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

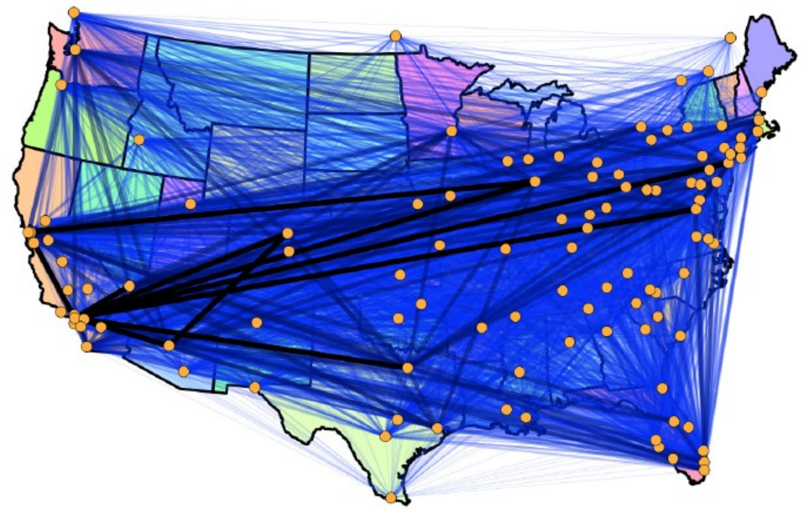
Yangqiu Song

Department of CSE, HKUST, Hong Kong

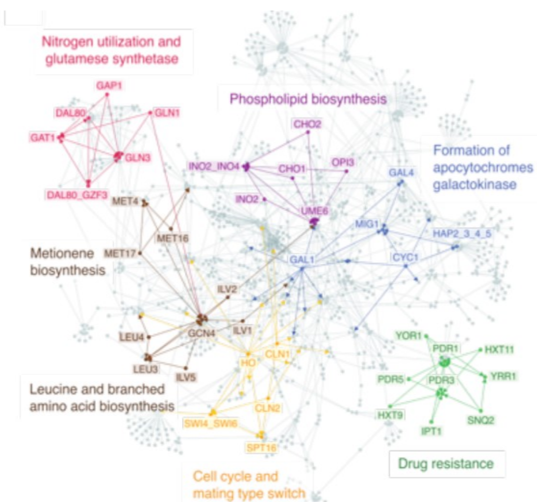
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Homogeneous Graph/Networks

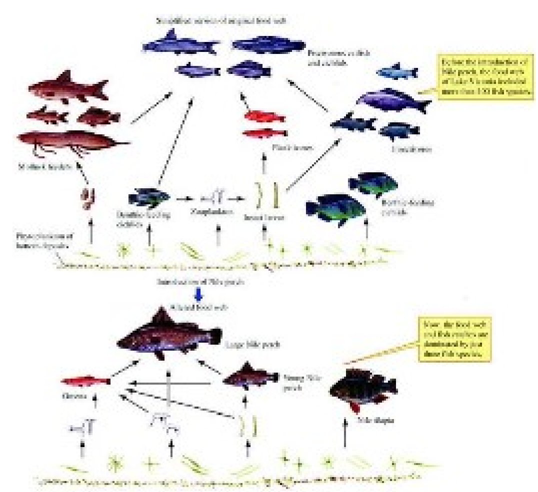
Social Network



Transportation Network



Gene Network

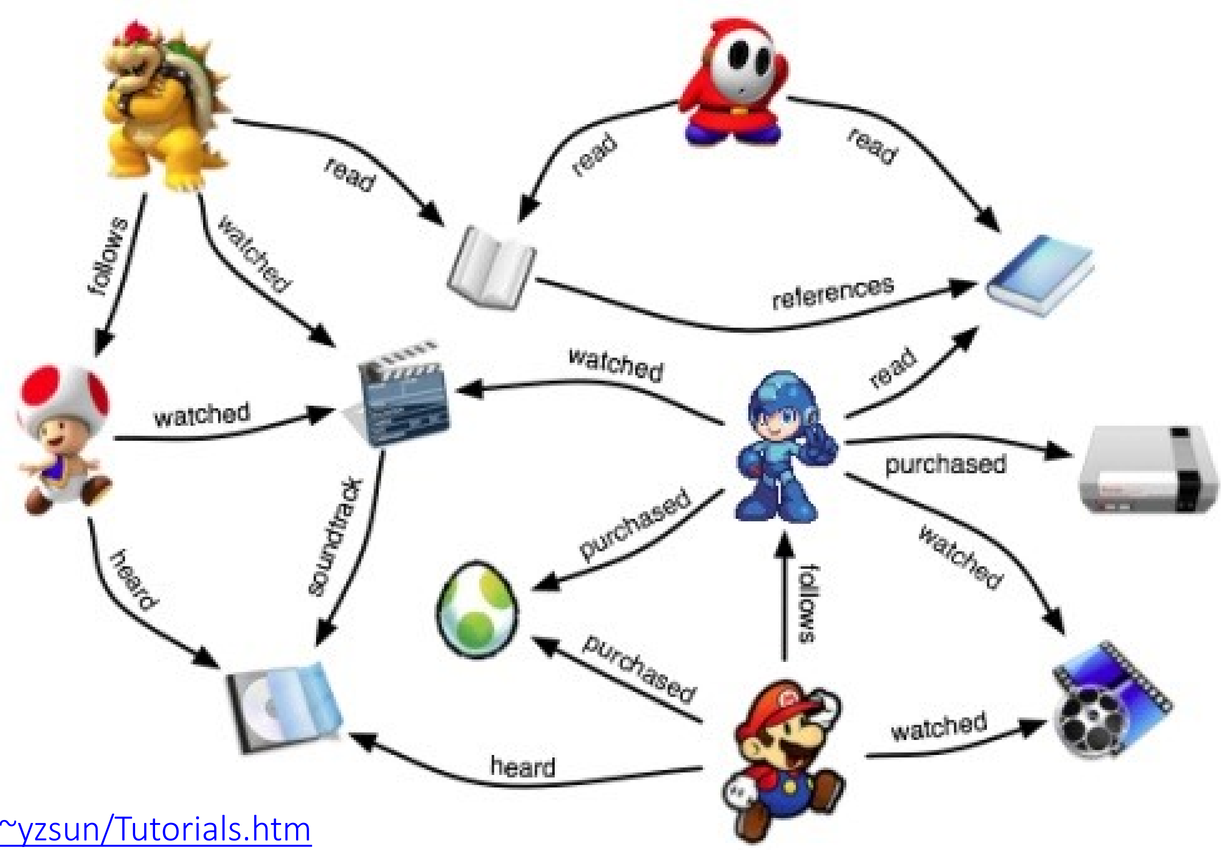


Food Network

http://snap.stanford.edu/higher-order/higher-order-SM-science16.pdf

Heterogeneous Information Networks

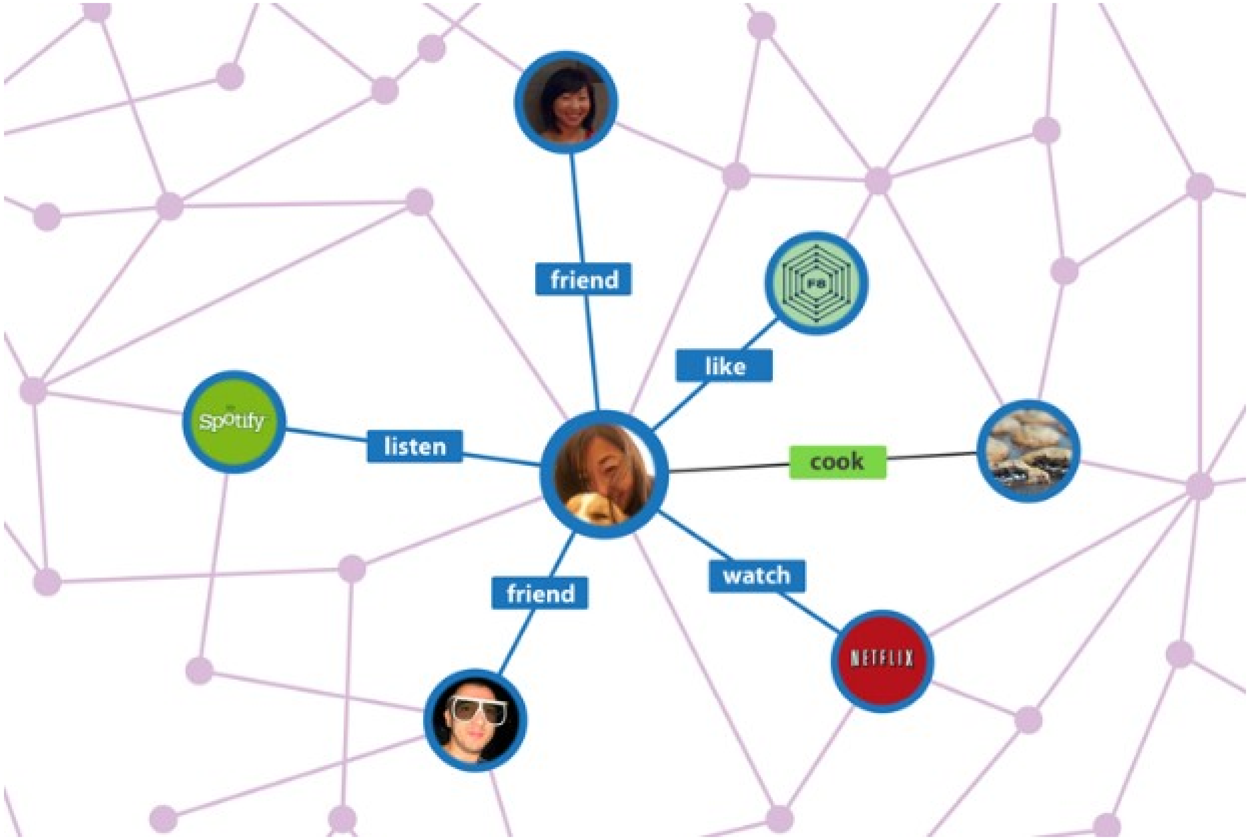
- Yizhou Sun, Jiawei Har)09-2012 (UIUC)
  - Entity type mappinc
  - Link type mapping: $\mathrm{E} \text{ -> } \mathrm{R}$



http://web.cs.ucla.edu/~yzsun/Tutorials.htm

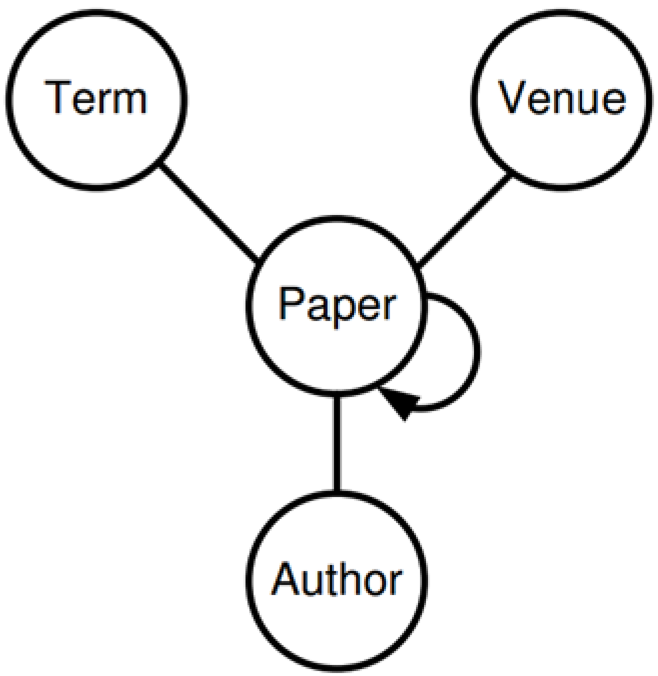
Modern Social Media

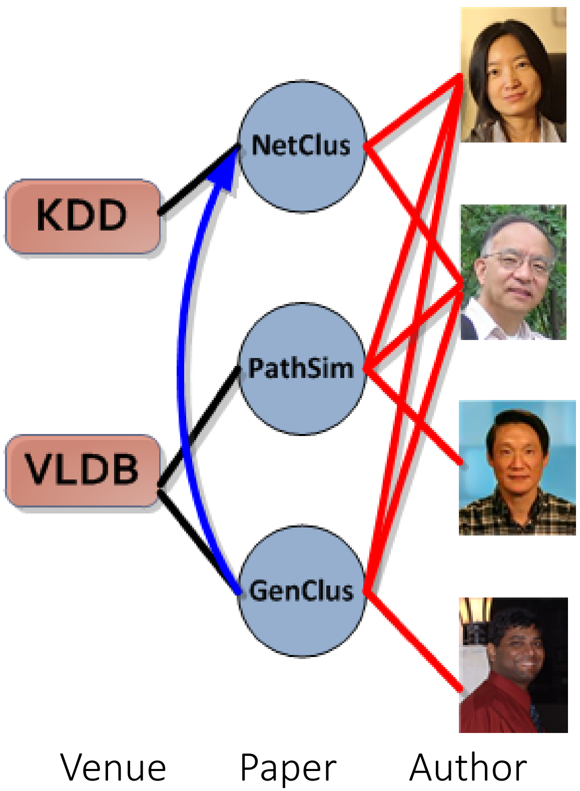
- Entities: Person, Check-in location, Articles, etc.
- Relations: Friends, Like, Check-in, etc.



Scholar Networks

- Entities: Paper, Venue, Author, Keyword, etc.
- Relations: Write, Attend, Contain, etc.



