Automatic Information Fusion with Heterogeneous Information Networks

Yangqiu Song

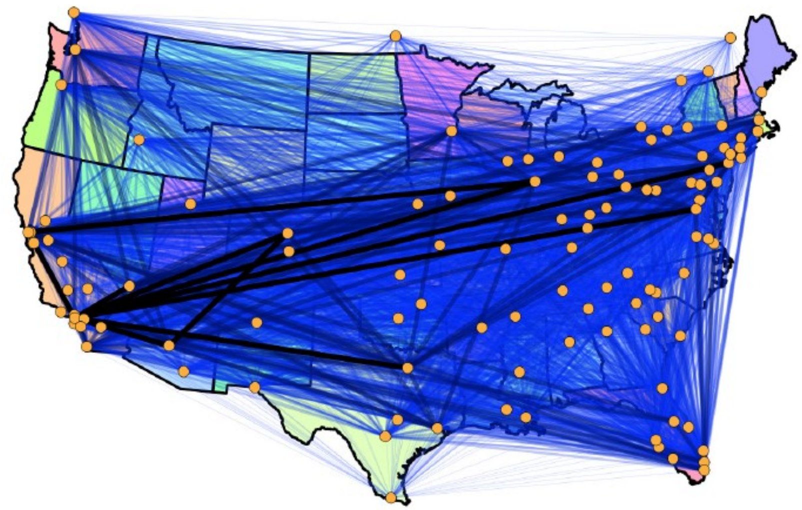
Department of CSE, HKUST, Hong Kong

香 港 科 技 大 學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

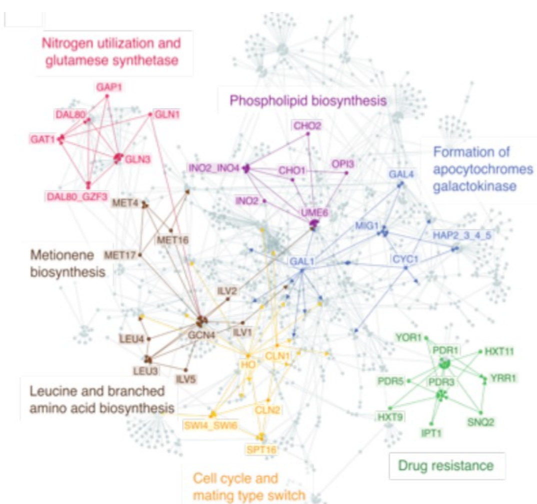
Joint work with many collaborators; Slides Credit: Chenguang Wang, Huan Zhao, Yanfang (Fanny) Ye

