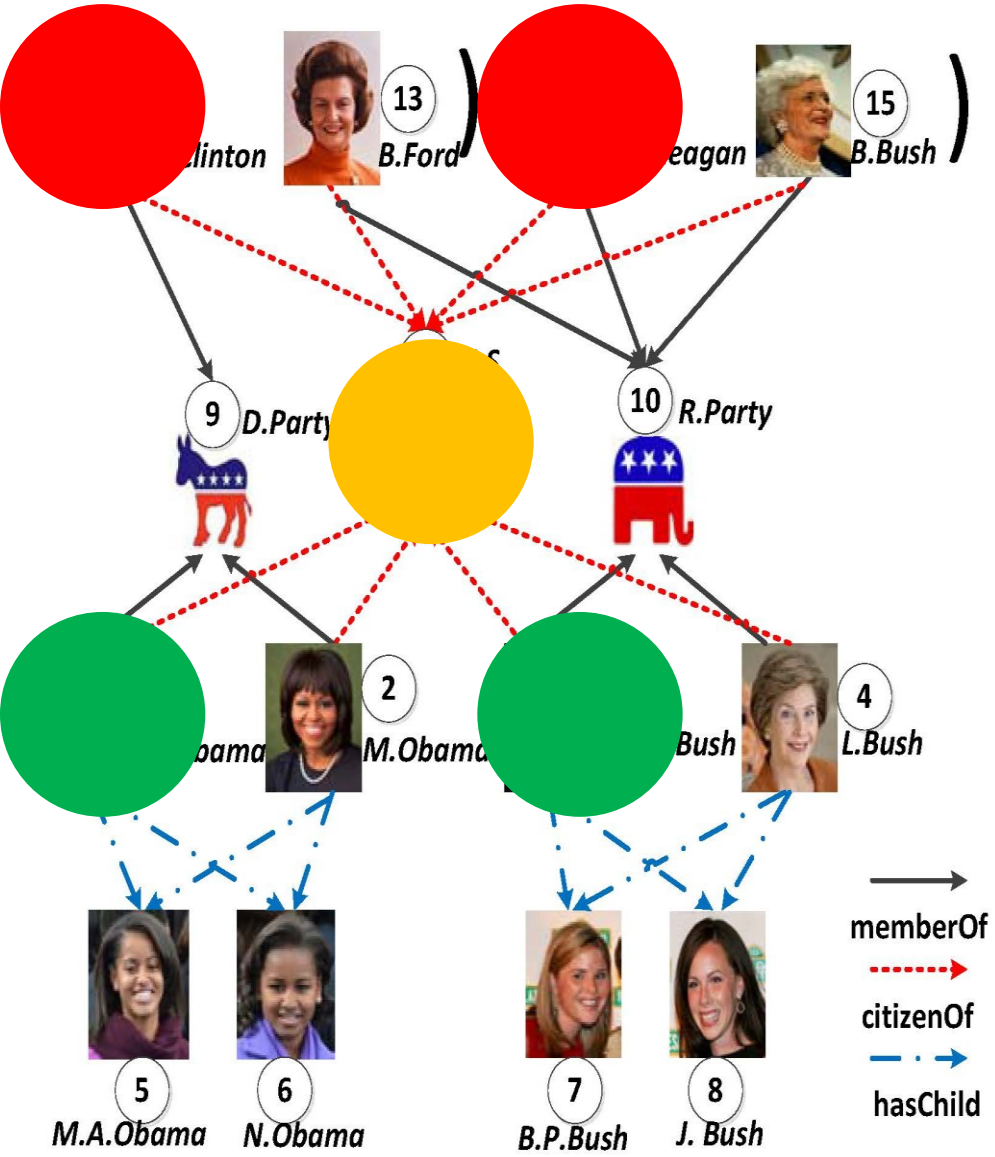
S: Priority Score	
(u,v)	BPCRW

Node Structure

S:0.5	
(1,1)	1
(3,3)	1
(12,12)	1
(14,14)	1

GreedyTree

Generating Meta-Paths



S: Priority Score	
(u,v)	BPCRW

Node Structure

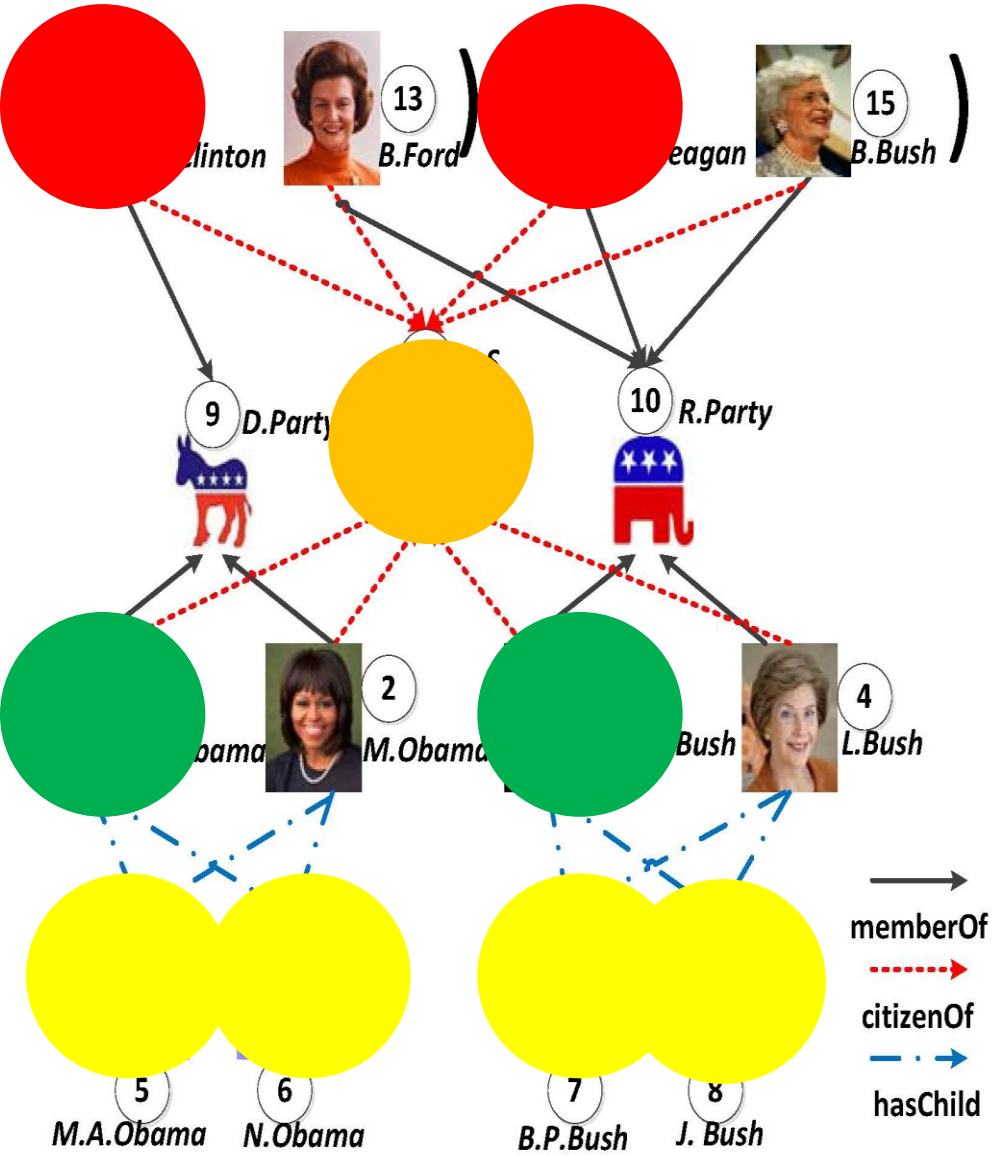
S:0.5	
(1,1)	1
(3,3)	1
(12,12)	1
(14,14)	1