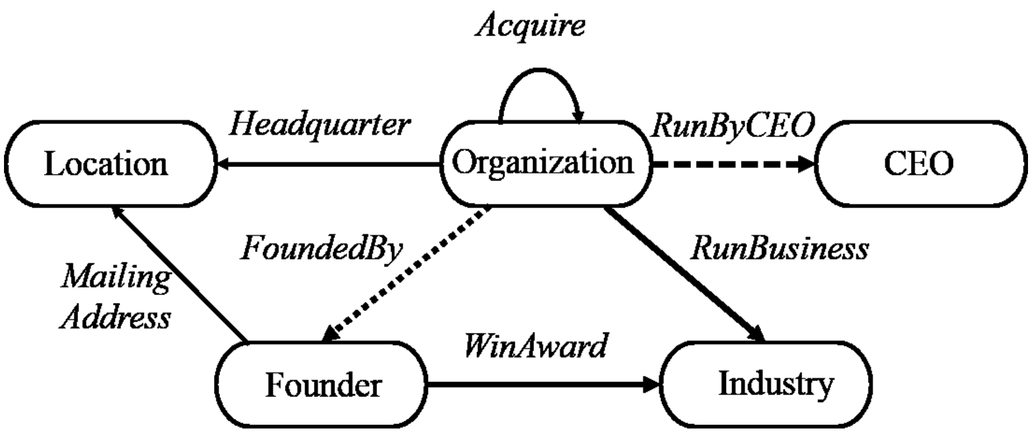


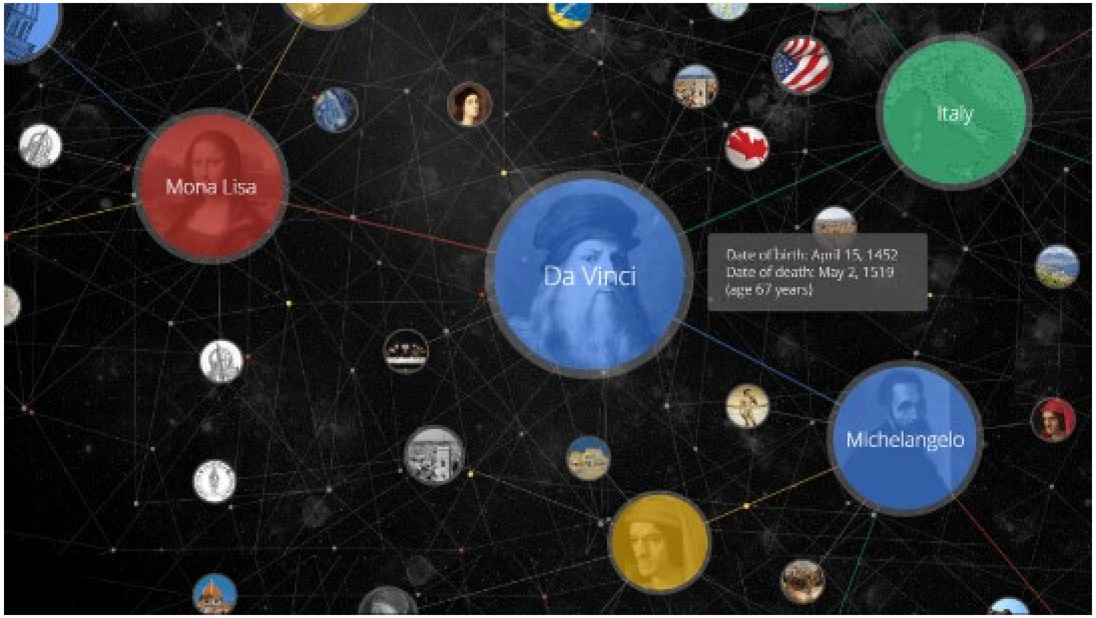
DBLP Bibliographic Network

http://web.cs.ucla.edu/~yzsun/Tutorials.htm

Knowledge Graphs

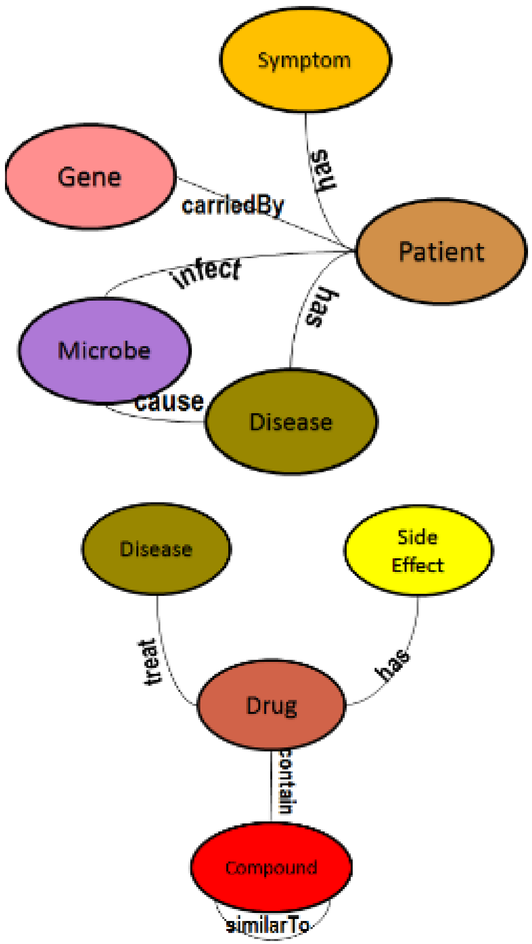
- Example of entities and their relations:

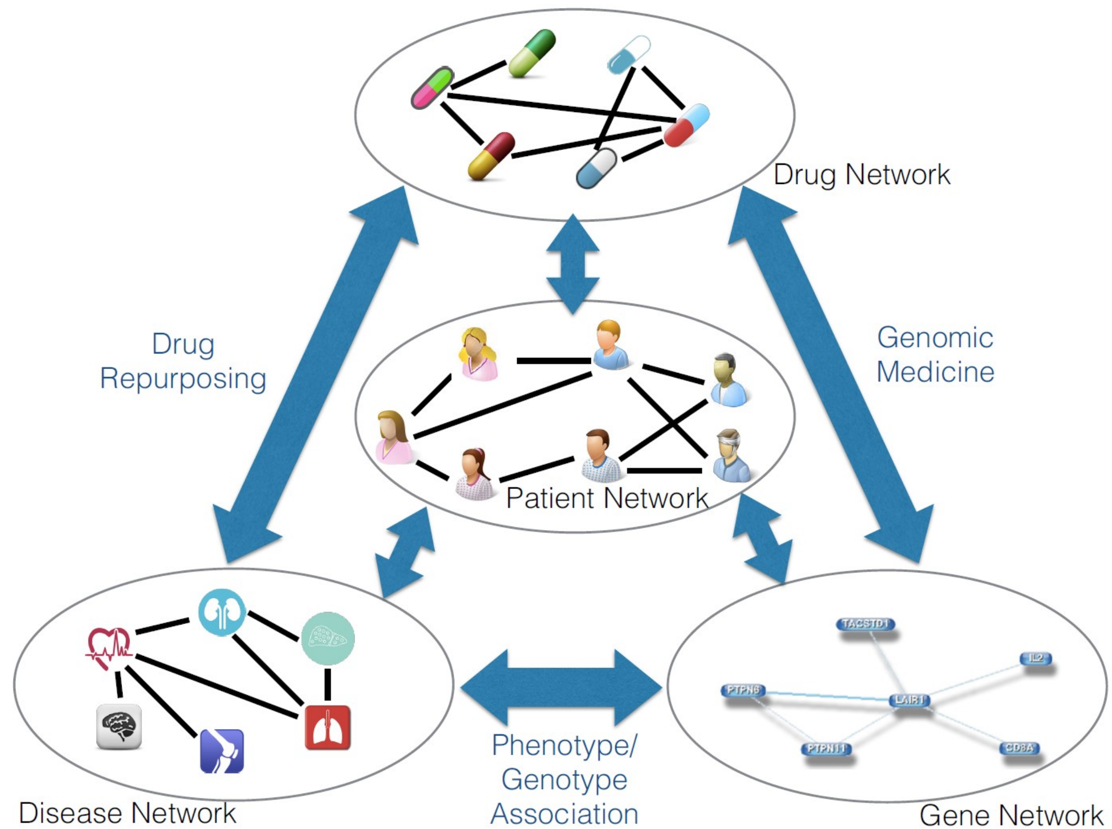




Bio-medical Network

- Entities: Gene, Patient, Drug, Disease, etc.
- Relations: Drug repurposing, Genotyping, etc.

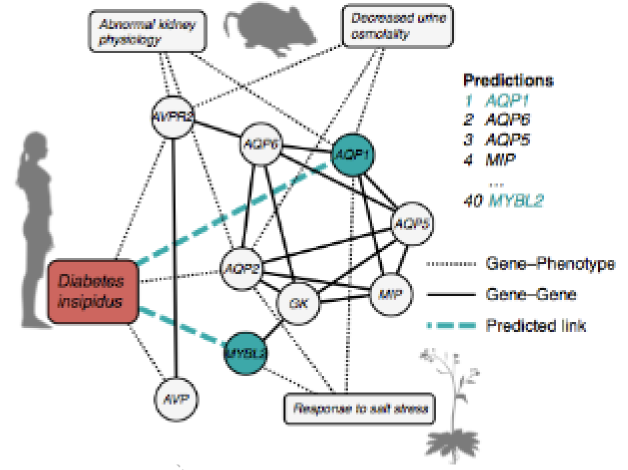




http://web.cs.ucla.edu/~yzsun/Tutorials.htm
https://sites.google.com/site/feiwang03/talks

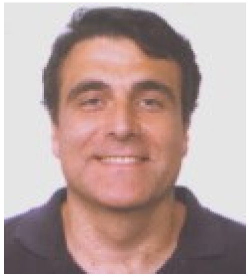
Problems in HIN

- Link Prediction
  - Homogeneous
  - Heterogeneous: recommendation



- Entity Typing/Profiling

- Similarity Search



Meta-Path: *Author-Paper-Author*

Rank	Author	Score
1	Christos Faloutsos	1
2	Spiros Papadimitriou	0.127
3	Jimeng Sun	0.12
4	Jia-Yu Pan	0.114
5	Agma J. M. Traina	0.110
6	Jure Leskovec	0.096
7	Caetano Traina Jr.	0.096
8	Hanghang Tong	0.091
9	Deepayan Chakrabarti	0.083
10	Flip Korn	0.053

Christos' students or close collaborators

http://bigdata.ices.utexas.edu/project/gene-disease/
http://web.cs.ucla.edu/~yzsun/Tutorials.htm
http://xren7.web.engr.illinois.edu/tutorial.html

Explicit vs. Implicit "Flat" Semantics

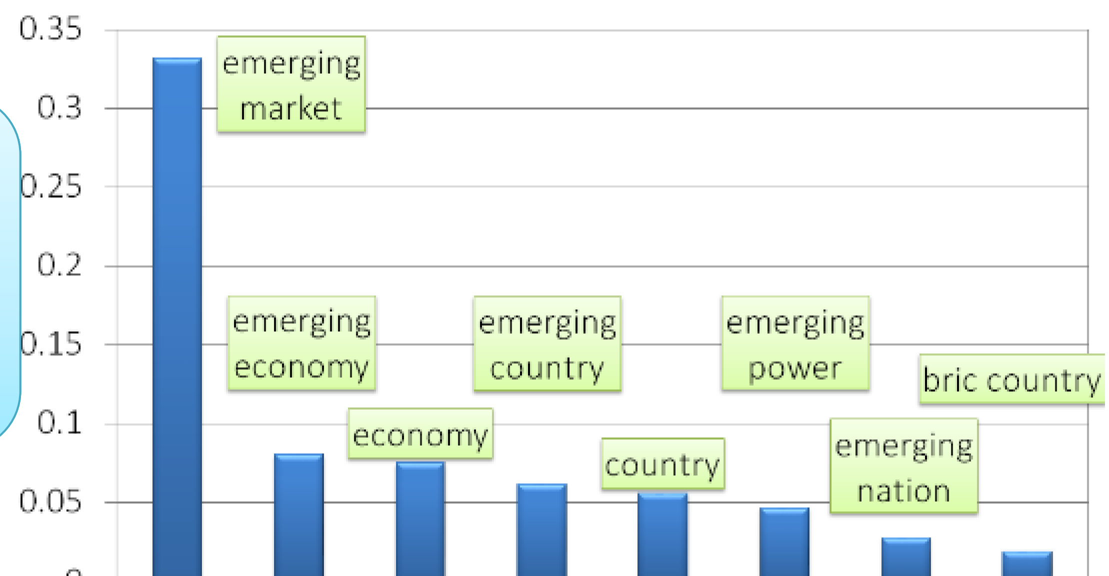
- Explicit Semantic Analysis [Gabrilovich and Markovitch '06, '07, '09]

Represent text as bag of Wikipedia titles

Barack Obama
Timeline of the presidency of Barack Obama (2009)
Family of Barack Obama
Barack Obama citizenship conspiracy theories
Barack Obama
Barack Obama presidential primary campaign 2008

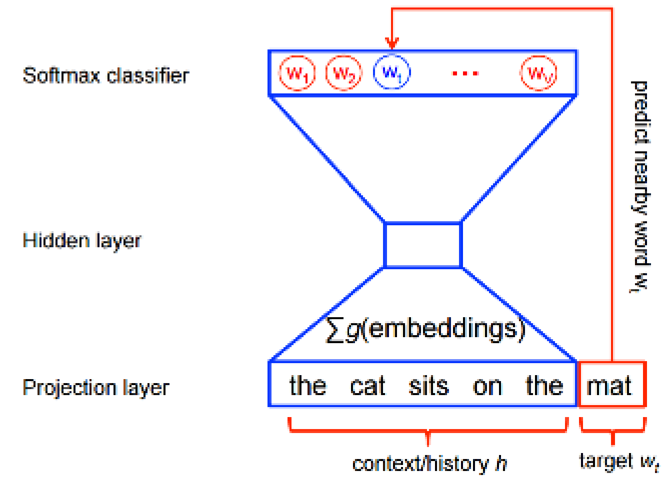
- Probabilistic Conceptualization [Song et al., '11,'15]

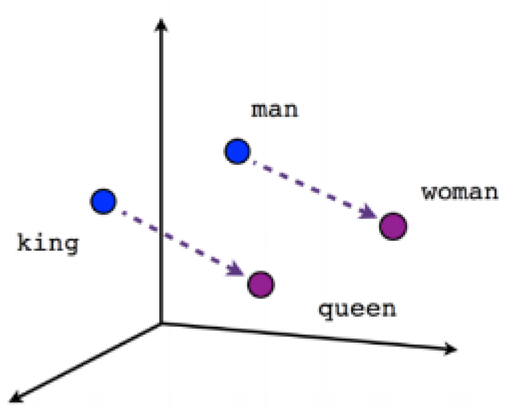
Given "China, India, Russia, Brazil", retrieve concepts from Probase [Wu et al., SIGMOD'12]



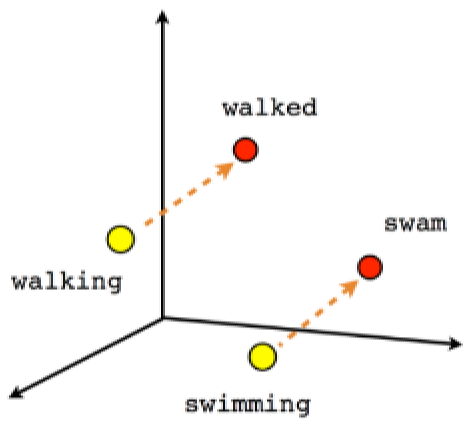
Explicit vs. Implicit "Flat" Semantics

- Implicit Semantic Analysis
  - SVD [Deerwester et al., JASIS'90]
  - PLSA [Hofmann, NIPS'99]
  - LDA [Blei et al., JMLR'03]
  - Word2vec [Mikolov et al., NIPS'13]
  - …

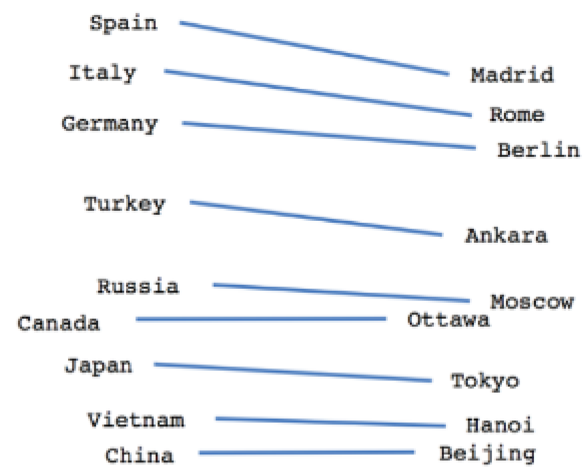




Male-Female



Verb tense

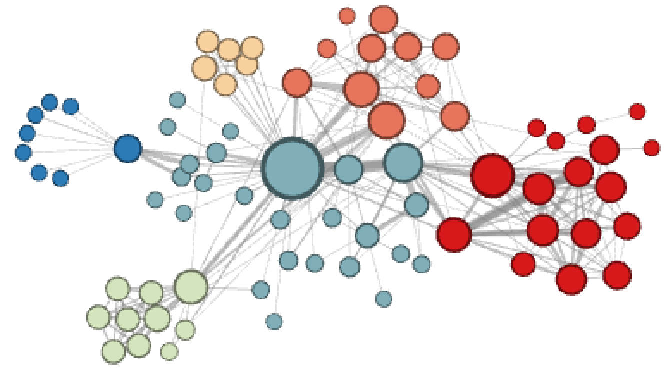


Country-Capital

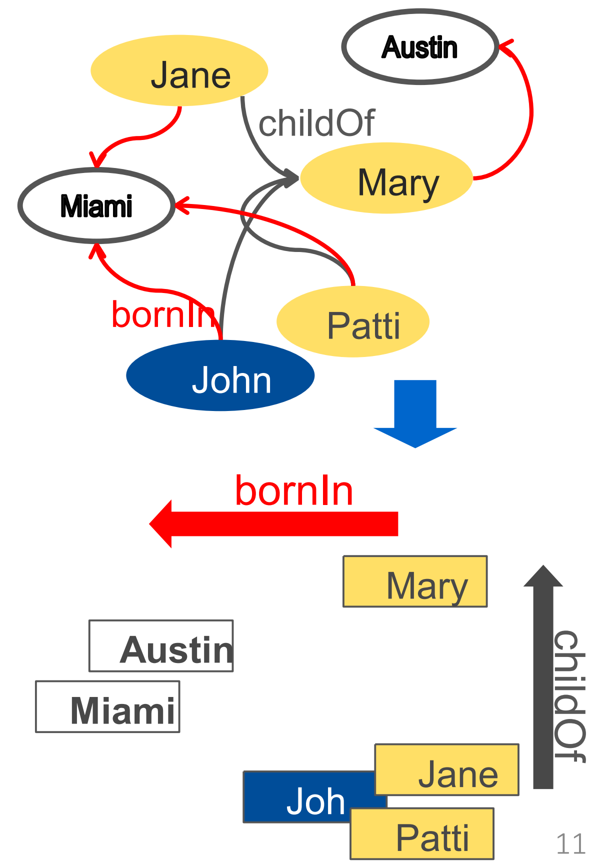
https://www.tensorflow.org/versions/r0.7/tutorials/word2vec/index.html