Homogeneous Graph/Networks

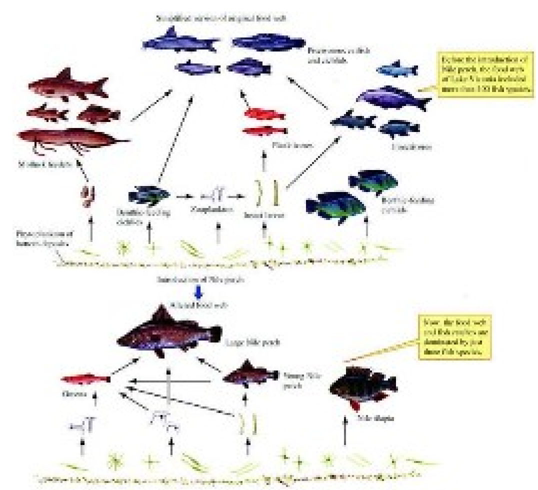
Social Network



Transportation Network



Gene Network

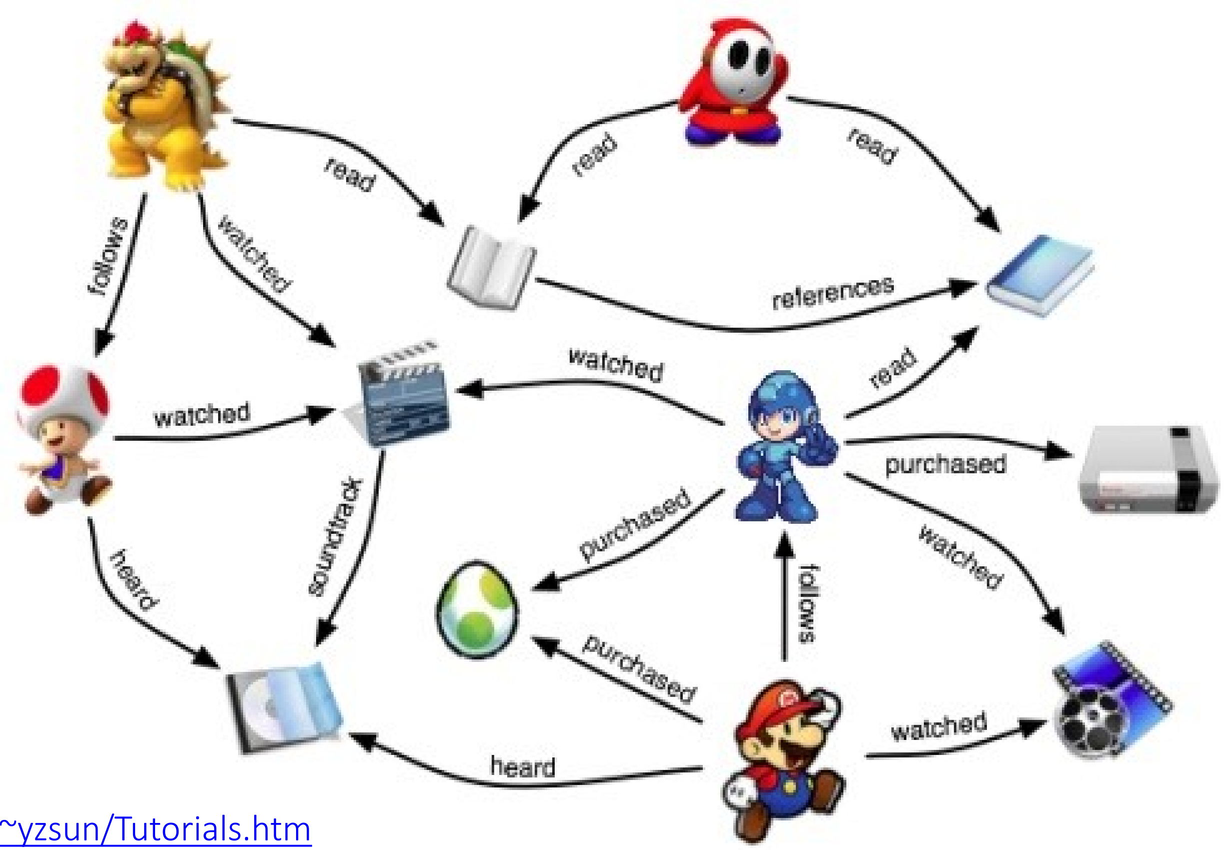


Food Network

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

Heterogeneous Information Networks

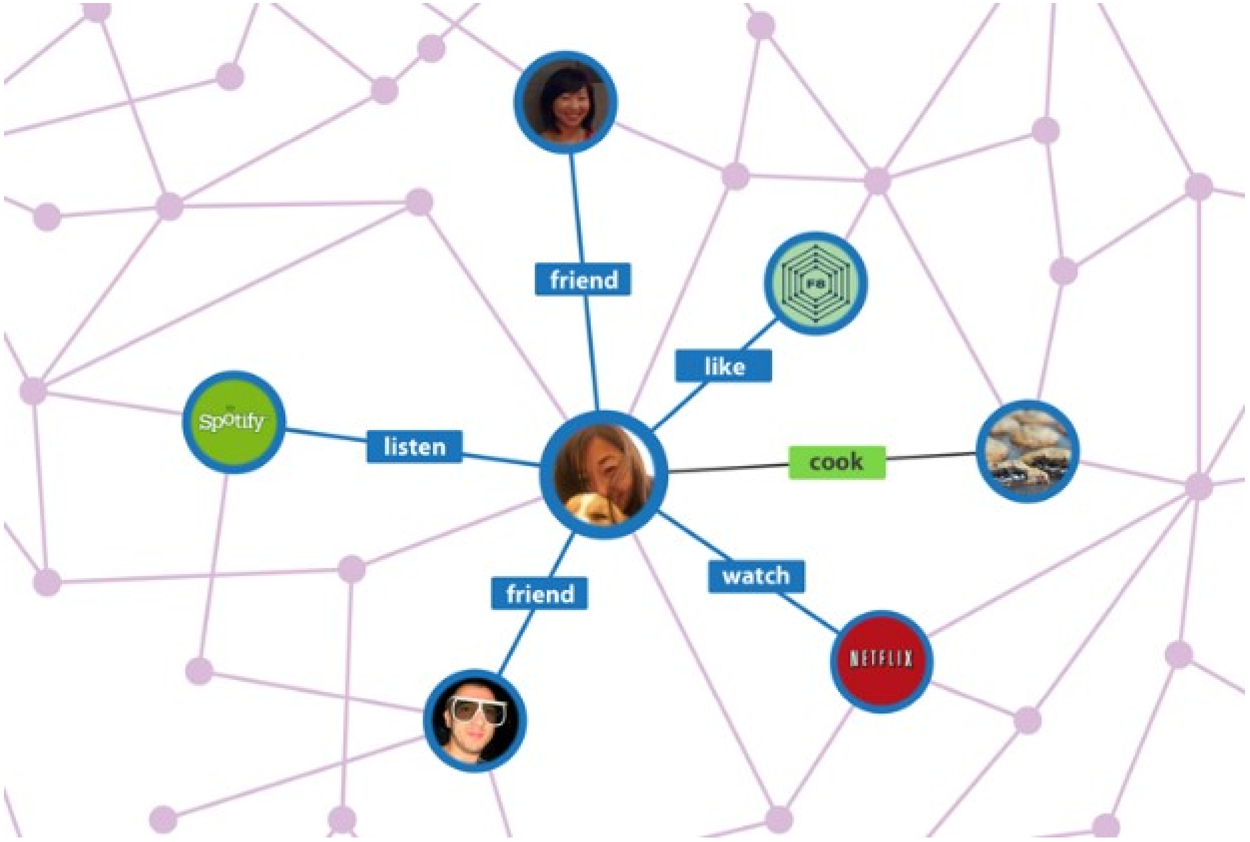
- Yizhou Sun, Jiawei Har)09-2012 (UIUC)
  - Entity type mappinc
  - Link type mapping: E -> R



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

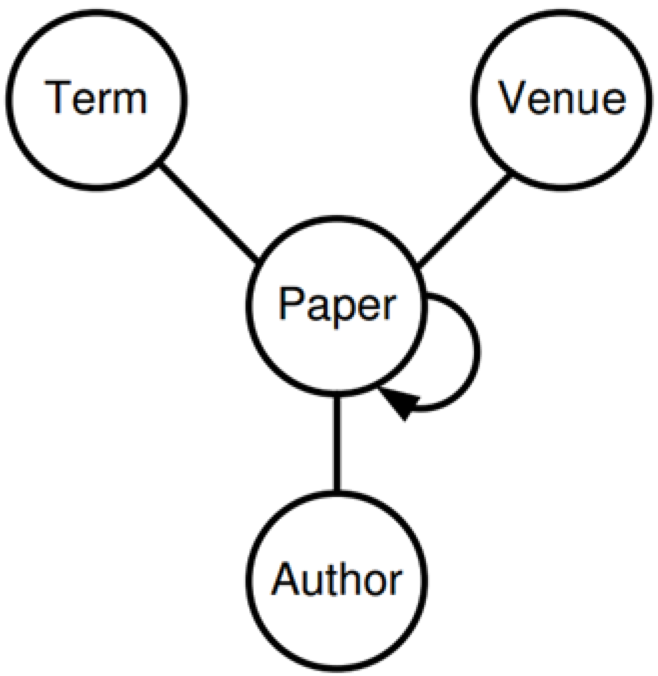
Modern Social Media

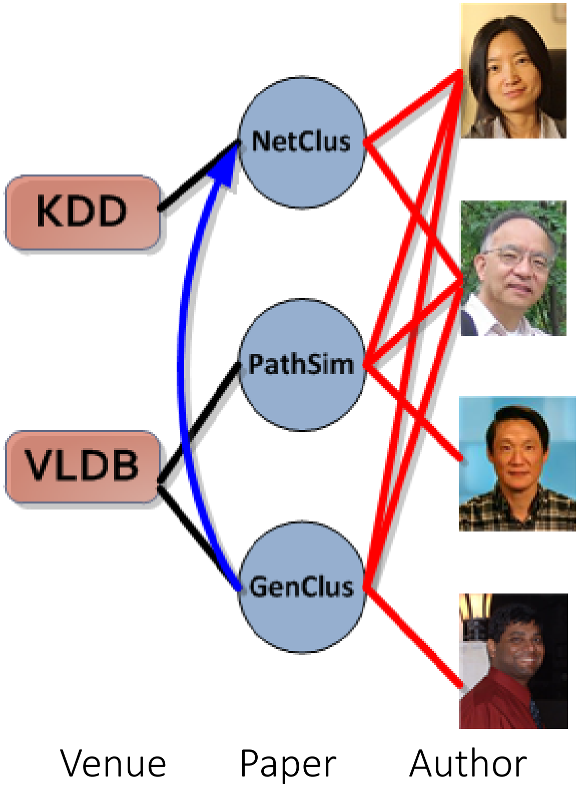
- Entities: Person, Check-in location, Articles, etc.
- Relations: Friends, Like, Check-in, etc.