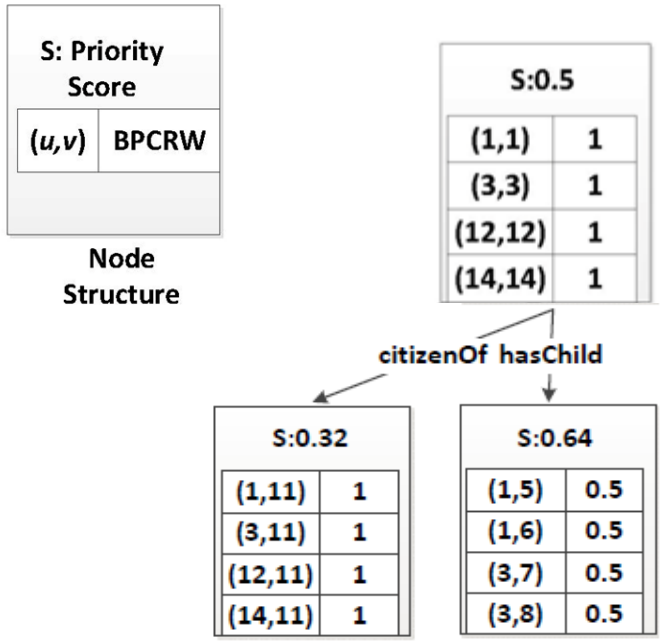
citizenOf

S:0.32	
(1,11)	1
(3,11)	1
(12,11)	1
(14,11)	1

GreedyTree

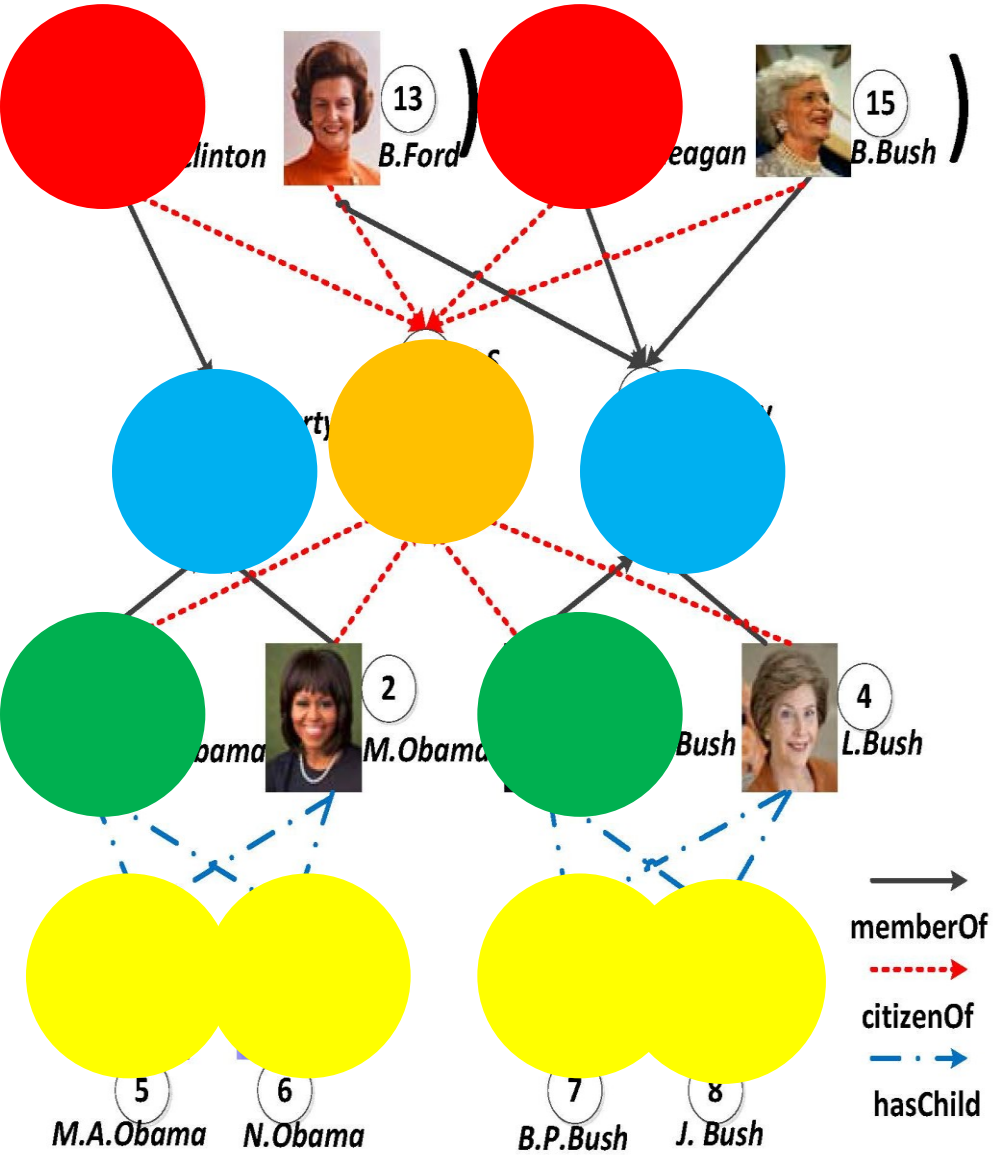
Generating Meta-Paths

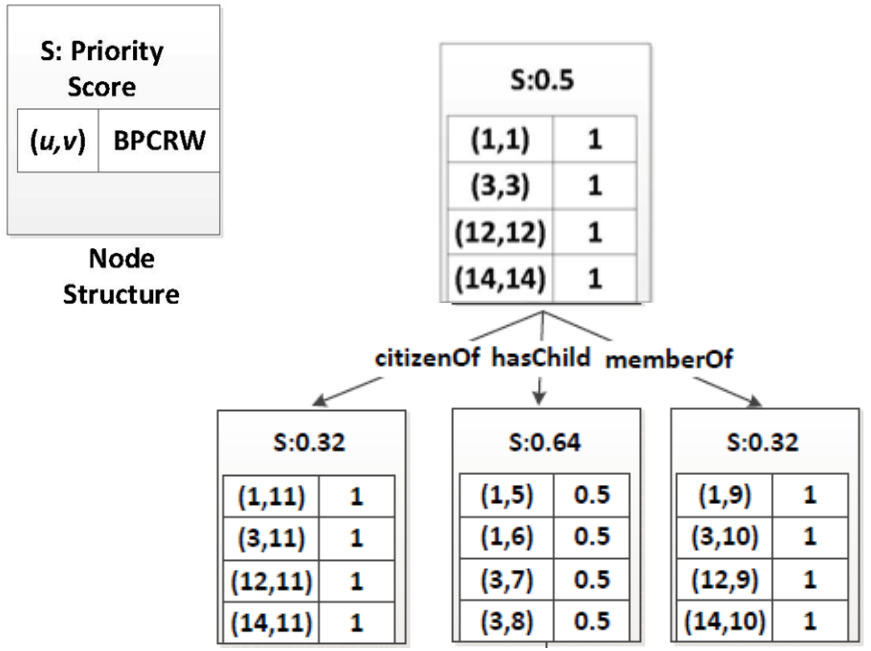




GreedyTree