Explicit vs. Implicit "Graph" Representation

- Graph Embedding
  - ISOMap [Tenenbaum et al., Science'00]
  - LLE [Roweis and Saul, Science'00]
  - Laplacian EigenMap [Belkin et al., NIPS'01]
  - (t)-SNE [Maaten and Hinton, JMLR'08]
  - Deepwalk [Perozzi et al., KDD'14]
  - LINE [Tang et al., WWW'15]
  - Node2vec [Grover and Leskovec, KDD'16]



- Knowledge Graph Embedding
  - TransE [Bordes et al., NIPS'13]
  - TransH [Wang et al., AAAI'14]
  - TransR [Lin et al., AAAI'15]
  - PathEmbedding [Guu et al., and Lin et al., EMNLP'15]
  - ATranB…



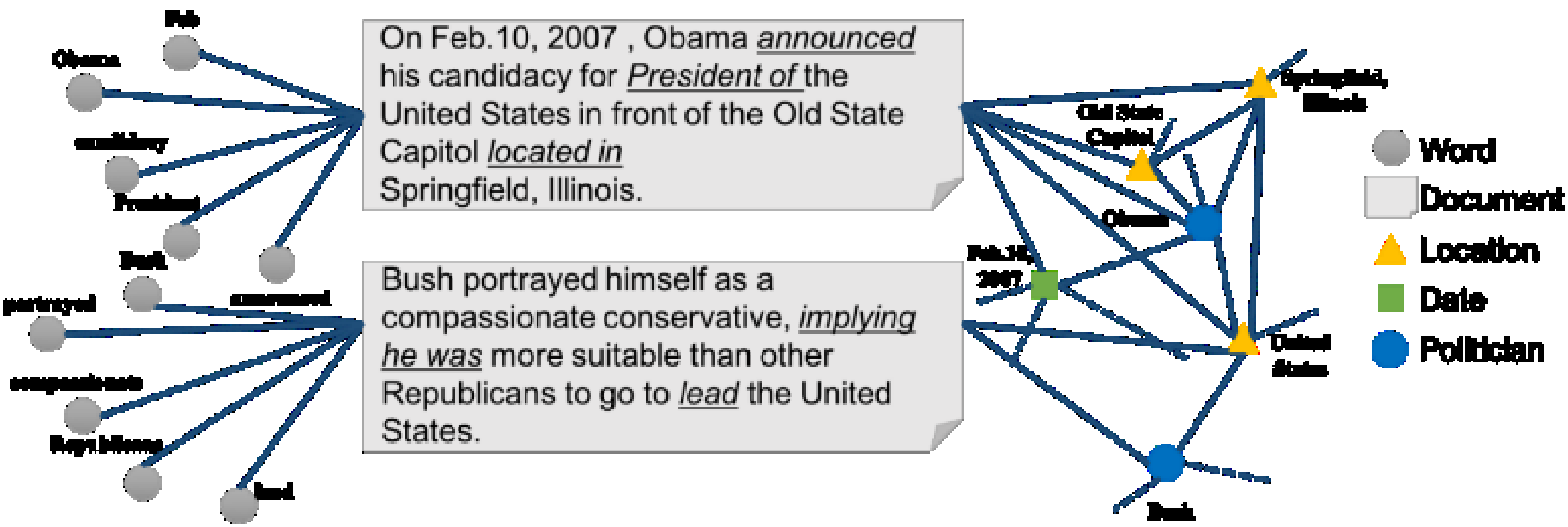
Explicit vs. Implicit Representation

Representation	Implicit	Explicit
Flat/Homogenous	LDA, word2vec	ESA
Graph/Heterogeneous	TransE	

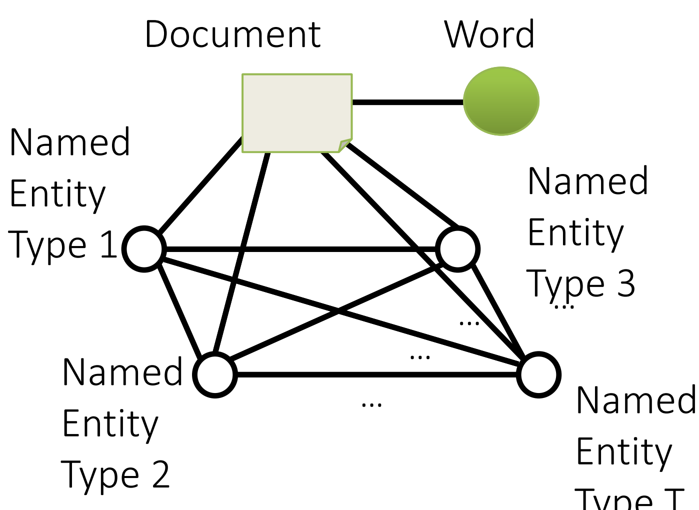
This talk

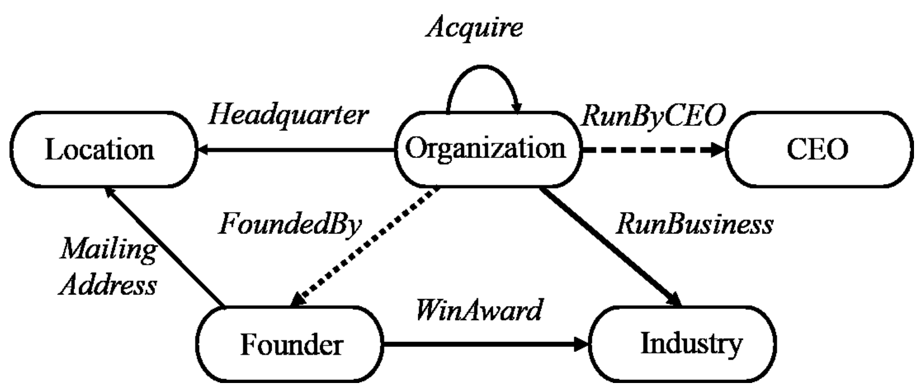
- From meta-path to meta-graphs
  - Semi-supervised learning [Jiang et al., IJCAI'17]
  - Recommendation [Zhao et al., KDD'17]
- Benefits
  - Have explicit semantics
    - Explainable
    - Knowledge discovery
  - Resolve different kinds of ambiguity

What Semantics Can HIN Provide?



HIN **network-schema**: network with multiple object types and/or multiple link types.

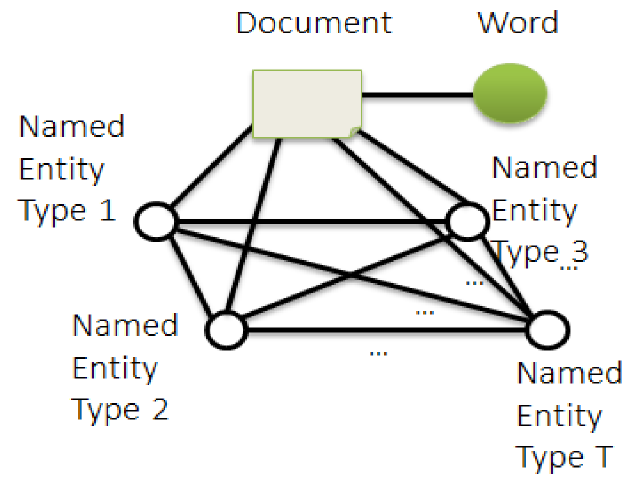




Meta-path, Commuting Matrix, and PathSim

- Meta-path defined over network schema.
  - [Sun et al., VLDB'11]
  - E.g.,

Document $\xrightarrow{\textit{Contains}}$ $\boxed{\text{word}}$ $\xleftarrow{\textit{Contains}}$ Document

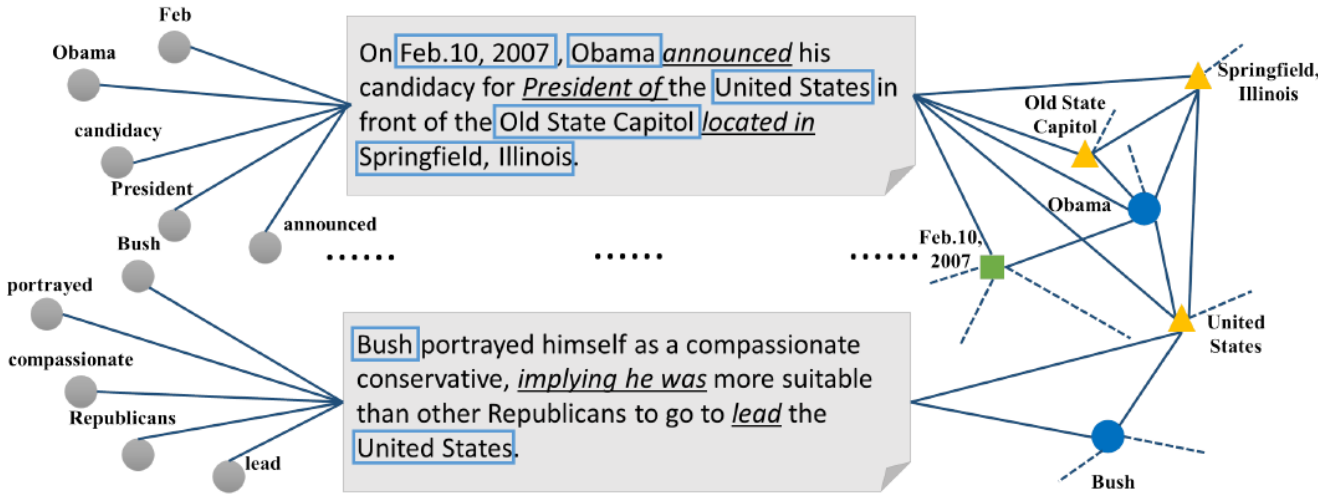


- Commuting matrix:
  - e.g., document->word binary occurrence matrix: W

- Un-normalized similarity: $W^T W$: dot product
  - Overall normalization: PathSim [Sun et al., VLDB'11]
  - Individual normalization: Path Ranking Algorithm [Lao et al., ML'10, EMNLP'11]

What Distinct Semantics Can HIN Provide?

- The semantics of entities and their relations



- What can context cover?
- What cannot?
  - Higher order relations

\`\`New York'' vs. \`\`New York Times''

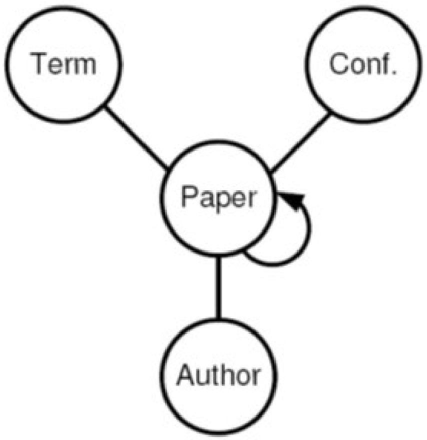
\`\`George Washington'' vs. \`\`Washington''

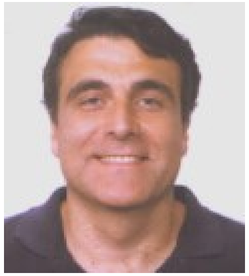
Document $\xrightarrow{\textit{Contains}}$ Basketball $\xrightarrow{\textit{Affiliation In}}$ NBA $\xleftarrow{\textit{Affiliation In}}$ Basketball $\xleftarrow{\textit{Contains}}$ Document

Document $\xrightarrow{\textit{Contains}}$ Basketball $\rightarrow$ Olympics $\leftarrow$ Basketball $\xleftarrow{\textit{Contains}}$ Document

Entity Search

- Who are most similar to Christos Faloutsos?
  - [Sun et al., 2011]





(a) Path: APA

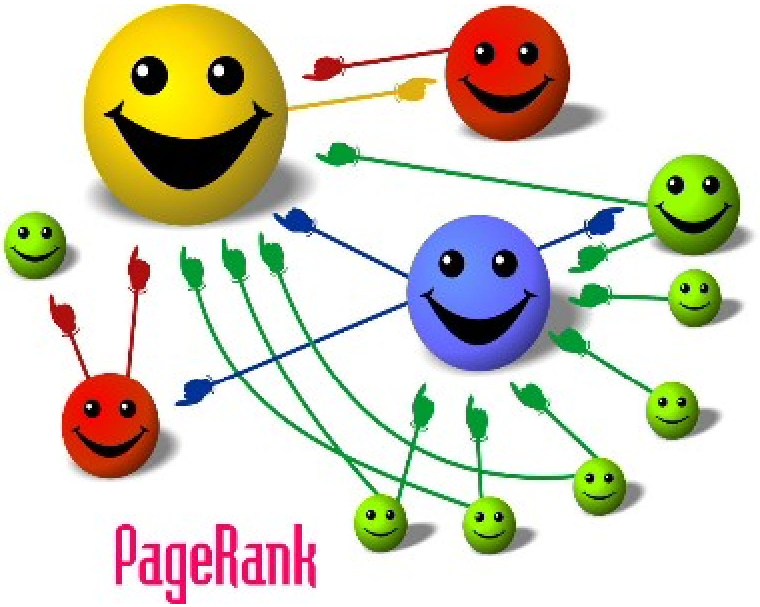
Rank	Author	Score
1	Christos Faloutsos	1
2	Spiros Papadimitriou	0.127
3	Jimeng Sun	0.12
4	Jia-Yu Pan	0.114
5	Agma J. M. Traina	0.110
6	Jure Leskovec	0.096
7	Caetano Traina Jr.	0.096
8	Hanghang Tong	0.091
9	Deepayan Chakrabarti	0.083
10	Flip Korn	0.053

(c) Path: $APTPA$

Rank	Author	Score
1	Christos Faloutsos	1
2	Jian Pei	0.661
3	Srinivasan Parthasarathy	0.600
4	Jeffrey Xu Yu	0.587
5	Ming-Syan Chen	0.579
6	Jiawei Han	0.576
7	Mohammed Javeed Zaki	0.571
8	Hans-Peter Kriegel	0.563
9	Yannis Manolopoulos	0.548
10	Rakesh Agrawal	0.545

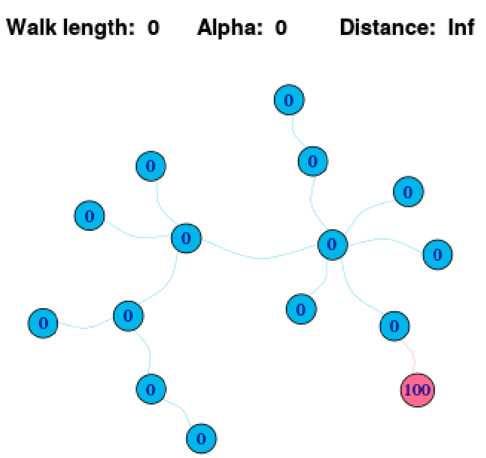
What's Still Missing/Unachievable?