Scholar Networks

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



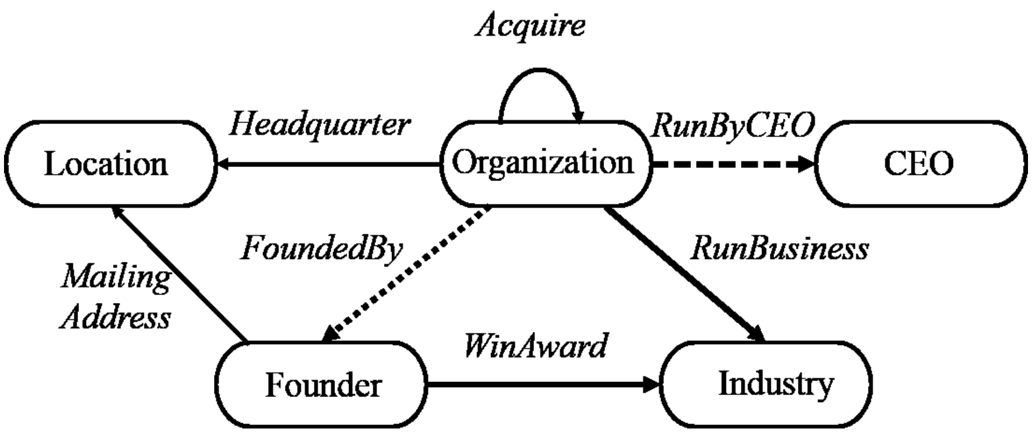


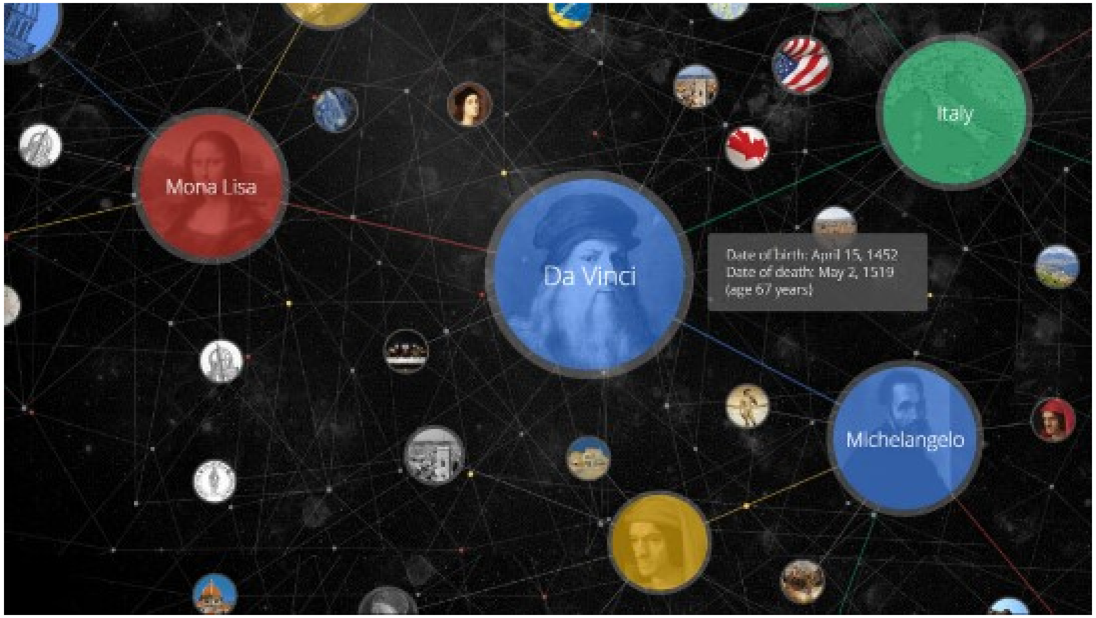
DBLP Bibliographic Network

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

Knowledge Graphs

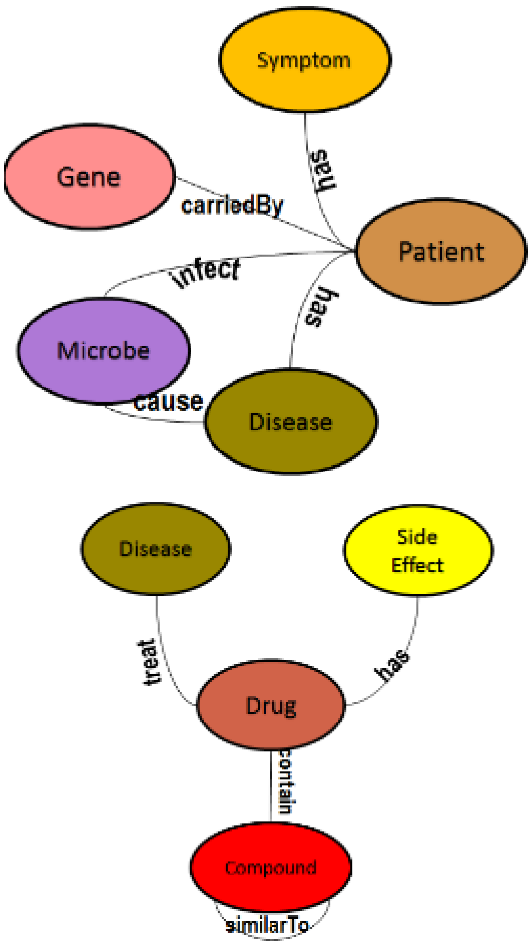
- Example of entities and their relations:

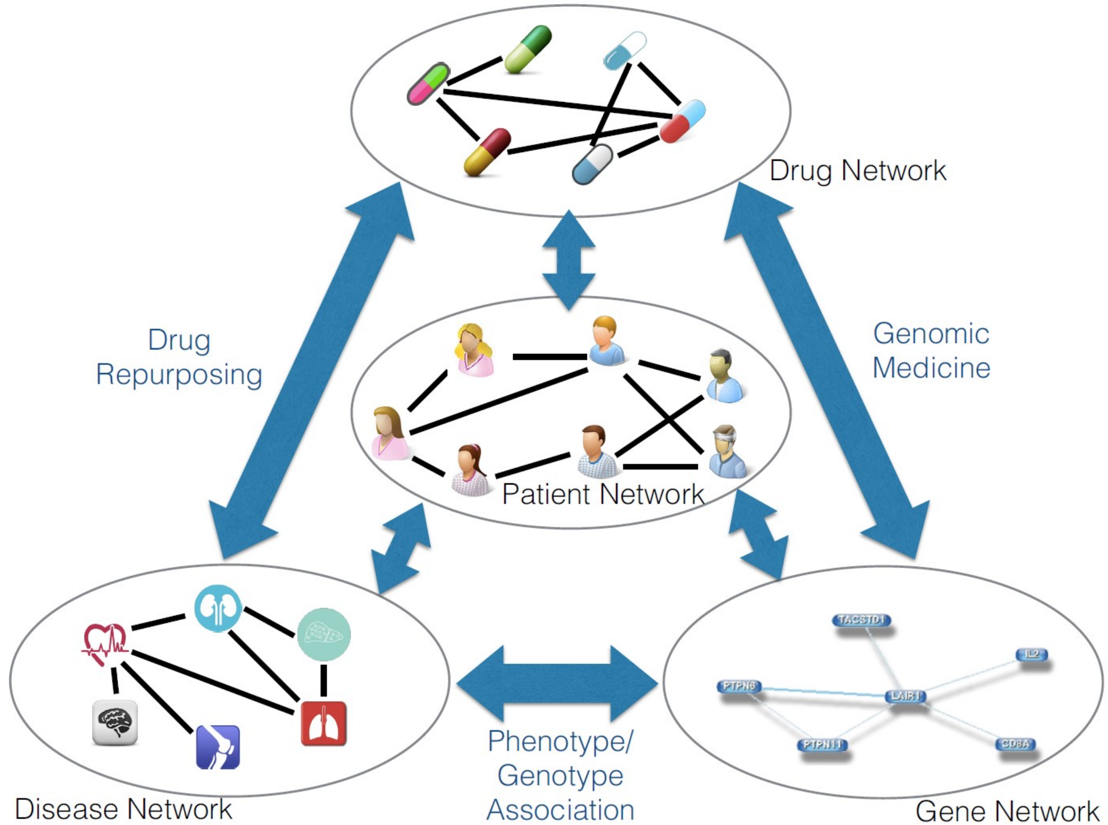




Bio-medical Network

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

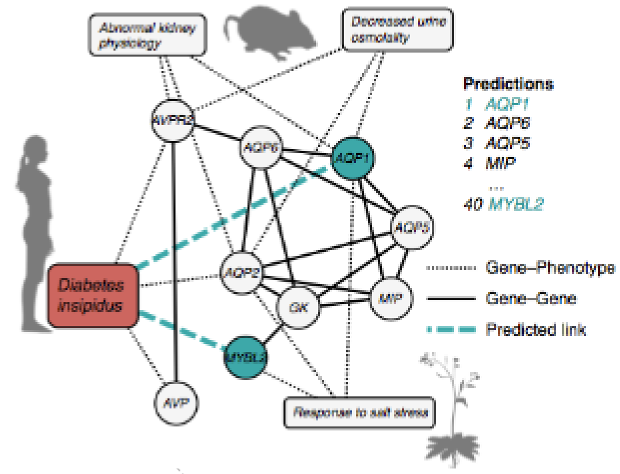




http://web.cs.ucla.edu/~yzsun/Tutorials.htm
https://sites.google.com/site/feiwang03/talks

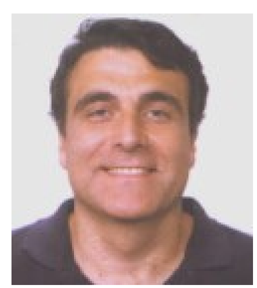
Problems in HIN

- 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

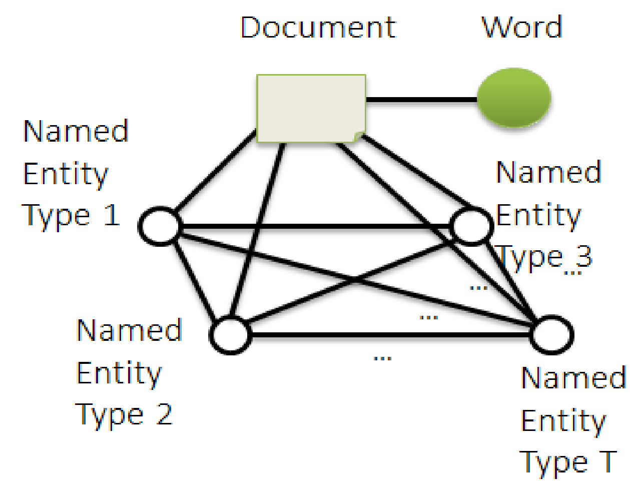
Christos' students or close collaborators

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

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]

If there are many meta-paths, how to integrate them into a machine learning algorithm?