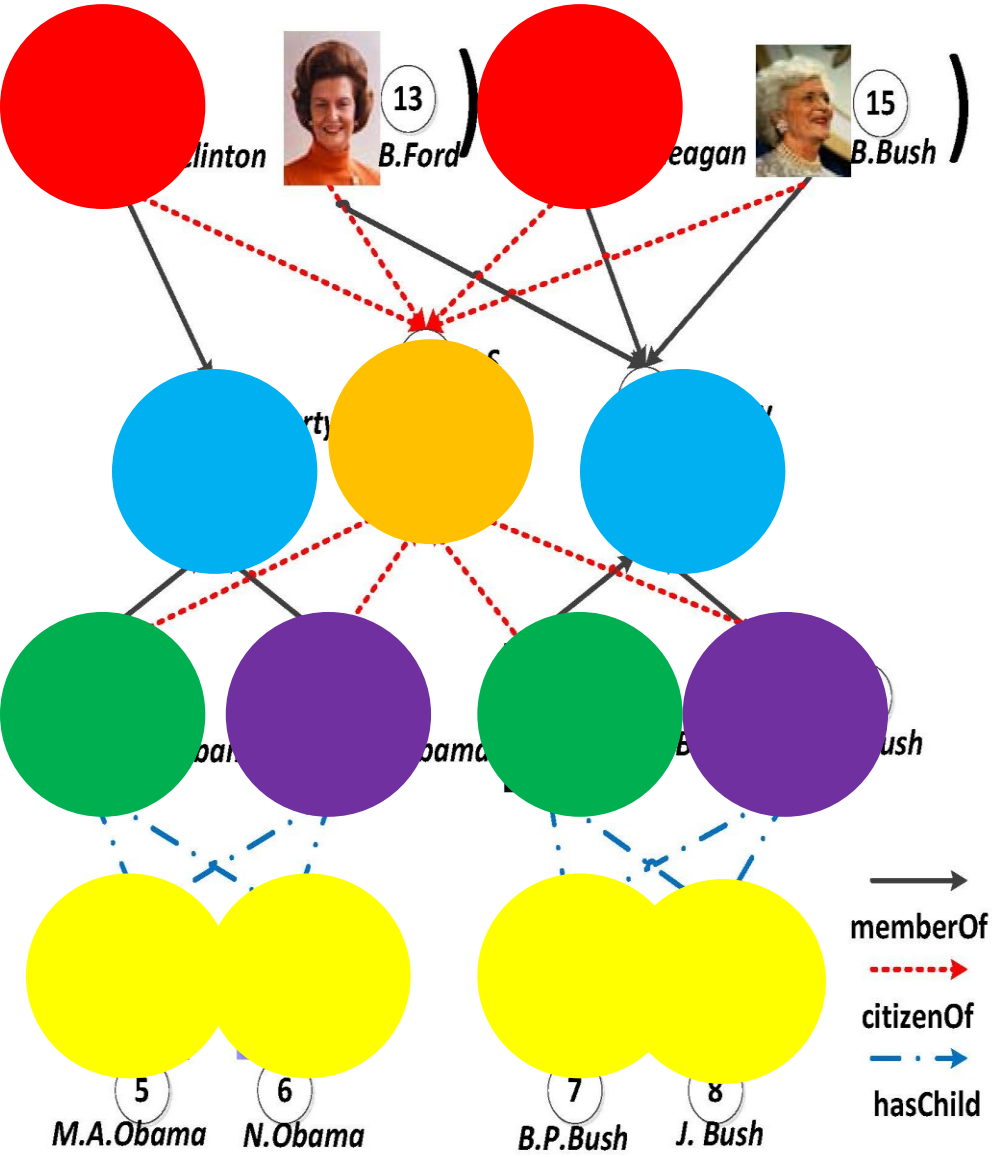
Generating Meta-Paths





GreedyTree

Generating Meta-Paths



S: Priority Score

(u,v)	BPCRW

Node Structure

S:0.5

(1,1)	1
(3,3)	1
(12,12)	1
(14,14)	1

citizenOf hasChild memberOf