- Let's consider a random walk on graph
  - Construct $n * n$ adjacency matrix $\mathbf{M}$
  - Normalize $\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2}$ ($\mathbf{D}$: degree matrix))
  - One step random walk: $\mathbf{p}^{t+1} = \mathbf{W}\mathbf{p}^{t}$
  - Stationary distribution follows: $\mathbf{p} = \mathbf{W}\mathbf{p}$

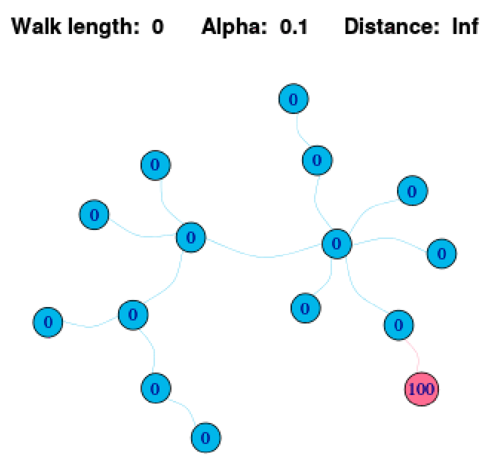


Personalized PageRank

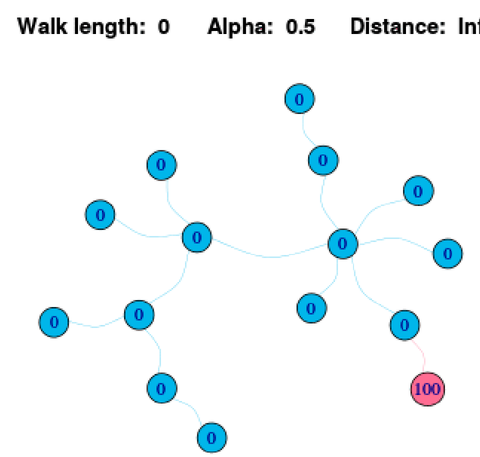
- PageRank [Page et al., '98]
  - $\mathbf{p}^{t+1} = (\alpha\mathbf{E} + (1-\alpha)\mathbf{W})\mathbf{p}^{t}$
  - With a probability to randomly/lazily jump
- Personalized PageRank/semi-supervised learning
  - [Haveliwala et al., TKDE'03, Jeh and Widom, WWW'03]
  - [Zhu et al., ICML'03, Zhou et al., NIPS'03]
  - $\mathbf{p}^{t+1} = \alpha\mathbf{q} + (1-\alpha)\mathbf{W}\mathbf{p}^{t}$
  - With a probability to restart with a label: prior



Lazy Random Walk



PPR (alpha = 0.1)

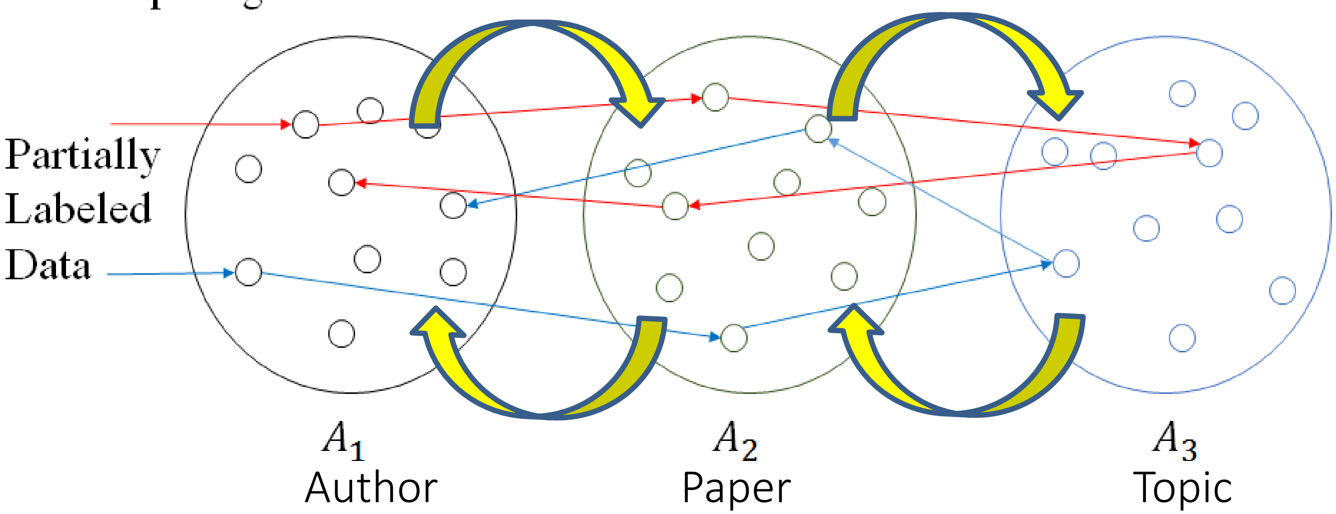


PPR(alpha = 0.5)

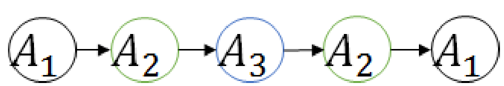
HIN: Path Constrained Random Walk

- In Path Ranking Algorithm
  - [Lao et al., ML'10, EMNLP'11]

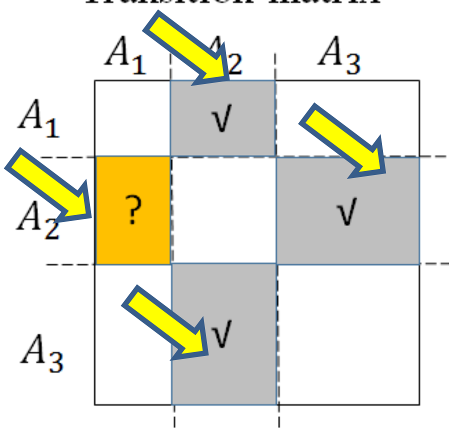
Meta-path guided Random Walk



Desired meta-path

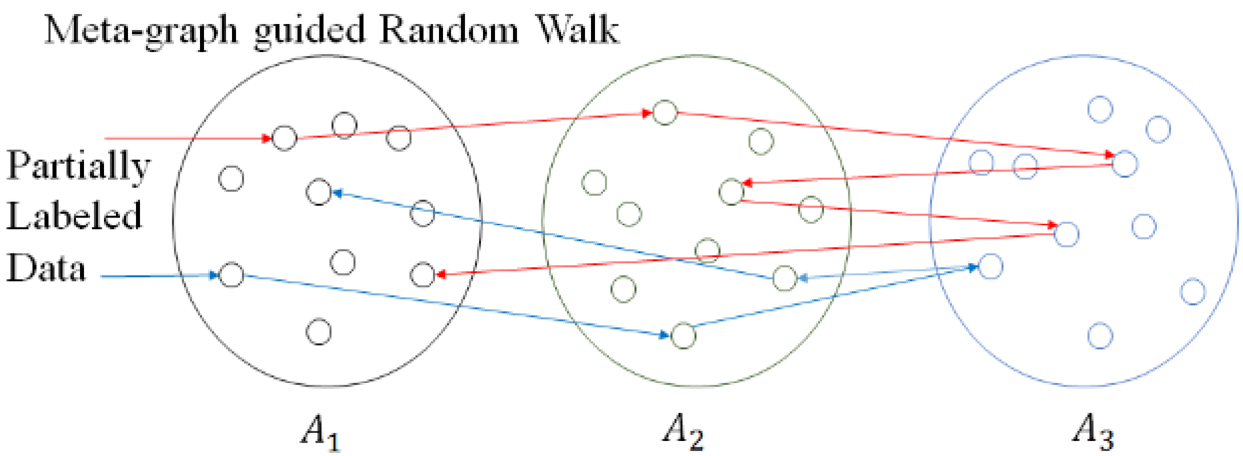


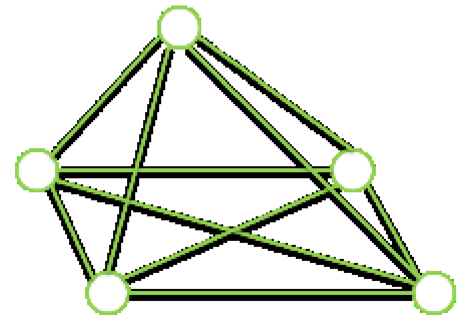
Transition matrix



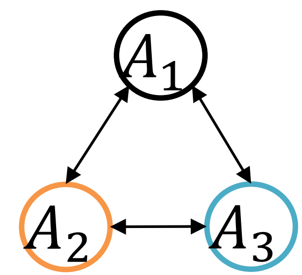
Meta-graph vs. Meta-path

- Meta-graph: [Fang et al., ICDE'16; Huang et al., KDD'16].
  - A sub-graph of network schema





Meta-graph

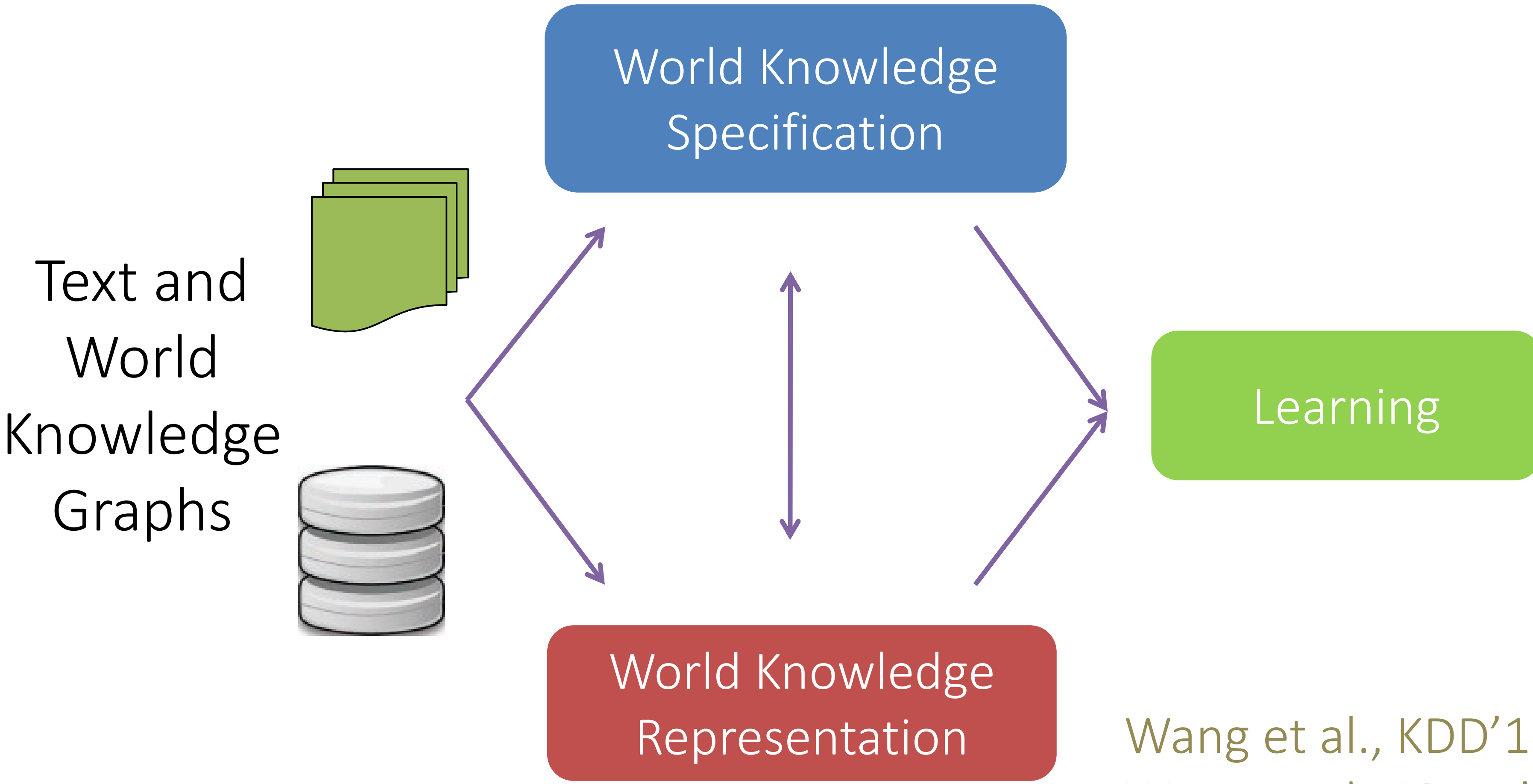


  - We get a stationary distribution!

Transition matrix

	A_1	A_2	A_3
A_1		√	√
A_2	√		√
A_3	√	√	

Application: Semi-supervised Text Classification



Wang et al., KDD’15
Wang et al., ICDM’15
Wang et al., TKDD’16
Wang et al., AAAI’16

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Trump is the president of the United States of America

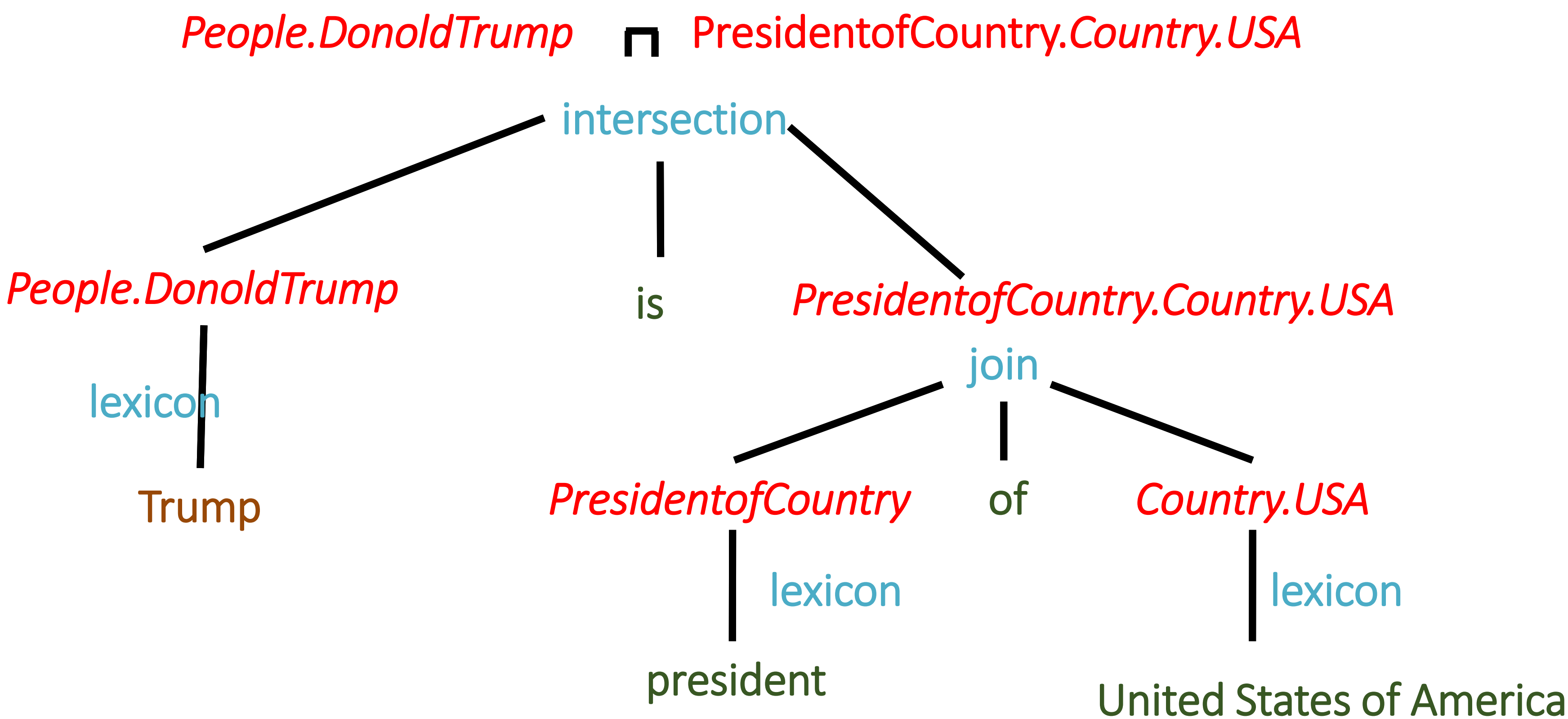
Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Logic form *People.DonoldTrump* $\sqcap$ PresidentofCountry.*Country.USA*