Representative Applications

- Text Classification
  – Unsupervised fusion for many meta-paths
- Recommender System
  – Feature based instead of similarity based fusion for heterogeneous linking
- Malware Detection
  – Supervised fusion using multi-kernel learning

Text Categorization: Two Challenges

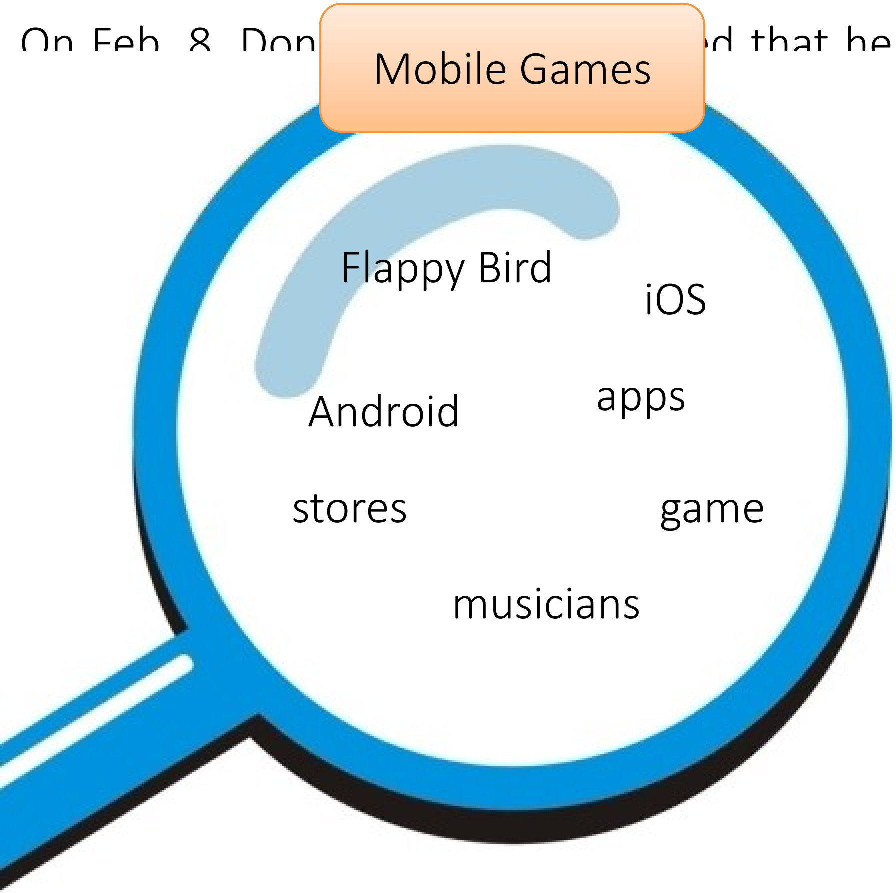
- Impacts many applications!
  - ✓ Social network analysis, health care, machine reading …
- Traditional approach:

Label data → Train a classifier → Make prediction

- Two challenges:
  - ✓ Representation
  - ✓ Labels

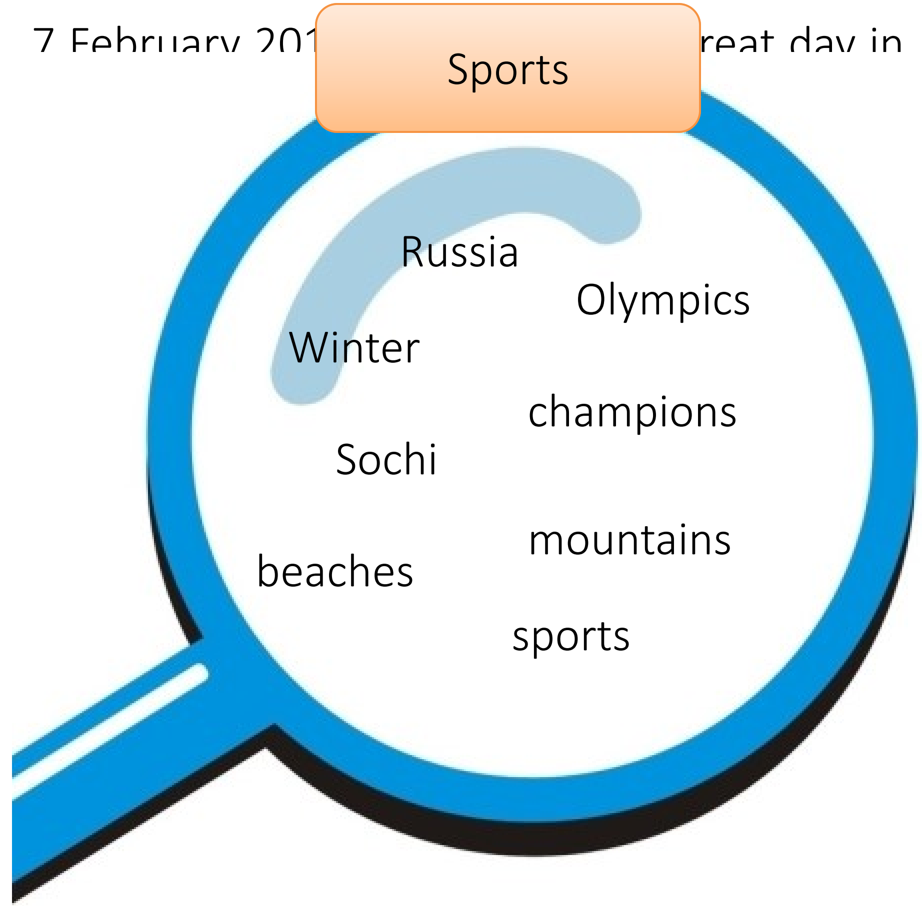
Representation: Bag-of-words

On Feb. 8, Don ... d that he



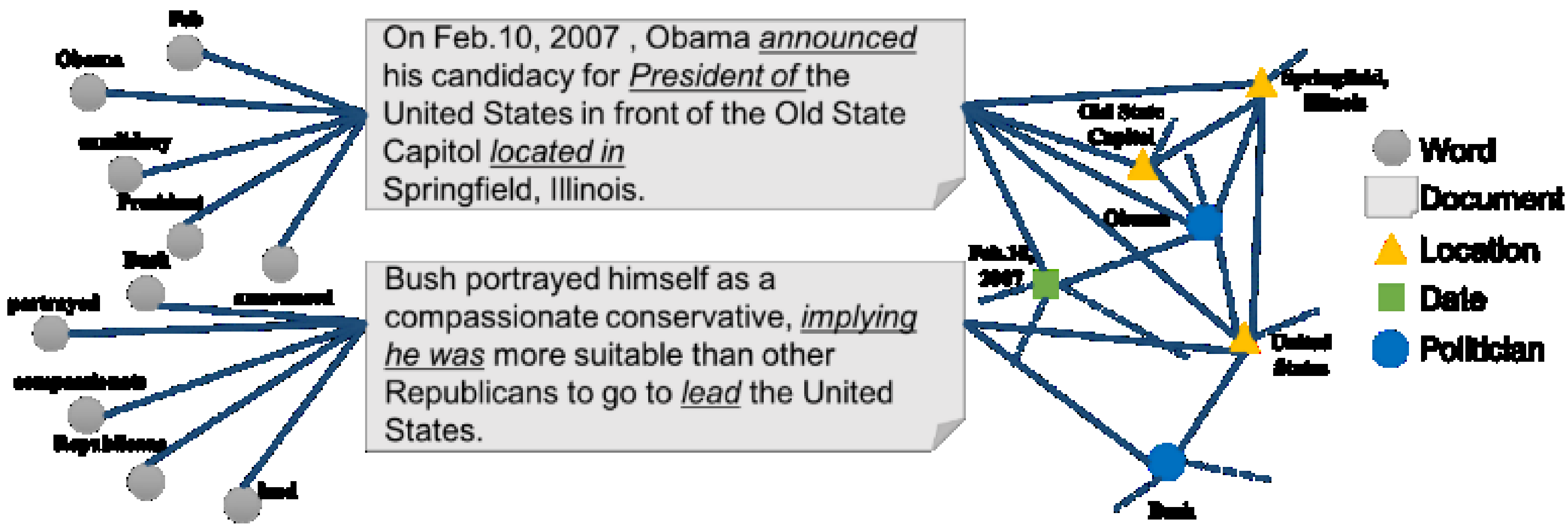
Internet trolls."

7 February 201 ... reat day in

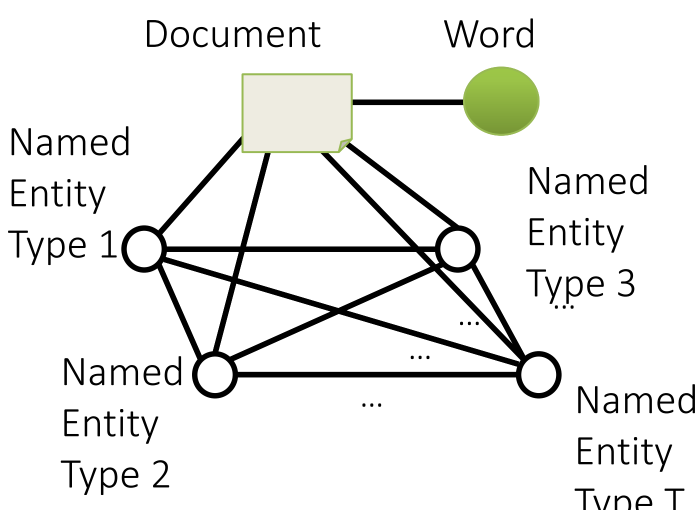


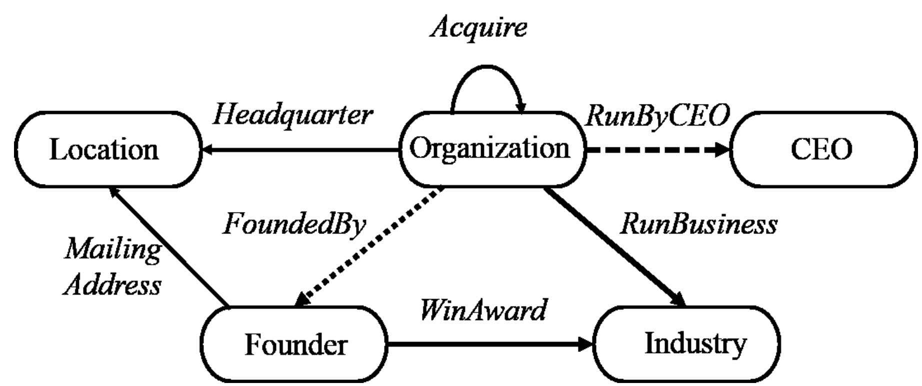
from 7 to 23 February 2014.

What Semantics Can HIN Provide for Text?

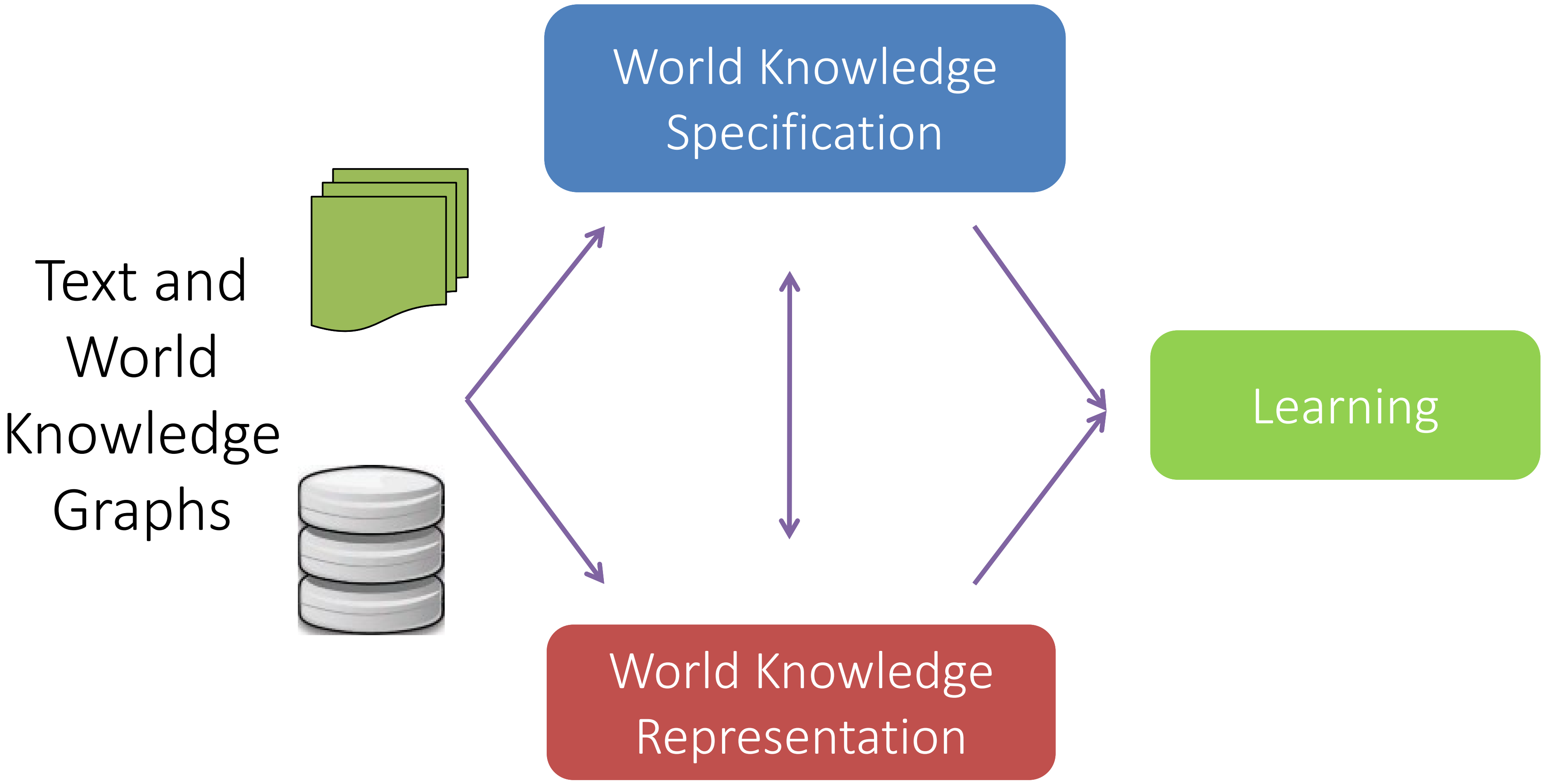


HIN **network-schema**: network with multiple object types and/or multiple link types.