S:0.32

(1,11)	1
(3,11)	1
(12,11)	1
(14,11)	1

S:0.64

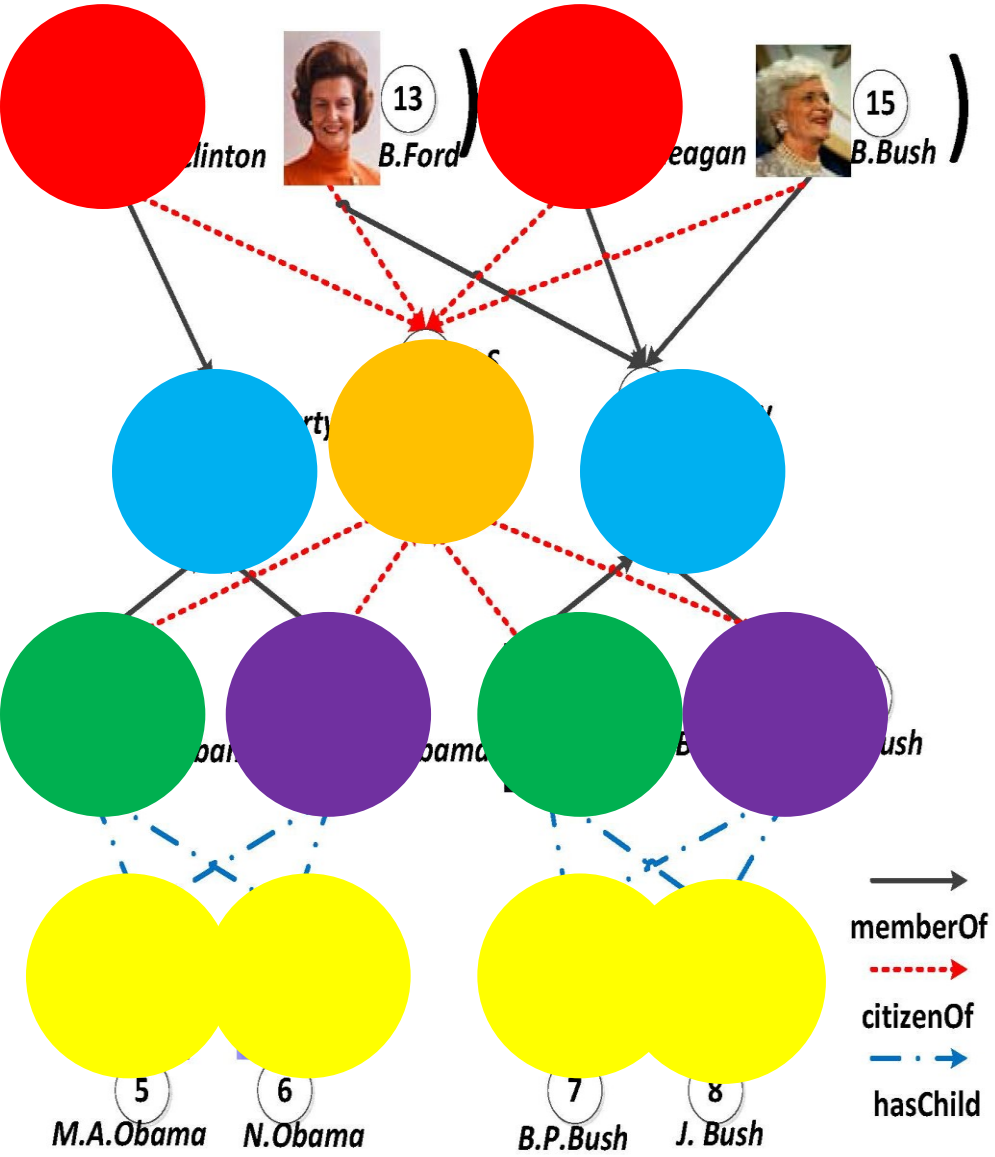
(1,5)	0.5
(1,6)	0.5
(3,7)	0.5
(3,8)	0.5

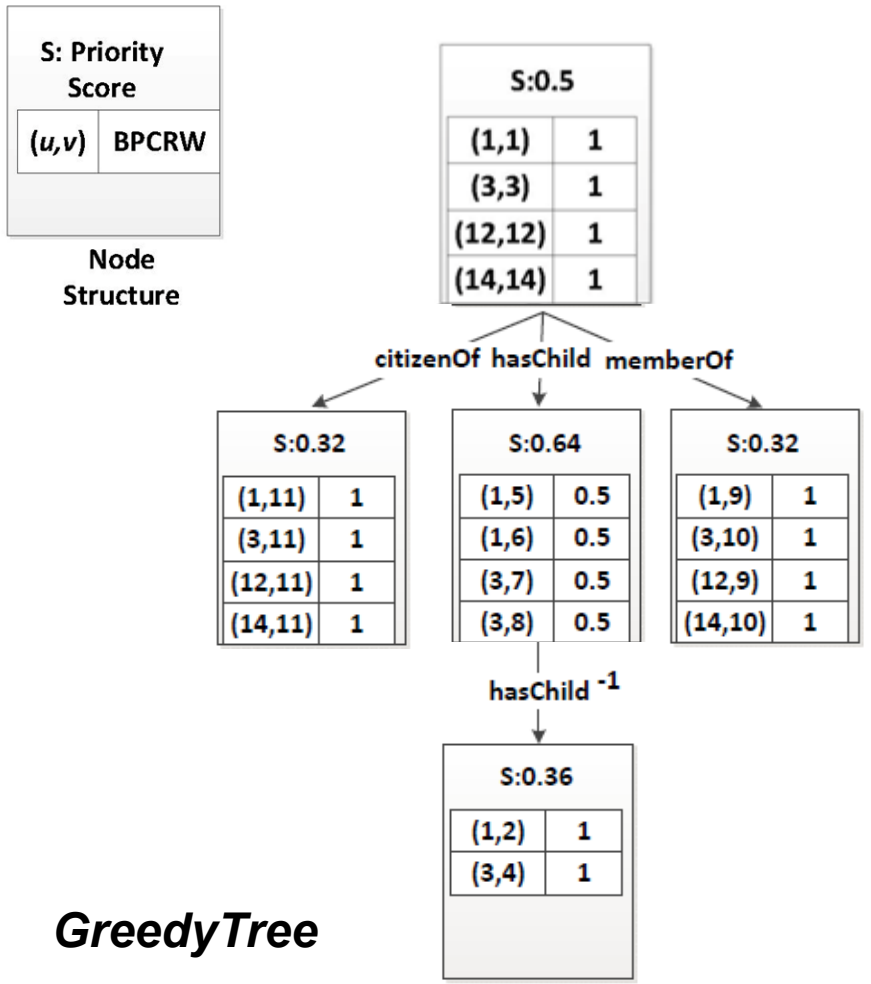
S:0.32

(1,9)	1
(3,10)	1
(12,9)	1
(14,10)	1

GreedyTree

Generating Meta-Paths



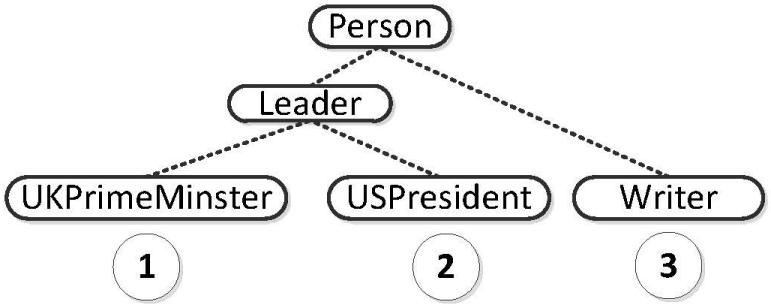


Meta-path Generation

- Phase 2: Node Class Generation
  - Why node classes are needed
    - A link only meta path may introduce some unrelated result pairs.