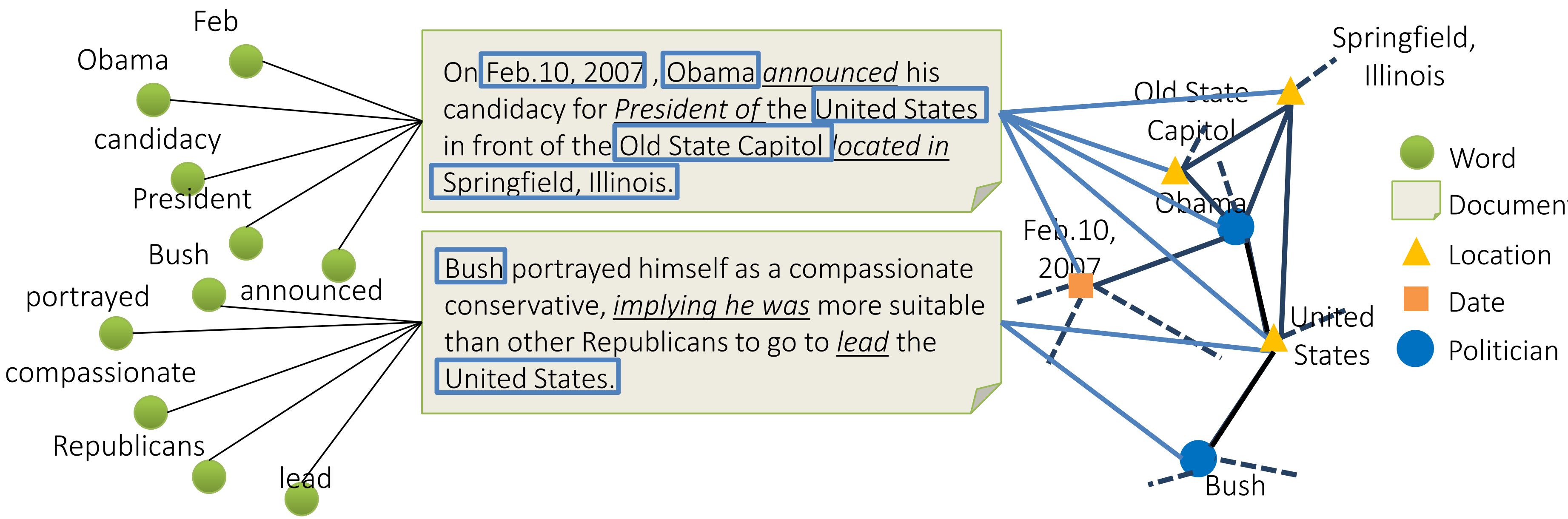
- Motivation: [Berant et al. EMNLP'13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase
  - Existing semantic parsing approaches, that require expert annotation
  - Scales to large scale knowledge-bases, supervised by the QA pairs
- We extend it to document analysis.

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Trump is the president of the United States of America



Example Meta-paths in Text HIN



Capturing higher-order relations

Document $\xrightarrow{\textit{Contains}}$ Politician $\xrightarrow{\textit{PresidentOf}}$ Country $\xleftarrow{\textit{PresidentOf}}$ Politician $\xleftarrow{\textit{Contains}}$ Document

Document $\xrightarrow{\textit{Contains}}$ Baseball $\xrightarrow{\textit{Affiliation In}}$ Sports $\xleftarrow{\textit{Affiliation In}}$ Baseball $\xleftarrow{\textit{Contains}}$ Document

Document $\xrightarrow{\textit{Contains}}$ Military $\xrightarrow{\textit{DepartmentOf}}$ Government $\xleftarrow{\textit{DepartmentOf}}$ Military $\xleftarrow{\textit{Contains}}$ Document

Algorithm

- Input:
  - Partially labeled documents
  - HIN based on semantic parsing

- Algorithm:
  - Step 1: extract transition matrices of different meta-graphs
  - Step 2: run personalized random walk based semi-supervised learning
  - Step 3: Ensemble of different meta-graph guided random walk

- Output:
  - Labels of all unlabeled data

Ensemble

- Supervised learning (SVM)
  - Input: meta-graph generated labels (soft labels)
  - Output: ground truth labels (partially labeled ones)
- EM [Dawid and Skene, 1979]
  - E-step: estimate posterior of label assignment of each meta-graph label
  - M-step: estimate label cluster probabilities, and likelihood of label assignment of each meta-graph label
- Co-training [Wan et al., SDM'15]
  - Train the weight of each meta-graph
  - Update the label assignment of each random walk

	Meta-graph 1		Meta-graph 2		…		Meta-graph G	
	Label 1	Label 2	Label 1	Label 2			Label 1	Label 2
Doc 1	0.9	0.1	0.1	0.8			0.9	0.2
Doc 2	0.9	0.2	0.8	0.1			0.6	0.5
…								
Doc N	0.2	0.7	0.1	0.6			0.3	0.6

Dataset

- 4 sub-datasets derived from 20-newsgroups and RCV1

20NewsGroup

RCV1-GCAT

Document datasets				
Sub-datasets	#(Document)	#(word)	#(Entity)	#(Types)
20NG-SIM	3,000	8,010	11,192	219
20NG-DIF	3,000	9,182	13,297	251
GCAG-SIM	3,596	11,096	10,540	233
GCAT-DIF	2,700	13,291	13,179	261
Each sub-datasets consists of three similar or distinct topics.				

Results

- BOW: bag-of-words
- Entity: entities extracted by semantic parsing
- NB: naïve Bayes
- SVM: support vector machines
- LP: label propagation
  - LP+Meta-graph: co-training [Wan et al., SDM'15]
  - KnowSim: unsupervised ensemble of meta-paths [Wang et al., ICDM'16]

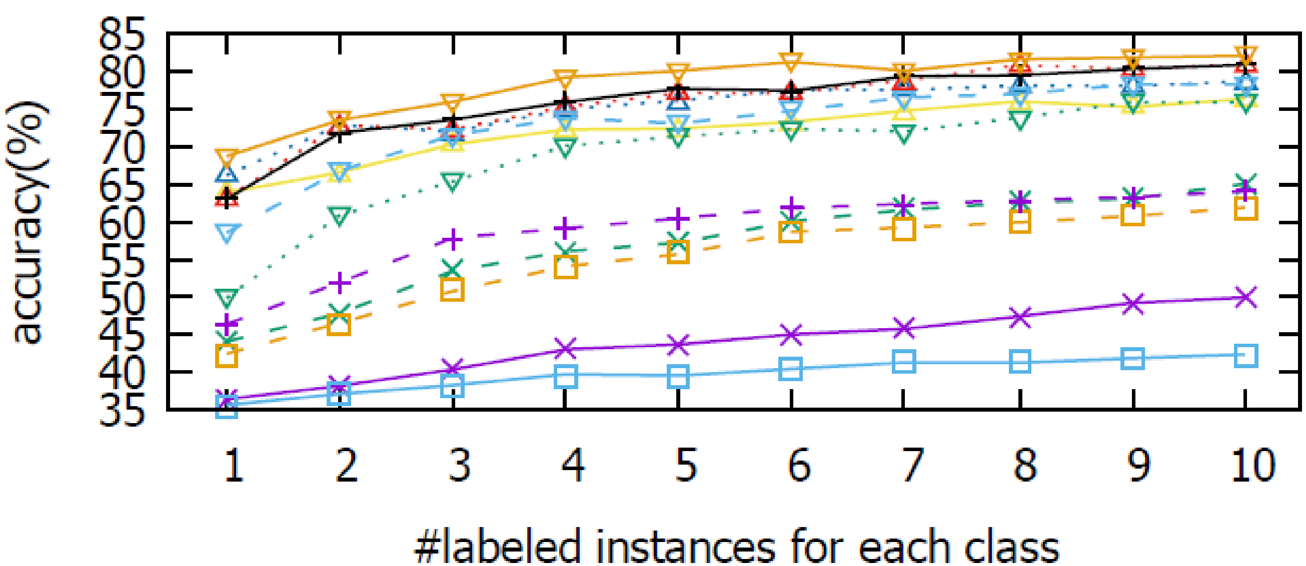
	NB		SVM			LP		SemiHIN		Ensemble		
Settings / Datasets	BOW	BOW+ Entity	BOW	BOW+ Entity	BOW+ Entity	Meta-path	Know-Sim	DWD Graph	Full-Graph	SVM	EM	Co-train
20NG-SIM	39.02	48.46	37.34	49.67	54.53	57.75	56.87	48.94	58.46	52.04	54.44	**60.99**
20NG-DIF	43.74	57.24	39.57	55.71	72.40	76.13	77.14	61.31	77.69	71.36	73.08	**80.08**
GCAT-SIM	71.24	71.24	73.92	74.64	70.97	71.05	60.59	79.14	**81.02**	68.79	69.96	80.97
GCAT-DIF	56.60	56.66	63.52	63.91	61.95	61.37	51.64	64.32	65.05	57.48	58.19	**66.95**

- We show our results of five labeled training data for each class. All the numbers are averaged accuracy (in percentage %) over 50 random trials.

Results

- BOW: bag-of-words
- Entity: entities extracted by semantic parsing
- NB: naïve Bayes
- SVM: support vector machines
- LP: label propagation
  - LP+Meta-graph: co-training [Wan et al., SDM'15]
  - KnowSim: unsupervised ensemble of meta-paths [Wang et al., ICDM'16]

20NG-DIF

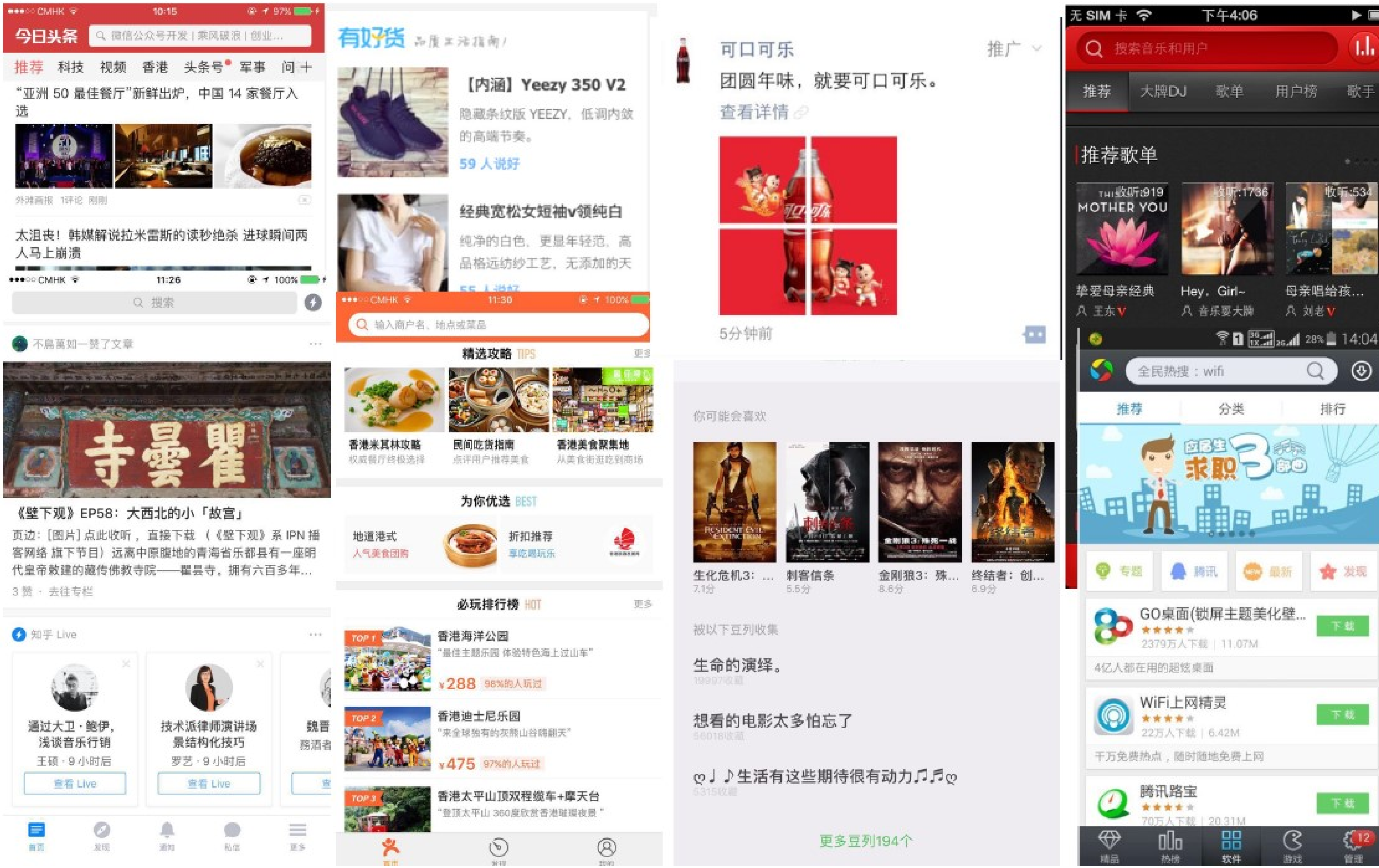


Explicit vs. Implicit “Graph” Representation

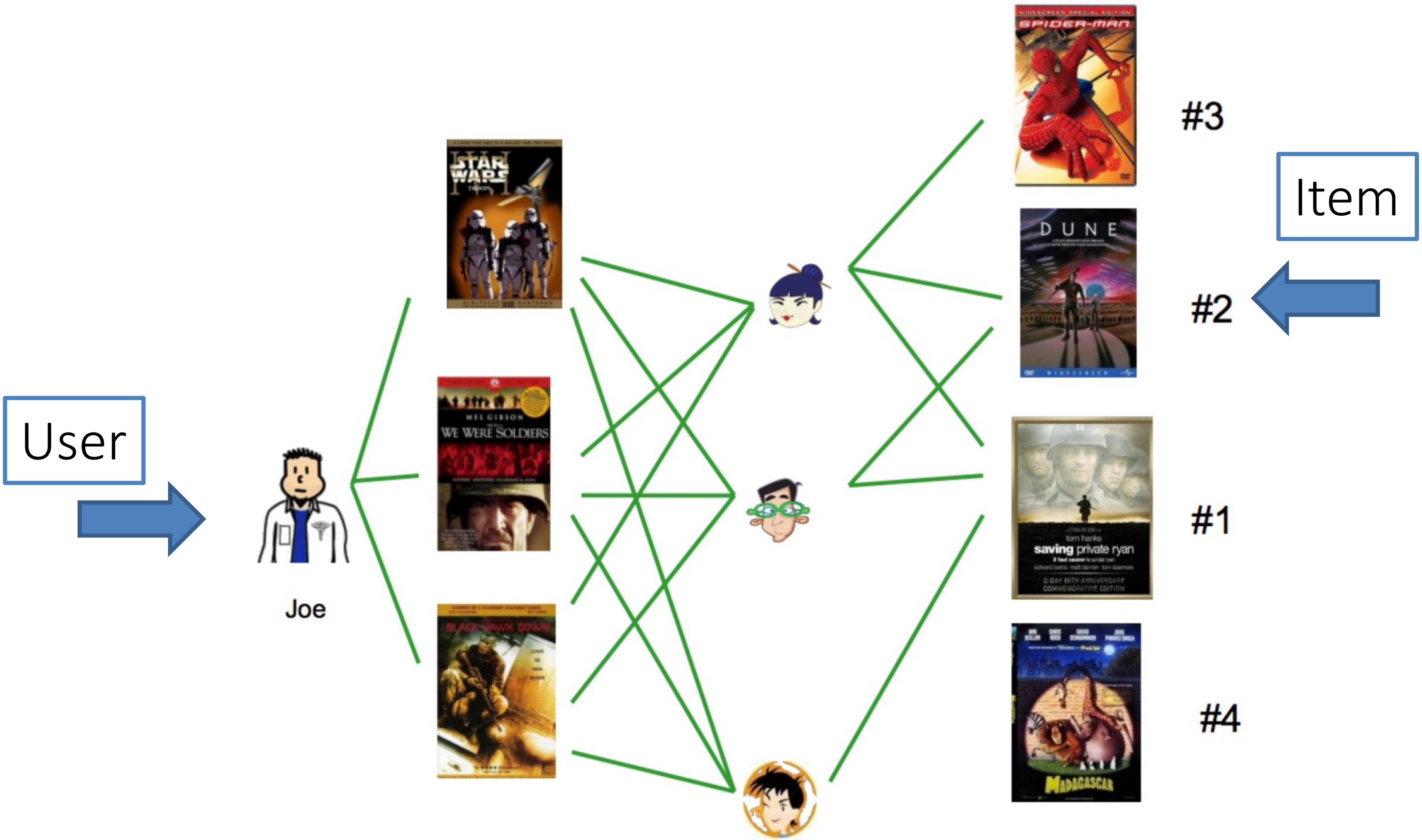
Representation	Implicit	Explicit
Flat/Homogenous	LDA, word2vec	ESA
Graph/Heterogeneous	TransE	This talk