Knowledge Empowered Text Classification



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

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Trump is the president of the United States of America

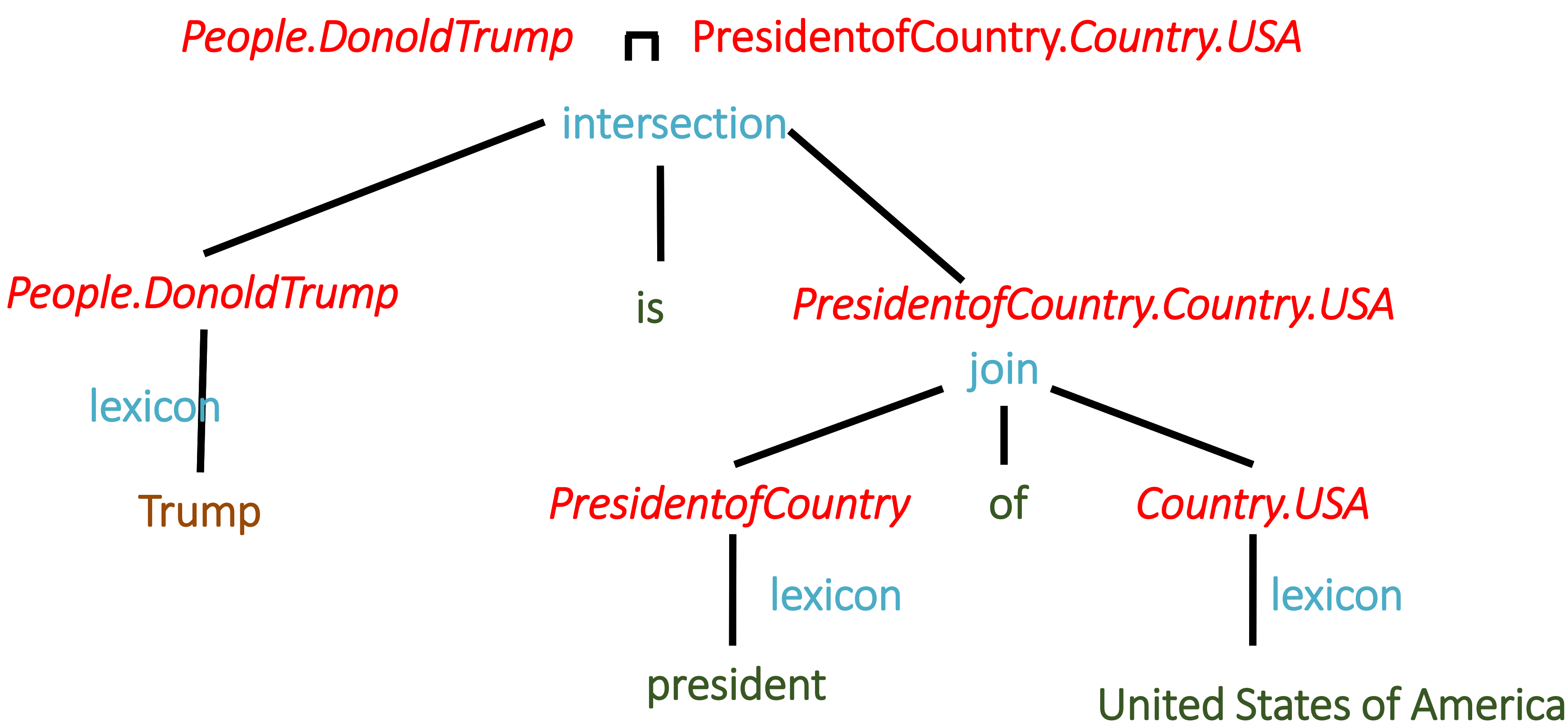
Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

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

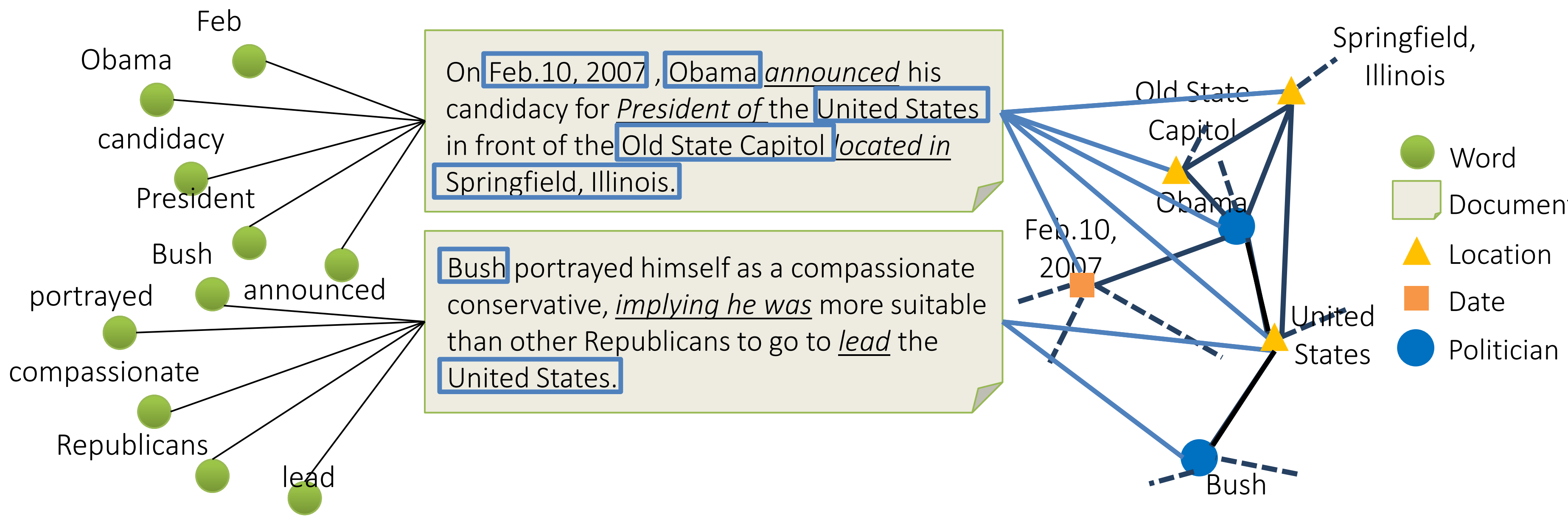
- Motivation: [Berant et al. EMNLP'13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase
  - Existing semantic parsing approaches, that require expert annotation
  - Scales to large scale knowledge-bases, supervised by the QA pairs
- We extend it to document analysis.