$? \xrightarrow{\text{liveIn}} ?$ is less specific tha $\text{Scientist} \xrightarrow{\text{liveIn}} \text{CapitalCity}$

  - Solution 1: Lowest Common Ancestor(LCA)
    - Record the LCA in the node of GreedyTree



Meta-path Generation

- Solution 3: TFOF
  - $score(\phi) = \frac{tf(\phi)}{\log of(\phi)}$ ***tf*** is the frequency of label in positive examples. ***of*** is the overall count in KB

$tf(\text{USPresident})$ =2 $of(\text{USPresident})$ =42

$score(\text{USPresident}) = 1.23$

  - ***of*** can be pre-computed

Meta-path Generation

- Solution 3: TFOF
  - $score(\phi) = \frac{tf(\phi)}{\log of(\phi)}$ ***tf*** is the frequency of label in positive examples. ***of*** is the overall count in KB

$tf(\text{USPresident})$ =2 $of(\text{USPresident})$ =42

$score(\text{USPresident}) = 1.23$

S:0.5	
(1,1)	1
(3,3)	1
(12,12)	1
(14,14)	1

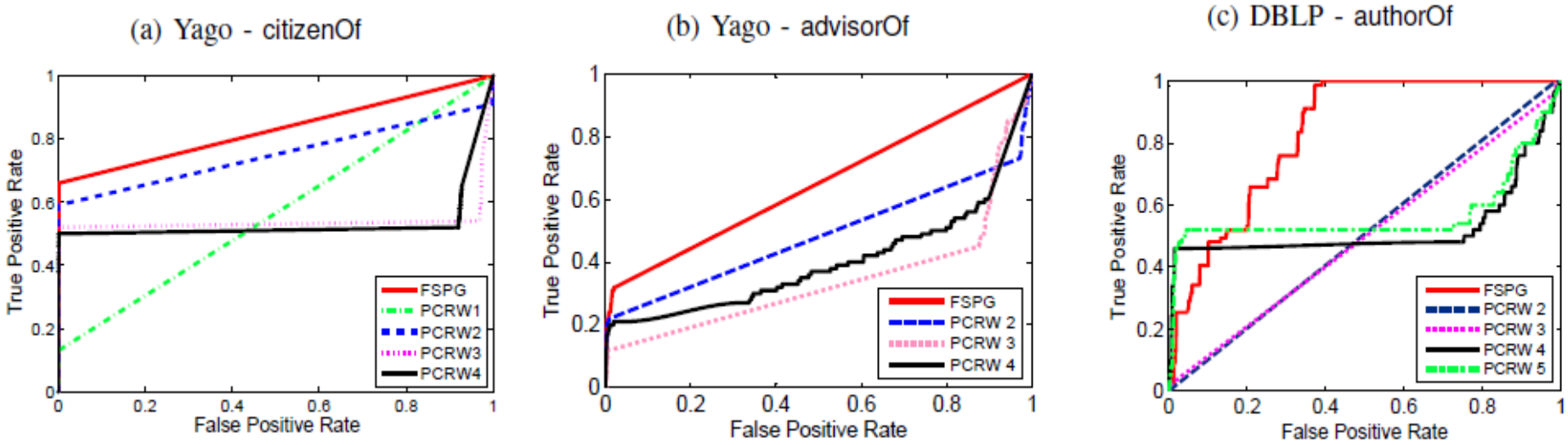
  - ***of*** can be pre-computed

Experiments

- Dataset
  - DBLP (four areas)
    - (Database, Data Mining, Articial Intelligence and Information Retrieval).
    - 14376 papers, 14475 authors, 8920 topics, 20 venues.
  - Yago
    - A Knowledge Base 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

Experiments

- Effectiveness

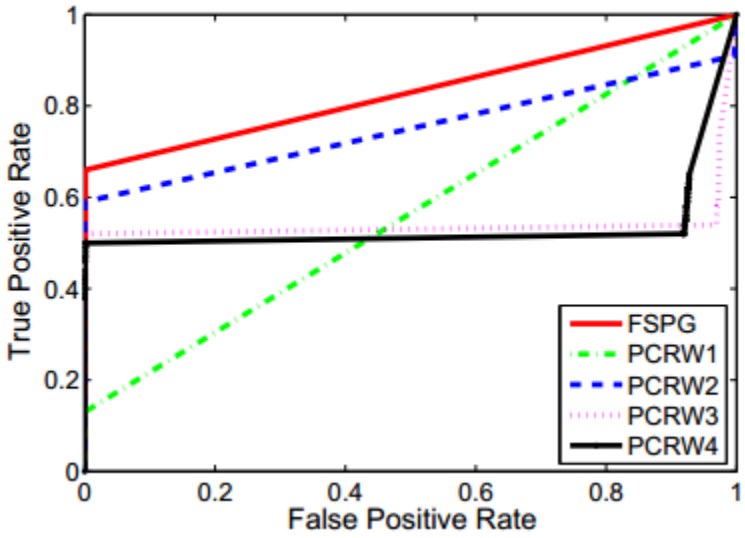


ROC for link prediction

- Baseline method: enumerating all meta-paths within a given max length L.
  - L is small, low recall.
  - L is large, low precision.