- From meta-path to meta-graphs
  - Semi-supervised learning [Jiang et al., IJCAI’17]
  - Recommendation [Zhao et al., KDD’17]
- Benefits
  - Have explicit semantics
    - Explainable
    - Knowledge discovery
  - Resolve different kinds of ambiguous

RS is Everywhere Nowadays



Typical Recommendation Problem

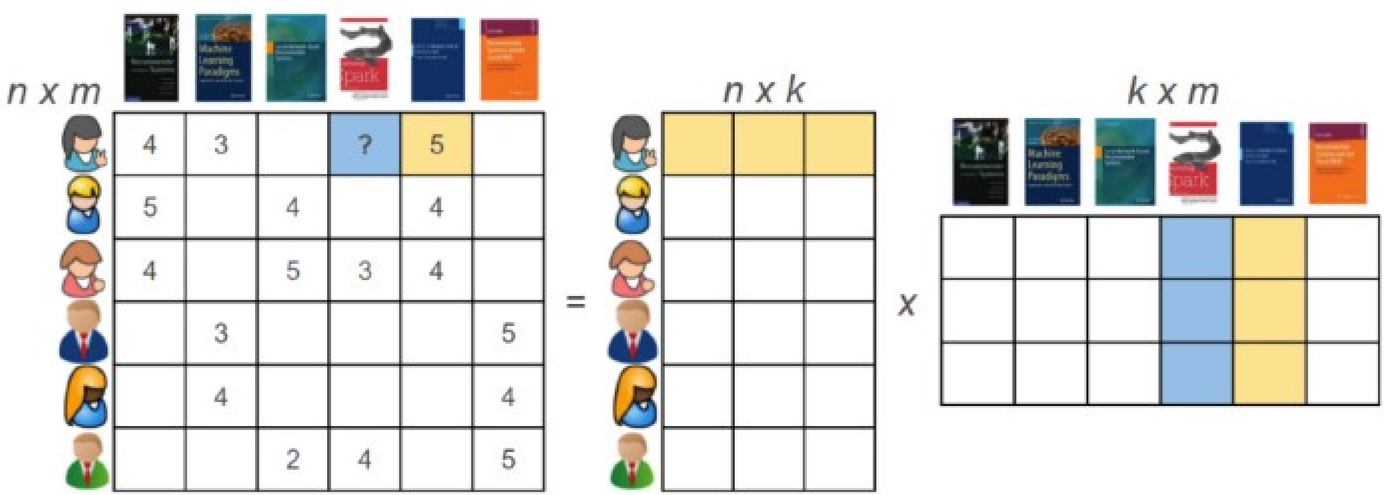


Matrix Factorization

- Matrix Factorization is one of the most popular methods for collaborative filtering
  - Given matrix $R \in \mathbb{R}^{n*m}$
  - each row represents an user *i*
  - While each column an item *j*

$$MAE = \frac{\sum_{(i,j)\in\mathcal{R}_{test}} |R_{ij} - \hat{R}_{ij}|}{|\mathcal{R}_{test}|},$$

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in\mathcal{R}_{test}} (R_{ij} - \hat{R}_{ij})^2}{|\mathcal{R}_{test}|}}.$$

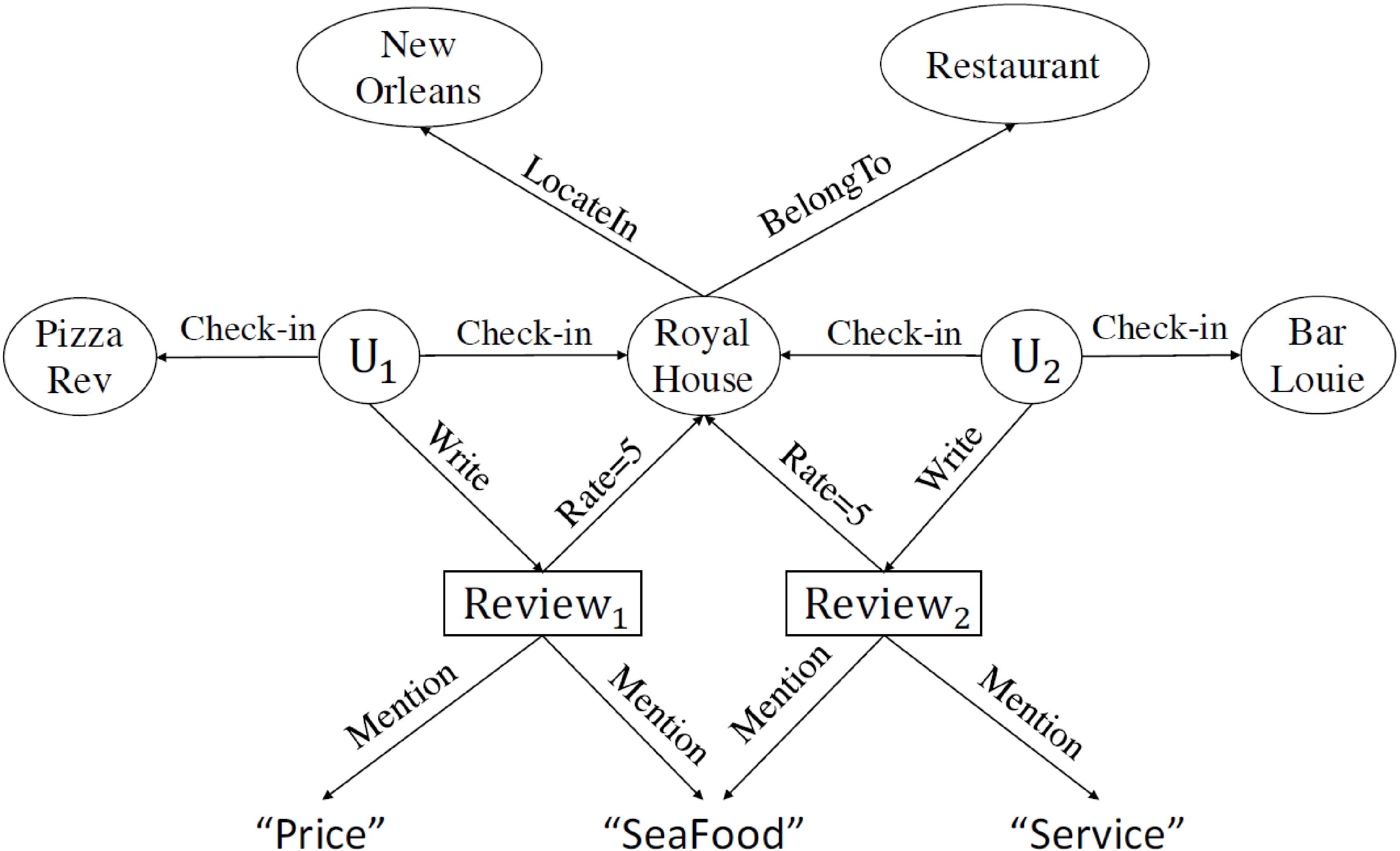


$$\min_{\mathbf{U},\mathbf{B}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{u}_i\mathbf{b}_j)^2 + \frac{\lambda_1}{2}||\mathbf{U}||_F^2 + \frac{\lambda_2}{2}||\mathbf{B}||_F^2$$

Other Existing Approaches

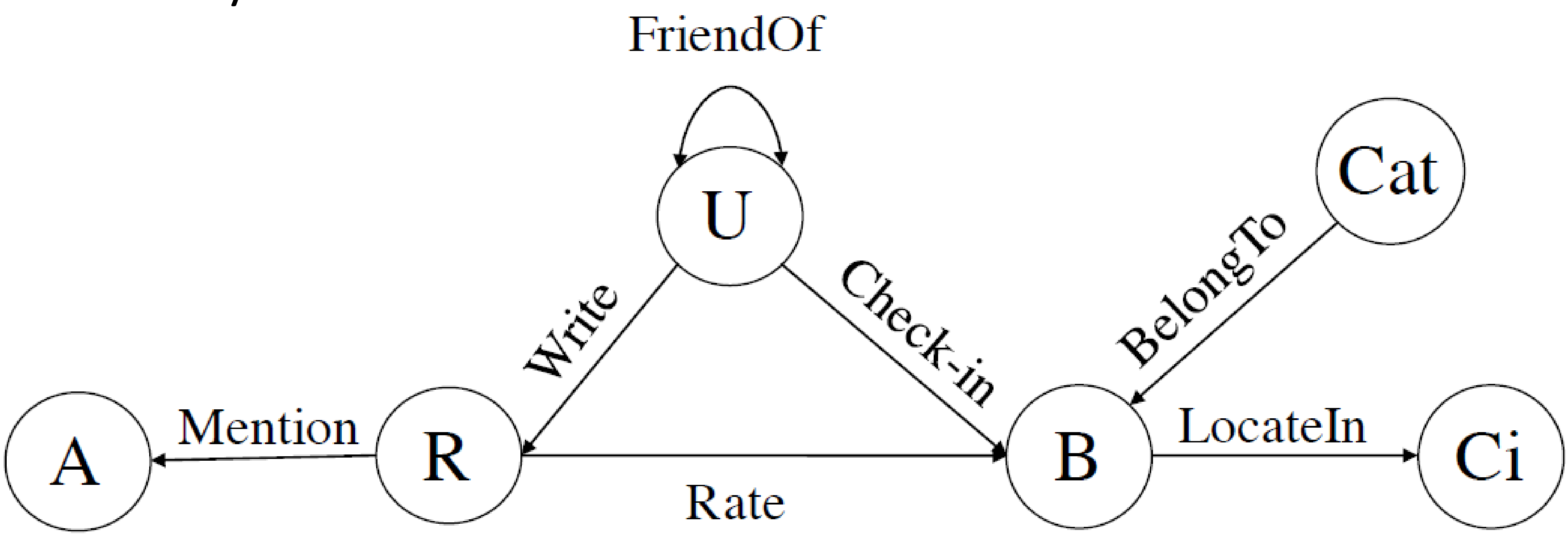
- Collaborative Filtering: Recommend items based only on the users past behavior
  - User based: find similar users for what they liked
  - Item based: find similar items which I have liked
- Content based: extract features for items
- Personalized learning to rank
- Demographic: user profiling
- Social recommendation: trust based
- Hybrid

It’s a Heterogeneous Information Network!