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

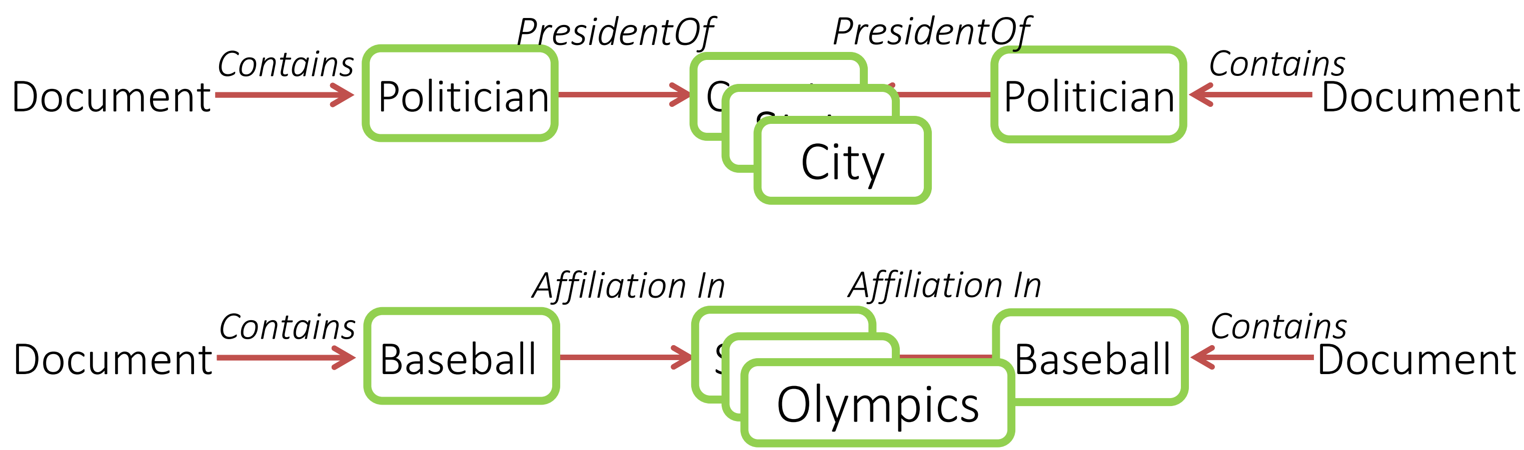
Document Trump is the president of the United States of America



Example Meta-paths in Text HIN



Capturing higher-order and more subtle relations



KnowSim

An ensemble of similarity measures defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

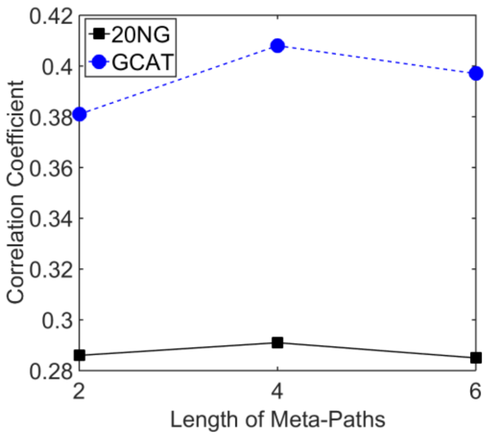
$$KS(d_i, d_j) = \frac{2 \times \boxed{\sum_m^{M'} w_m \,|\{p_{i \to j} \in P_m\}|}}{\boxed{\sum_m^{M'} w_m \,|\{p_{i \to i} \in P_m\}| + \sum_m^{M'} w_m \,|\{p_{j \to j} \in P_m\}|}}$$

Semantic broadness: the number of total meta-paths between themselves.

- Intuition: The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.

- KnowSim is computed in nearly linear time.

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM'15.



Challenges

of meta-paths:
20NG (325) GCAT (1,682)

Number of meta-paths could be very large.

$$KS(d_i, d_j) = \frac{2 \times \sum_m^{\boxed{M'}} w_m \,|\{p_{i \to j} \in P_m\}|}{\sum_m^{M'} \boxed{w_m} \,|\{p_{i \to i} \in P_m\}| + \sum_m^{M'} w_m \,|\{p_{j \to j} \in P_m\}|}$$

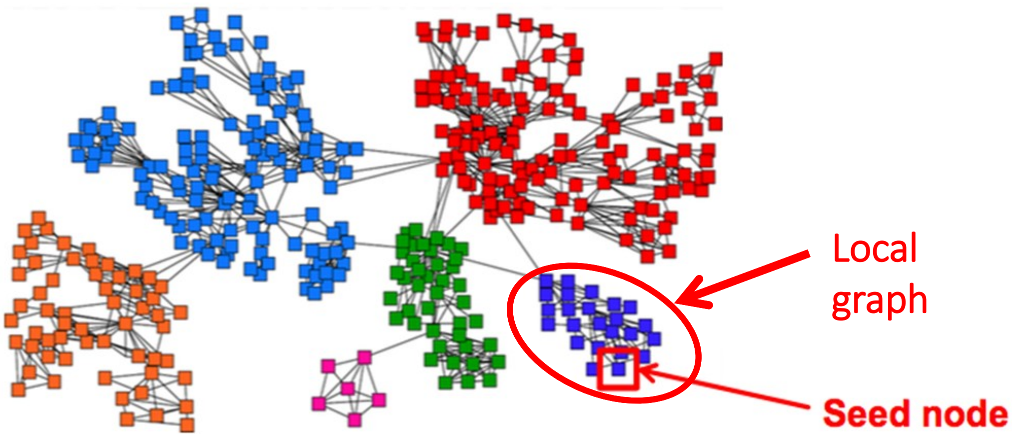
The weight/importance of each meta-path is different when the domain is different.

#1: How should we generate the large number of meta-paths at the same time?
Previous studies only focus on single meta-path, enumeration over the network is OK. In real world, what will happen when thousands of meta-paths are needed?

#2: How should we decide the weight of each meta-path?
Previous studies treat them equally. In real world, different meta-path should contribute differently in various domains.

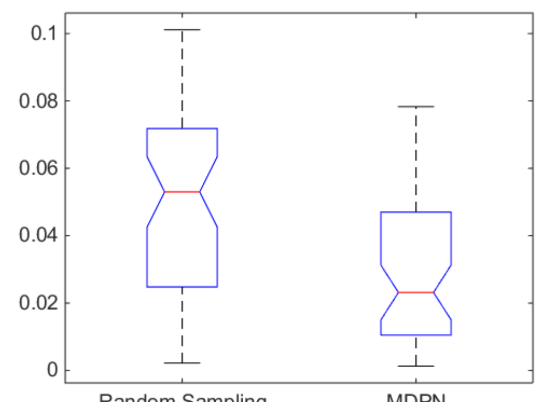
Meta-Path Dependent Random Walk

Intuition: Discovering compact sub-graph based on seed document nodes.



- Compute **Personalized PageRank (PPR)** around seed nodes.
- The random walk will get trapped inside the blue sub-graph.

- Algorithm outline
  - Run **PPR** (approximate connectivity to seed nodes) with teleport set = {**S**}
  - Sort the nodes by the decreasing **PPR** score
  - **Sweep** over the nodes and find compact **sub-graph**.
  - Use the sub-graph instead of the whole graph to compute # of meta-paths between nodes.



Frobenius norm of approximation of commuting matrices on 20NG dataset

Meta-Path Ranking

of meta-paths: **20NG (325)** and **GCAT (1,682)**

- Maximal Spanning Tree based Selection [Sahami, 1998]

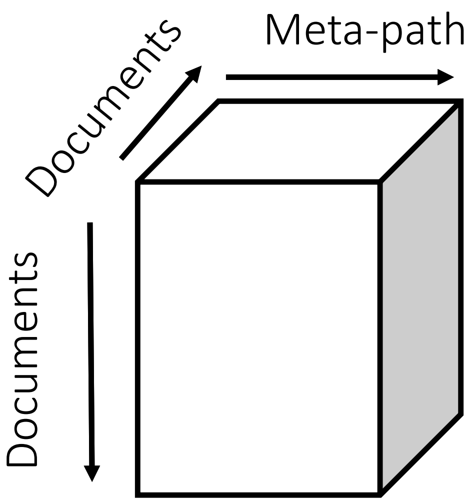
$$\frac{\sum_{j\neq i}^{M}\cos(\boldsymbol{D}_{.,j_1},\boldsymbol{D}_{.,j_2})}{M-1}$$

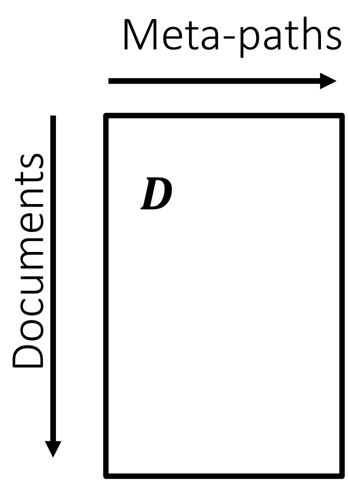
Select meta-paths with the largest dependencies with others