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

(a) Yago - citizenOf



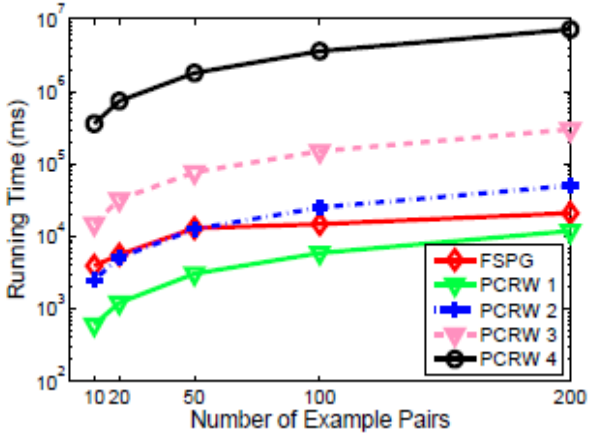
meta-path	w
Person $\xrightarrow{\text{bornIn}}$ City $\xrightarrow{\text{locatedIn}}$ Country	5.477
Person $\xrightarrow{\text{livesIn}}$ Country	0.361
Person $\xrightarrow{\text{graduateOf}}$ University $\xrightarrow{\text{locatedIn}}$ Country	0.023
Person $\xrightarrow{\text{diedIn}}$ City $\xrightarrow{\text{locatedIn}}$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

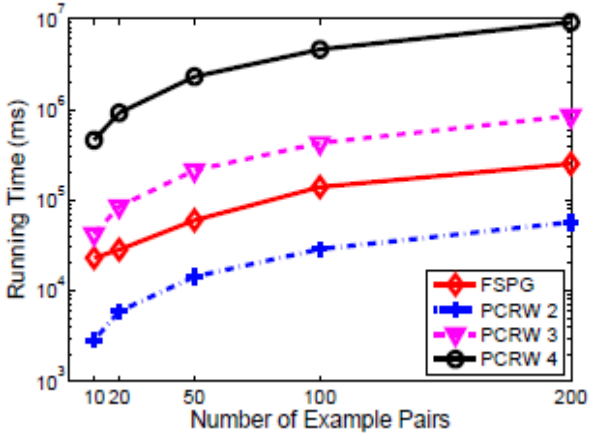
Experiments

- Efficiency

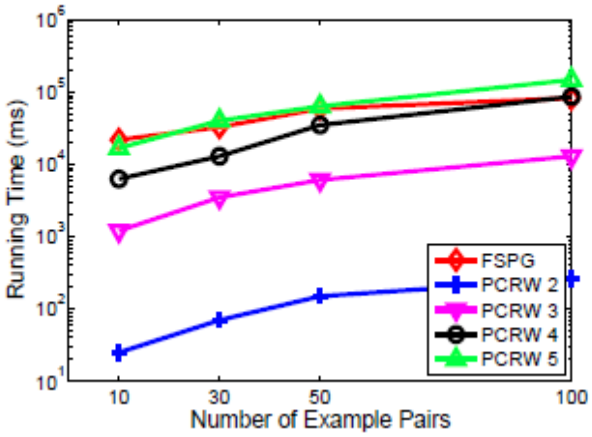
(a) Yago - citizenOf



(b) Yago - advisorOf



(c) DBLP - authorOf



Running time of FSPG

- In Yago, 2 orders of magnitude better.
- In DBLP, the running time is comparable to PCRW 5, but the accuracy is much better.

Conclusion

- We examined a novel problem of meta-paths generation which is highly needed to analyze and query KB.
- We proposed the ***FSPG*** algorithm, and developed ***GreedyTree*** to facilitate its execution.

References

- [Yu EDBT'13]:] C. Shi , P. Yu"Relevance Search in Heterogeneous Network" EDBT'13
- [Han VLDB'11] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks", VLDB'11
- [Han ASONAM'11] Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks", ASONAM'11
- [Han CIMK'12]Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "Meta-Path Selection with User Guided Object Clustering in Heterogeneous Information Networks ", CIKM'12
- [Sun VLDB'11]L. Sun, R. Cheng and etc On Link-based Similarity join, VLDB 2011
- [Han KDD'12]J. Han, Y. Sun, X. Yan, and P. S. Yu, "Mining Heterogeneous Information Networks"
- [Cohen ECML'11]W. Cohen, N. Lao "Relational Retrieval Using a Combination of Path-Constrained Random Walks" ECML 2011
- [Weinberger JMLR'09] Weinberger "Distance Metric Learning for Large Margin Nearest Neighbor Classification" Journal of Machine Learning Research 10 (2009) 207-244
- [Chopra CVPR'05]S.Chopra "Learning a similarity metric discriminatively, with application of face verification" CVPR 2005