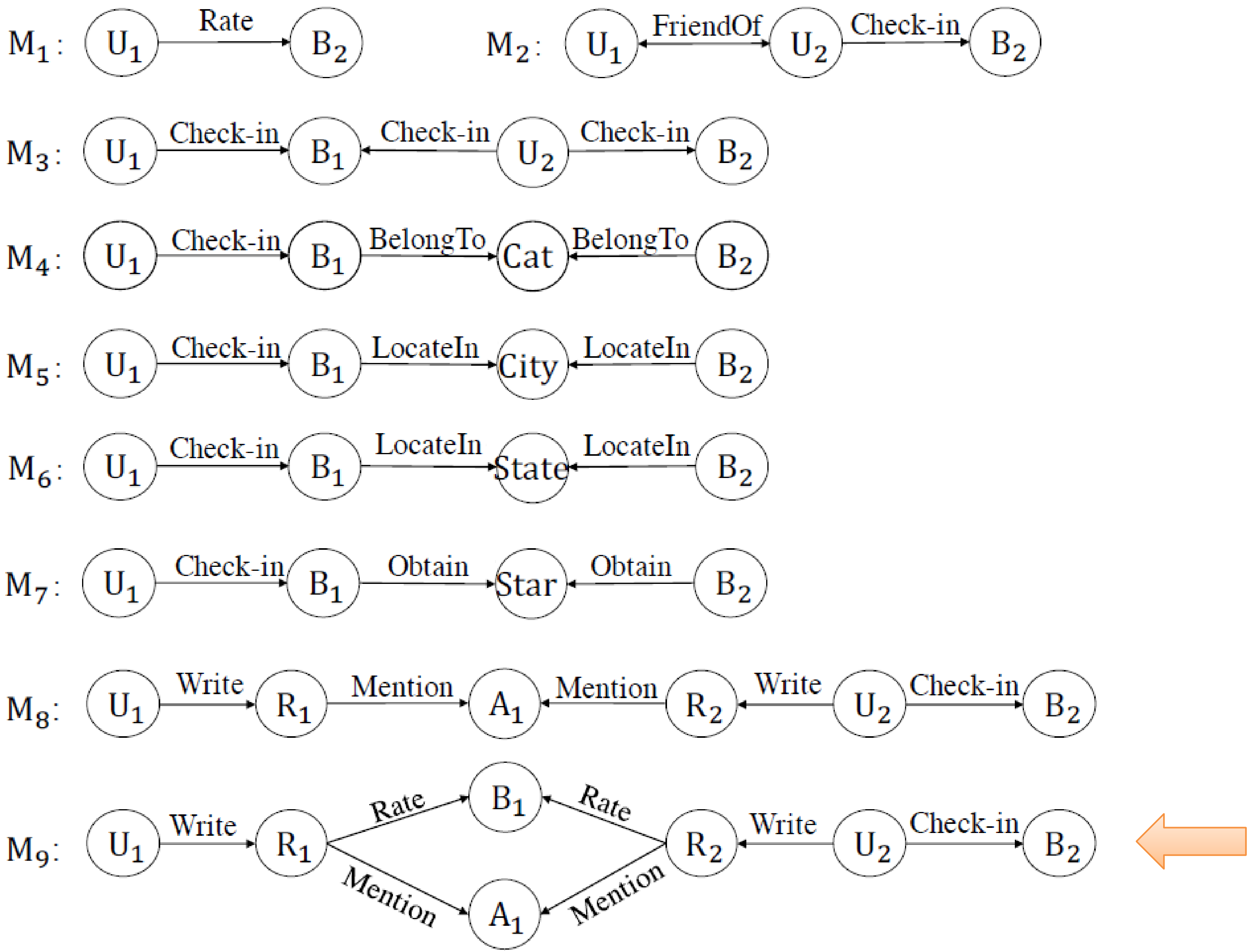


A Typical Network Schema of Yelp

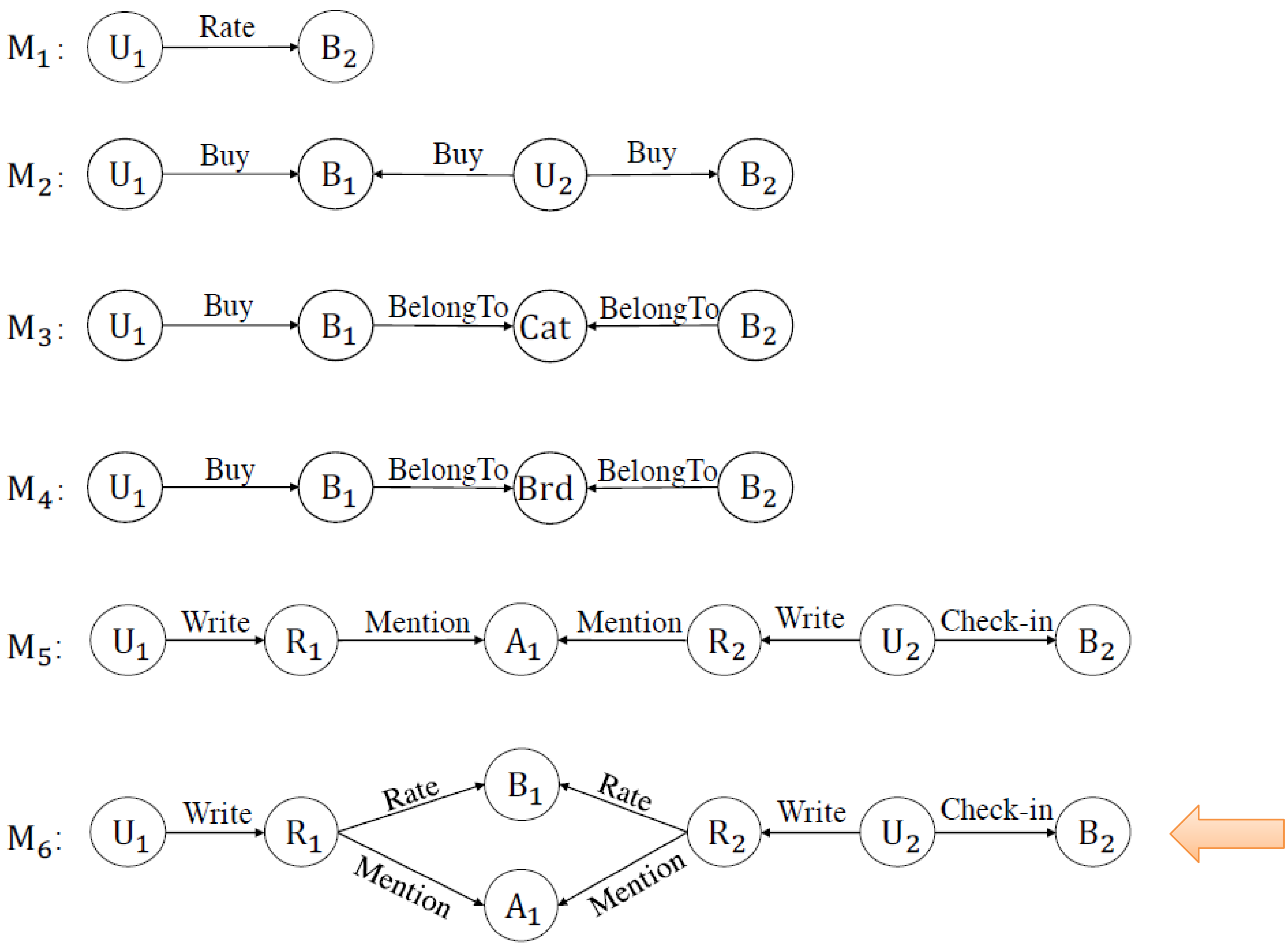
- R: reviews;
- U: users;
- B: business;
- Cat: category of item;
- Ci: city



Meta-graphs Extracted From Yelp

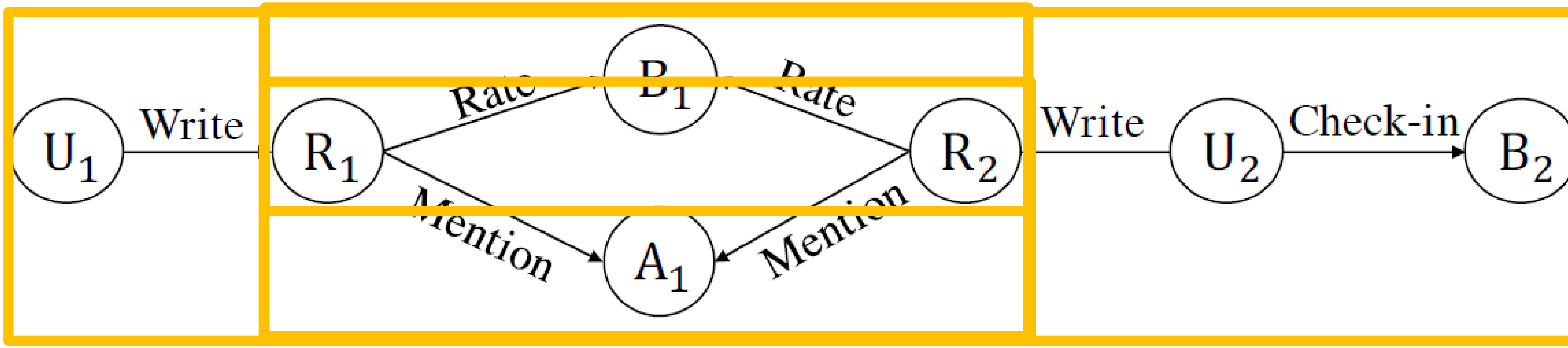


Meta-graphs Extracted From Amazon



Brd: brand of item

Compute a Similarity based on Meta-graph



Compute $\mathbf{C}_{P_1}$: $\mathbf{C}_{P_1} = \mathbf{W}_{RB} \cdot \mathbf{W}_{RB}^{\top};$

Compute $\mathbf{C}_{P_2}$: $\mathbf{C}_{P_2} = \mathbf{W}_{RA} \cdot \mathbf{W}_{RA}^{\top};$

Compute $\mathbf{C}_{S_r}$: $\mathbf{C}_{S_r} = \mathbf{C}_{P_1} \odot \mathbf{C}_{P_2};$

Compute $\mathbf{C}_{M_9}$ $\mathbf{C}_{M_9} = \mathbf{W}_{UR} \cdot \mathbf{C}_{S_r} \cdot \mathbf{W}_{UR}^{\top} \cdot \mathbf{W}_{UB};$

How to Assemble Different Meta-graphs?

- Factorization Machine [Rendle ICDM’10, TIST’12]
  - One of the state-of-art recommendation model recent years.

Feature vector **x** | Target y

	User				Movie					Other Movies rated					Time	Last Movie rated					Target y	
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...		
$\mathbf{x}^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y^{(1)}$
$\mathbf{x}^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y^{(2)}$
$\mathbf{x}^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y^{(2)}$
$\mathbf{x}^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y^{(3)}$
$\mathbf{x}^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y^{(4)}$
$\mathbf{x}^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y^{(5)}$
$\mathbf{x}^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y^{(6)}$

Matrix Factorization (MF)+Factorization Machine (FM)

- For each meta-graph, do MF:

$$\min_{\mathbf{U},\mathbf{B}} \frac{1}{2}||P_\Omega(\mathbf{U}\mathbf{B}^\top - \mathbf{R})||_2^2 + \frac{\lambda_u}{2}||\mathbf{U}||_2^2 + \frac{\lambda_b}{2}||\mathbf{B}||_2^2$$

- Given all MF latent features:
  - *L* meta-graphs
  - *F* dimension of MF

$$\mathbf{x}^n = \underbrace{\mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(l)}, \ldots, \mathbf{u}_i^{(L)}}_{L \times F} \underbrace{\mathbf{b}_j^{(1)}, \ldots, \mathbf{b}_j^{(l)}, \ldots, \mathbf{b}_j^{(L)}}_{L \times F}$$

- Do FM:

$$\hat{y}^n(\mathbf{w}, \mathbf{V}) = w_0 + \sum_{i=1}^{d} w_i x_i^n + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i^n x_j^n,$$

Automatic Meta-graph Selection

- The original cost function of FM

$$\min_{\mathbf{w},\mathbf{V}} \sum_{n=1}^{N} (y^n - \hat{y}^n(\mathbf{w},\mathbf{V}))^2$$

$$\hat{y}^n(\mathbf{w},\mathbf{V}) = w_0 + \sum_{i=1}^{d} w_i x_i^n + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i^n x_j^n$$

- + group lasso:

$$\Phi_{\mathbf{W}}(\mathbf{w}) = \sum_{l=1}^{2L} ||\mathbf{w}_l||_2 \qquad \Phi_{\mathbf{V}}(\mathbf{V}) = \sum_{l=1}^{2L} ||\mathbf{V}_l||_2$$

L meta-graphs
- In side meta-graph: L2 norm
- Between meta-graphs: L1 norm

nonmonotonous accelerated proximal gradient (nmAPG) algorithm [Li and Lin, NIPS'15]

Datasets

Yelp-200k				
Relations(A-B)	Number of A	Number of B	Number of (A-B)	Avg Degrees of A/B
User-Business	36,105	22,496	191,506	5.3/8.5
User-Review	36,105	191,506	191,506	5.3/1
User-User	17,065	17,065	140,344	8.2/8.2
Business-Category	22,496	869	67,940	3/78.2
Business-Star	22,496	9	22,496	1/2,499.6
Business-State	22,496	18	22496	1/1,249.8
Business-City	22,496	215	22,496	1/104.6
Review-Business	191,506	22,496	191,506	1/8.5
Review-Aspect	191,506	10	955,041	5/95,504.1

Amazon-200k				
Relations(A-B)	Number of A	Number of B	Number of (A-B)	Avg Degrees of A/B
User-Business	59,297	20,216	183,807	3.1/9.1
User-Review	59,297	183,807	183,807	3.1/1
Business-Category	20,216	682	87,587	4.3/128.4
Business-Brand	95,33	2,015	9,533	1/4.7
Review-Business	183,807	20,216	183,807	1/9.1
Review-Aspect	183,807	10	796,392	4.3/79,639.2

Comparison Results

Traditional Approaches

HIN Based Approaches

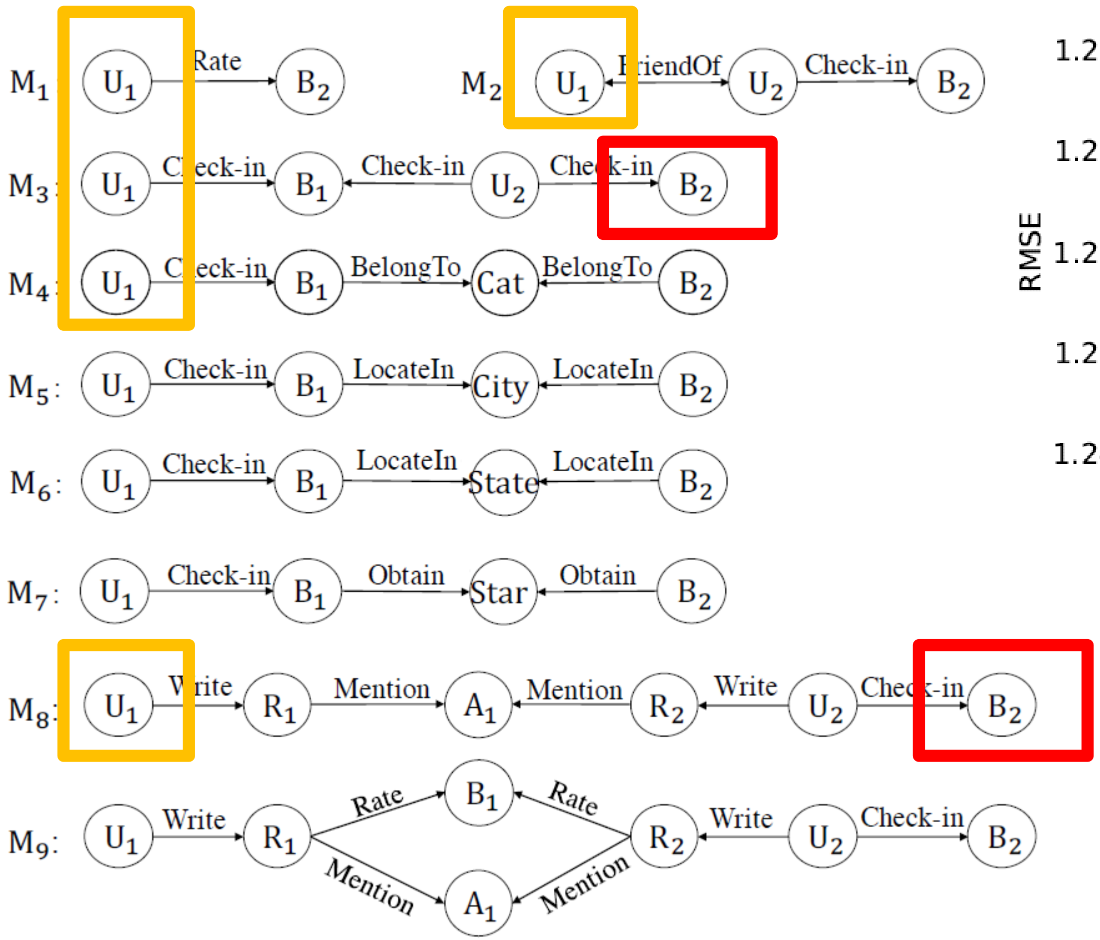
	Amazon-200k	Yelp-200k	CIKM-Yelp	CIKM-Douban
RegSVD	2.9656 (+60.0%)	2.5141 (+49.9%)	1.5323 (+27.7%)	0.7673 (+9.0%)
FMR	1.3462 (+11.9%)	1.7637 (+28.6%)	1.4342 (+22.8%)	0.7524 (+7.2%)
HeteRec	2.5368 (+53.2%)	2.3475 (+47.0%)	1.4891 (+25.6%)	0.7671 (+9.0%)
SemRec	- -	1.4603 (+13.8%)	1.1559 (+4.2%)	0.7216 (+3.2%)
FMG	**1.1864**	**1.2588**	**1.1074**	**0.6985**

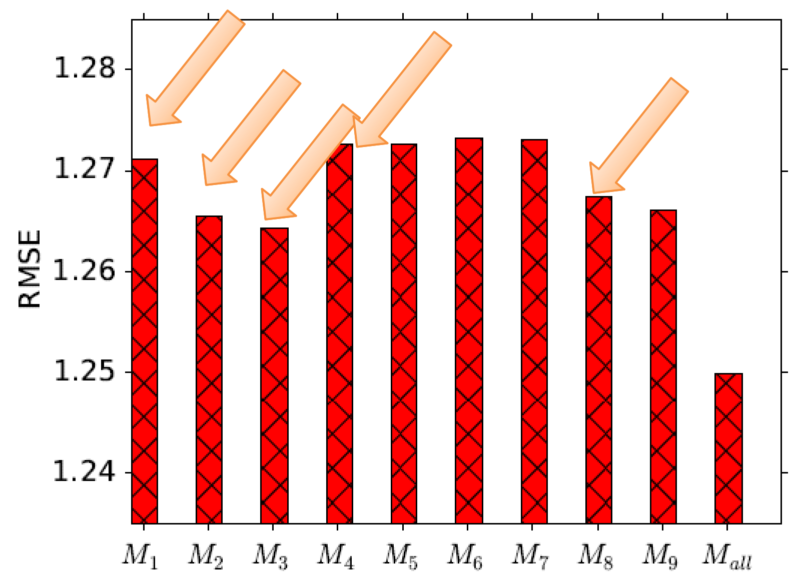
- HeteRec [Yu et al., WSDM'14]:
  - Factorize each meta-path
  - Ensemble using the recovered matrices
  - Item-based CF
- SemRec [Shi et al., CIKM'15]:
  - Ensemble of original similarity matrices based on different meta-paths
  - User based CF

	Amazon-200k	Yelp-200k	CIKM-Yelp	CIKM-Douban
Density	0.015%	0.024%	0.086%	0.630%

Selected Meta-graphs for Yelp

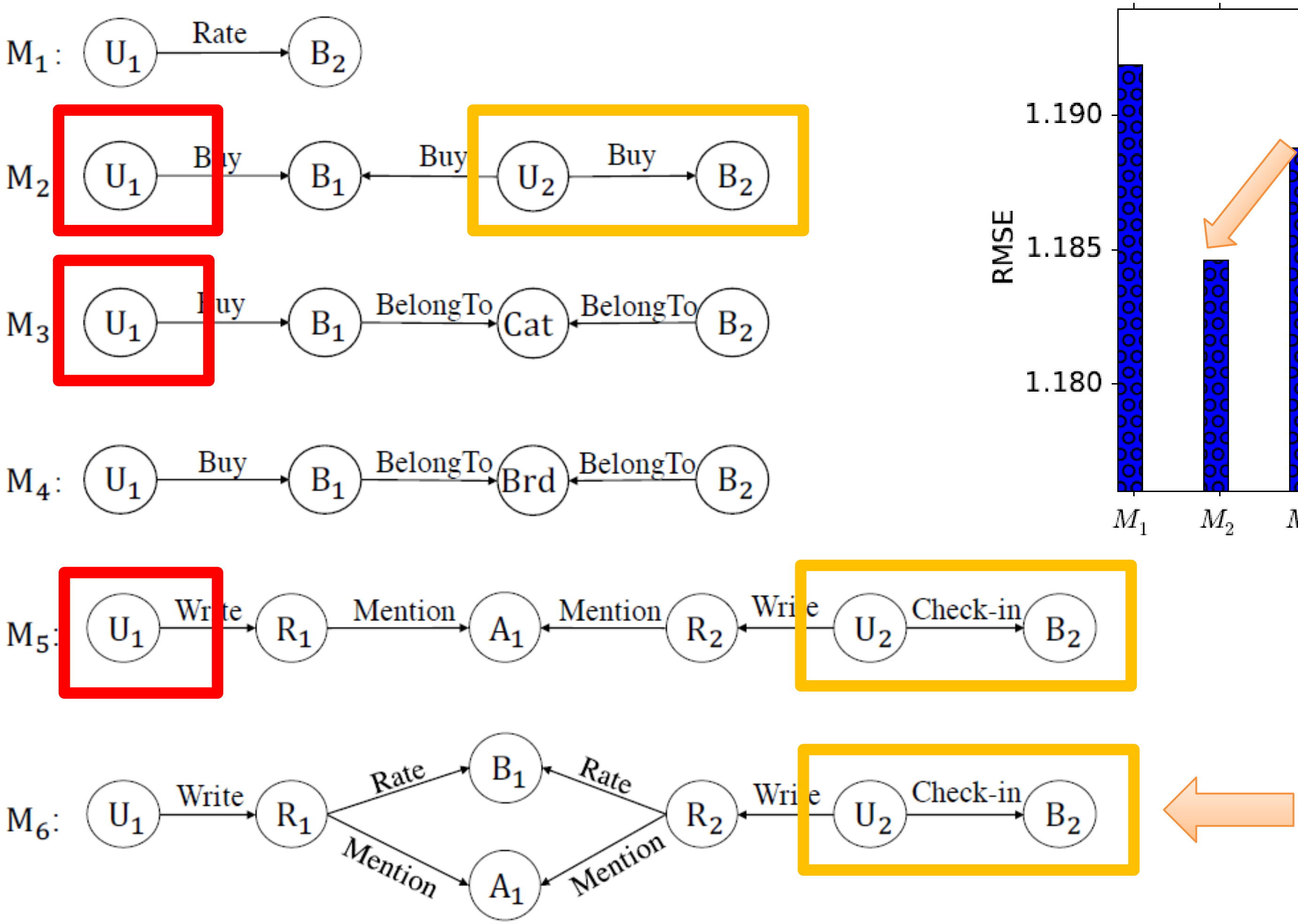
		User-Part		Item-Part	
		w	**V**	**w**	**V**
Yelp	Important	$M_1 - M_4, M_6, M_8$	$M_1 - M_3, M_5, M_8$	$M_1 - M_5, M_8, M_9$	M_3, M_8
	Useless	M_5, M_7, M_9	M_4, M_6, M_7, M_9	M_6, M_7	$M_1, M_2, M_4 - M_7, M_9$

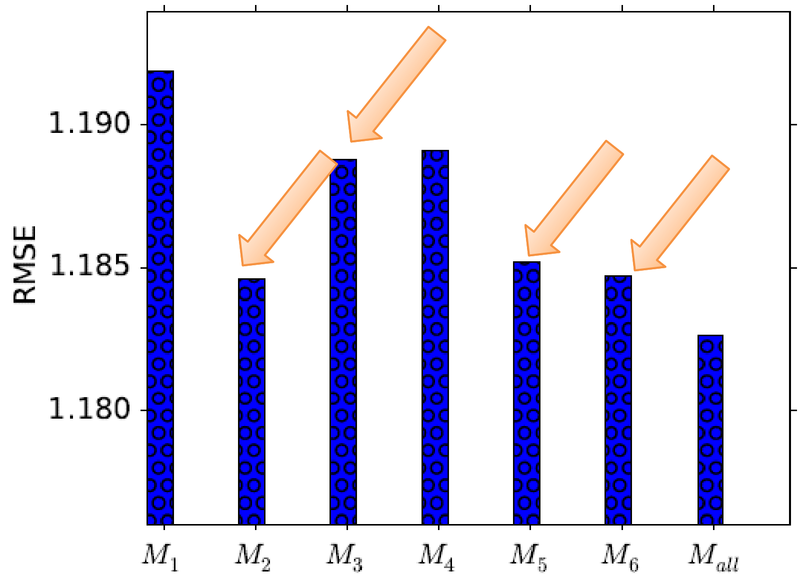




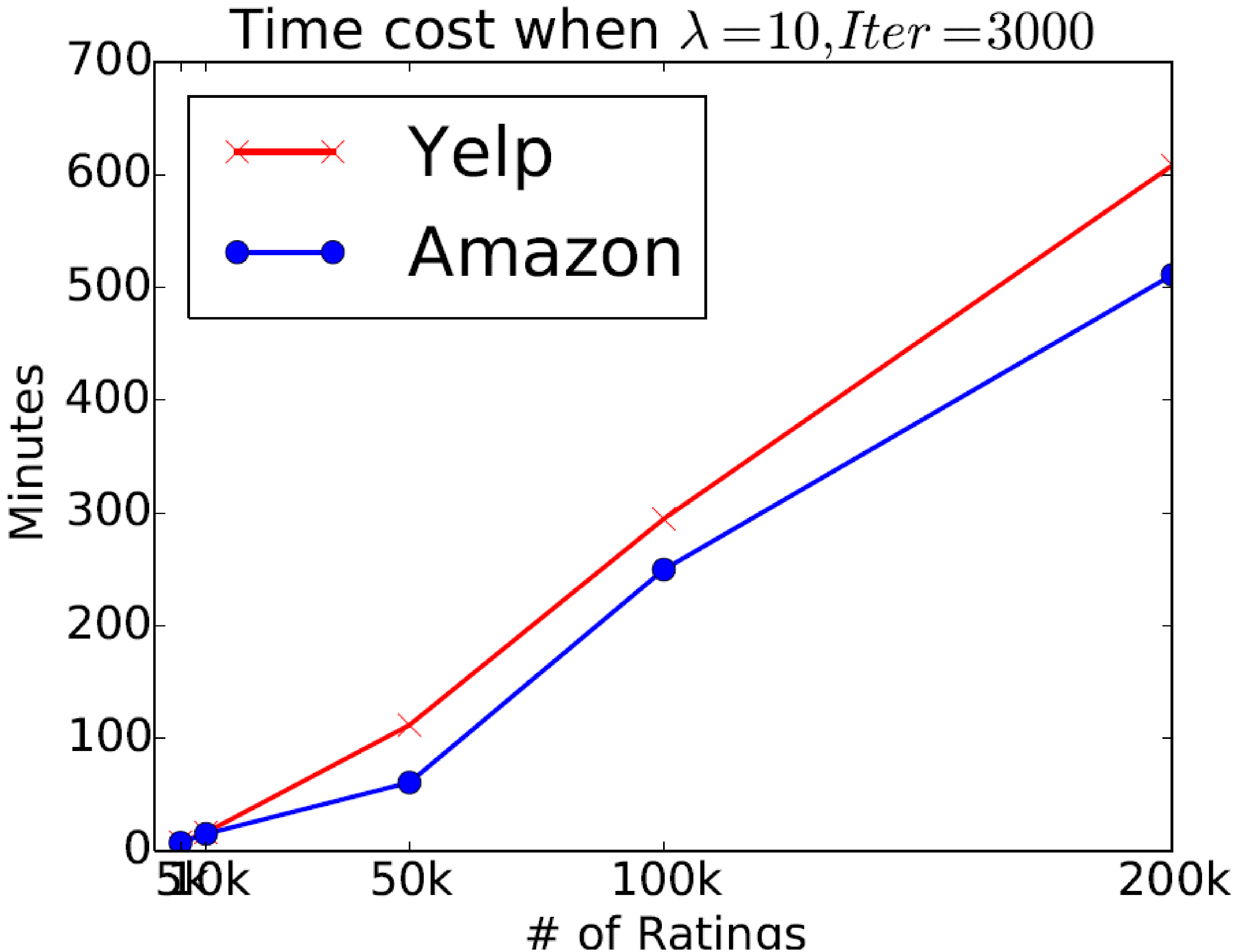
Selected Meta-graphs for Amazon

		User-Part		Item-Part	
		w	**V**	**w**	**V**
Amazon	Important	$M_1 - M_3, M_5$	$M_1 - M_6$	M_2, M_3, M_5, M_6	M_2, M_5, M_6
	Useless	M_4, M_6	-	M_1, M_4	M_1, M_3, M_4





Scalability of Algorithm



Collaborators

- He Jiang (HKUST)
- Huan Zhao (HKUST)
- Dik Lee (HKUST)
- Chenguang Wang (IBM)
- Ming Zhang (PKU)
- Yizhou Sun (UCLA)
- Jiawei Han (UIUC)
- Dan Roth (Upenn)

Slides Credit: Chenguang Wang, He Jiang, Huan Zhao

Conclusion

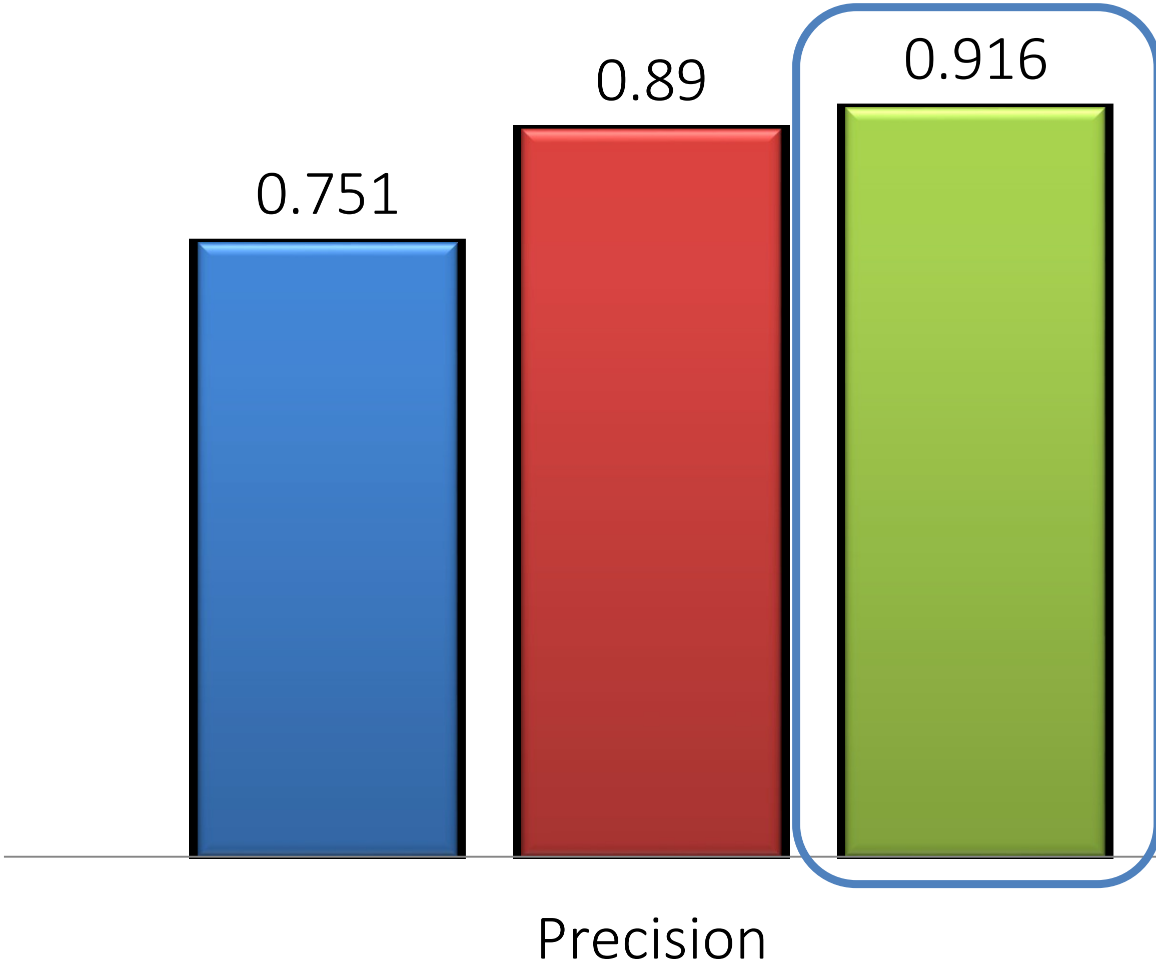
Heterogeneous information networks as explicit semantic analysis

From meta-path to meta-graph analysis

Code released at https://github.com/HKUST-KnowComp/FMG

Thank You! ☺

Precision of Different Semantic Filtering



FBSF

Frequency based semantic filter. Type is decided by the counts in one document.

DFBSF

Document frequency based semantic filter. Type is decided by the counts in whole document set.

CBSF

Conceptualization based semantic filter. Type is decided by the context in whole document set.

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15.
Wang et al., World knowledge as indirect supervision for document clustering. TKDD'16.

Error Analysis of Semantic Filtering

Type of error	Example sentence	Number and percentage of errors		
		FBSF (805)	DFBSF (359)	CBSF (272)
Entity Recognition	"Einstein 's theory of relativity explained mercury 's motion."	179 (22.2%)	129 (35.9%)	105 (38.6%)
Entity Disambiguation	"Bill said all this to make the point that Christianity is eminently."	537 (66.7%)	182 (50.7%)	130 (47.8%)
Subordinate Clause	"Bruce S. Winters, worked at United States Technologies Research Center, bought a Ford."	89 (11.1%)	48 (13.4%)	37 (13.6%)

Finding #1: Entity disambiguation is the major error factor.
Entity disambiguation is a tough research problem in NLP community. The type information of relations are not sufficient to further prune out mismatching entities during semantic filtering process.
Finding #2: CBSF performs the best.
For example, by using context, the number of incorrect entities caused by disambiguation can be dramatically reduced.

Classification Results

Average accuracy			
Model	Discrete		Embedding
Settings	BOW	BOW+ENTITY	Word2vec
20NG-SIM	90.81%	91.11%	91.67%
20NG-DIF	96.66%	96.90%	98.27%
GCAG-SIM	94.15%	94.29	96.81%
GCAT-DIF	88.98%	90.18%	90.64%

Mikolov 2013.
Window: 5
Dim: 400

Average accuracy					
Model	SVM^{HIN}	SVM^{HIN}+KnowSim		IndefSVM^{HIN}+KnwoSim	
Settings		DWD	DWD+other MetaPaths	DWD	DWD+other MetaPaths
20NG-SIM	91.60%	92.32%	92.68%	92.65%	93.38%
20NG-DIF	97.20%	97.83%	98.01%	98.13%	98.45%
GCAG-SIM	94.82%	95.29%	96.04%	95.63%	98.10%
GCAT-DIF	91.19%	90.70%	91.88%	91.63%	93.51%

Collective classification: Lu and Gatoor 2003; Kong et al. 2012