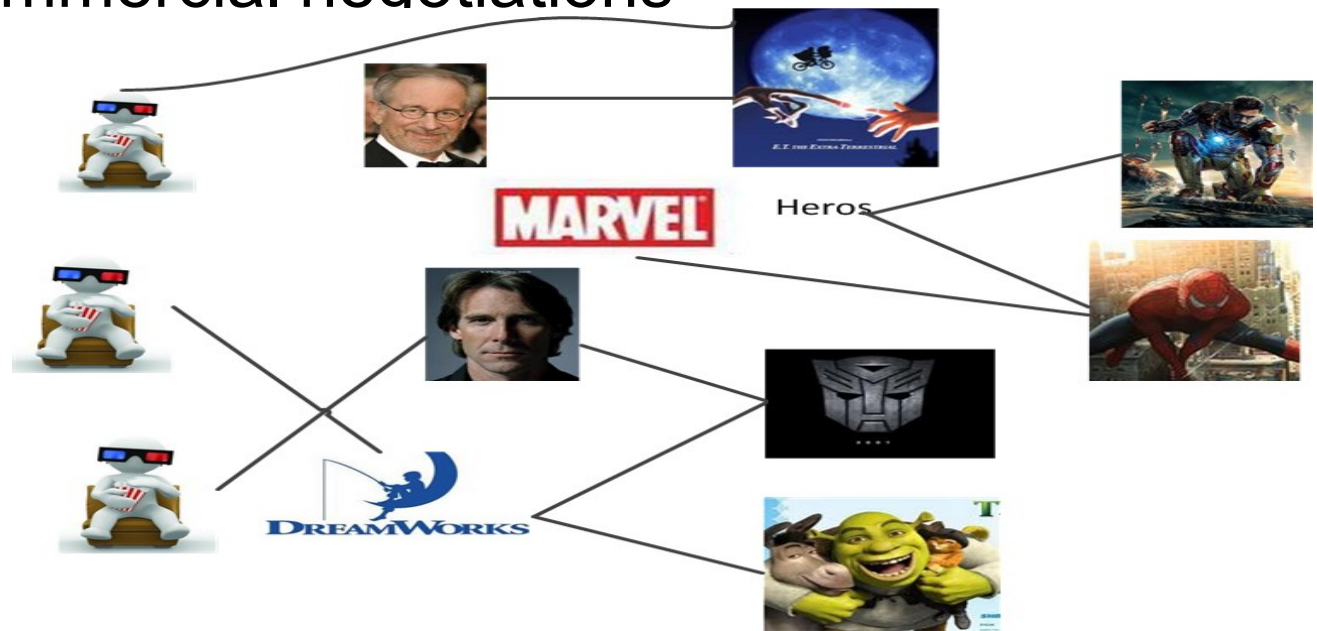
Thank you!

Changping Meng
meng40@purdue.edu

Introduction

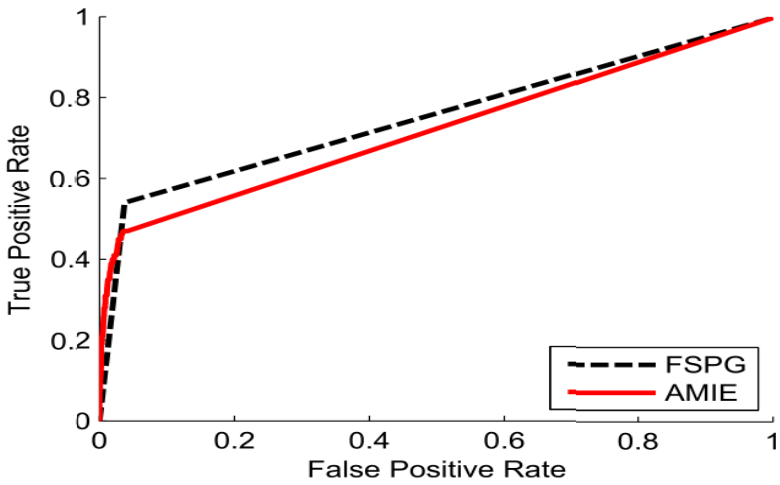
Applications

- 3. Recommendation
  - Promote Movies for customers
  - Choose representatives to Political or Commercial negotiations



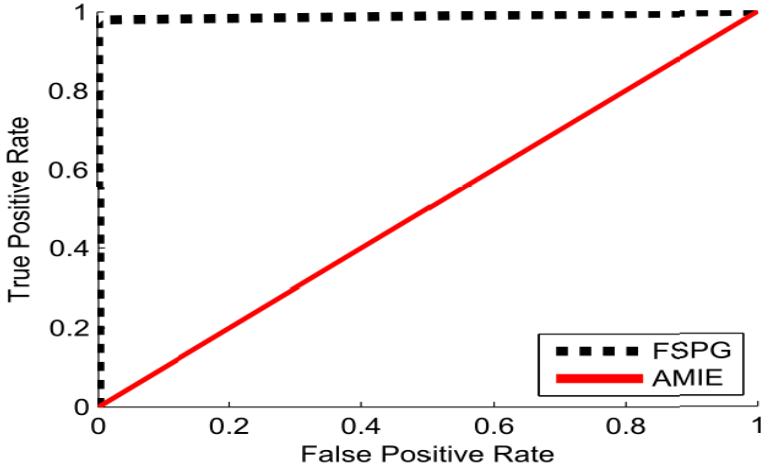
Association Rule mining compared with AMIE (Luis, WWW'13)

advisorOf



meta-path
$\text{Person} \xrightarrow{\text{influence}^{-1}} \text{Person}$
$\text{Person} \xrightarrow{\text{diedIn}} \text{City} \xrightarrow{\text{diedIn}^{-1}} \text{Person}$
$\text{Person} \xrightarrow{\text{workAt}} \text{University} \xrightarrow{\text{graduatedFrom}} \text{Person}$
$\text{Person} \xrightarrow{\text{liveIn}} \text{Country} \xrightarrow{\text{isCitizenOf}^{-1}} \text{Person}$

ivyLeagueAlumnus



AMIE does not consider the hierarchy of node types.
Failed to distinguish Ivy League alumni from the alumni of any other universities

Experiments

- Class label selection
  - The TFOF method of generating class labels is better for high precision queries
  - LCA is better than TFOF for higher recal rates.