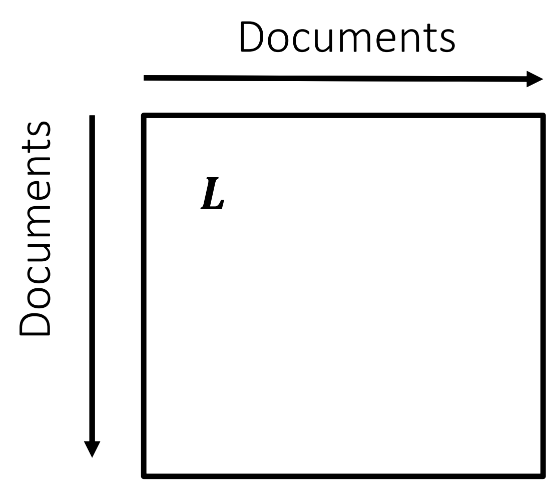
- Laplacian Score based Selection [He, 2006]

$$L_j = \frac{\widetilde{\boldsymbol{D}_{.,j}}^T \boldsymbol{L}\boldsymbol{D}_{.,j}}{\widetilde{\boldsymbol{D}_{.,j}}^T \Lambda \boldsymbol{D}_{.,j}}$$

Select a meta-path in **discriminating documents** from different clusters







Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

The number is reported in [X. Dong et al. KDD'14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.

Spectral Clustering with KnowSim (ICDM'15)

- Non-linear clustering (Ng et al., NIPS'01)
  - Construct k-NN graph based on pair-wise similarities
  - Perform k-means over Eigen vectors of the graph Laplacian

Datasets	Similarity Measures	BOW	BOW+TOPIC	BOW+TOPIC+ENTITY
20NG	Cosine	0.3440	0.3461	0.4247
	Jaccard	0.3547	0.3517	0.4292
	Dice	0.3440	0.3457	0.4248
GCAT	Cosine	0.3932	0.4352	0.4106
	Jaccard	0.3887	0.4292	0.4159
	Dice	0.3932	0.4355	0.4112

	KnowSim+UNIFORM	KnowSim+MST	KnowSim+LAP
20NG	0.4304	0.4304	**0.4461 (+3.9%)**
GCAT	0.4463	0.4653	**0.4736 (+8.8%)**

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM'15.

Classification Results with SVM needs a positive semi-definite(PSD) kernel matrix (AAAI'16)

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+other MetaPaths	DWD+other MetaPaths
20NG-SIM	91.60%	92.68%	93.38%
20NG-DIF	97.20%	98.01%	98.45%
GCAG-SIM	94.82%	96.04%	98.10%
GCAT-DIF	91.19%	91.88%	93.51%

Wang et al., Text Classification with Heterogeneous Information Network Kernels. AAAI'16.

Results on Semi-supervised Learning (IJCAI'17)

- 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**

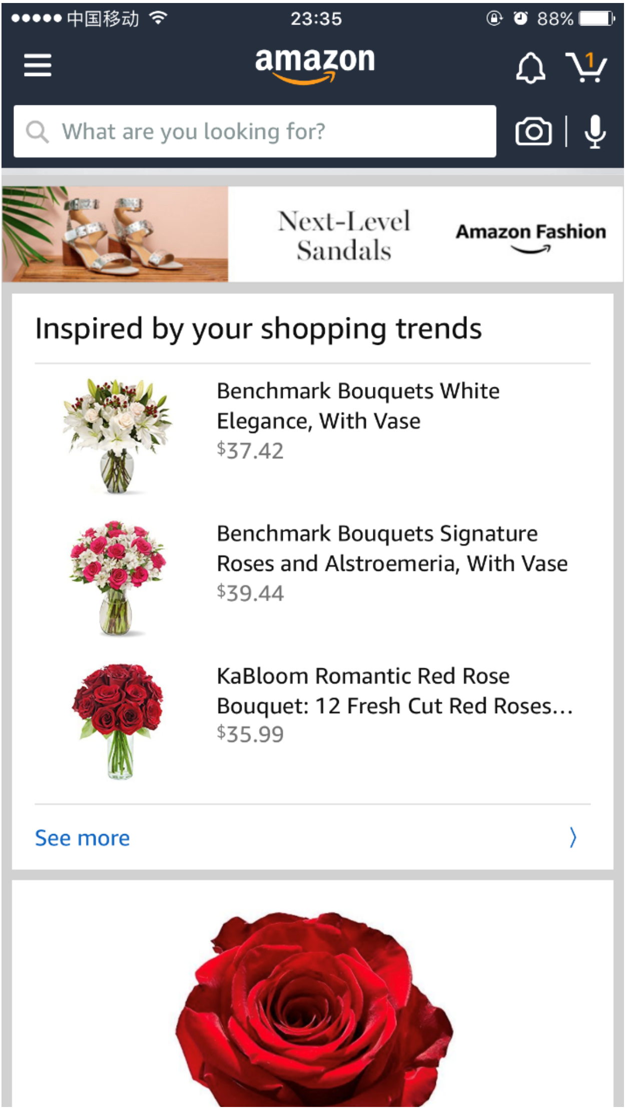
- We show our results of five labeled training data for each class. All the numbers are averaged accuracy (in percentage %) over 50 random trials.

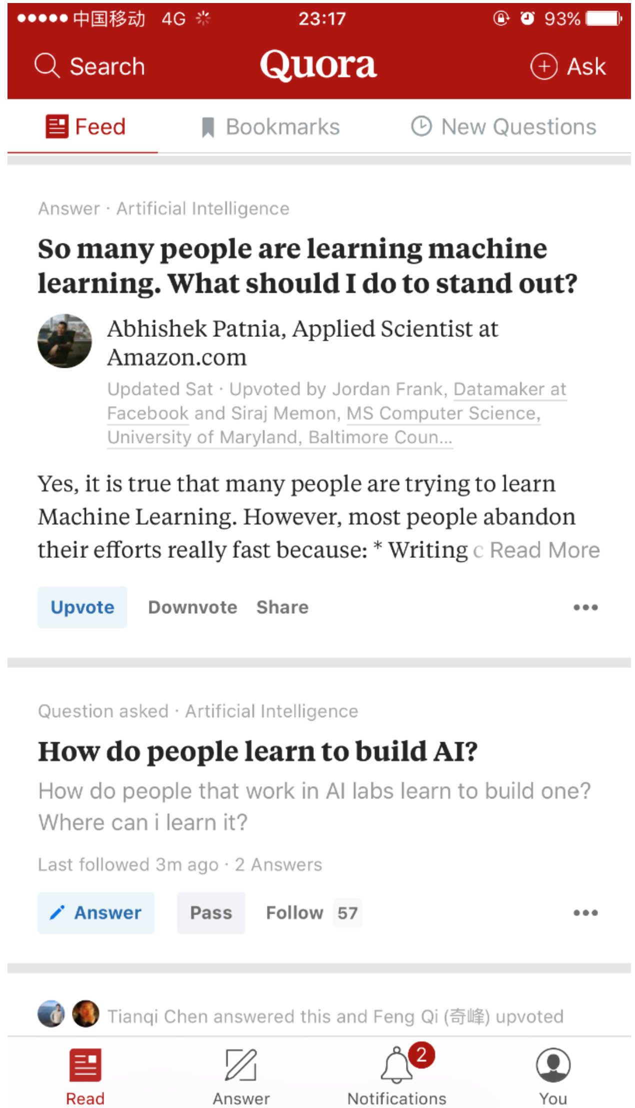
Jiang et al., Semi-supervised Learning over Heterogeneous Information Networks by Ensemble of Meta-graph Guided RandomWalks. IJCAI'17.

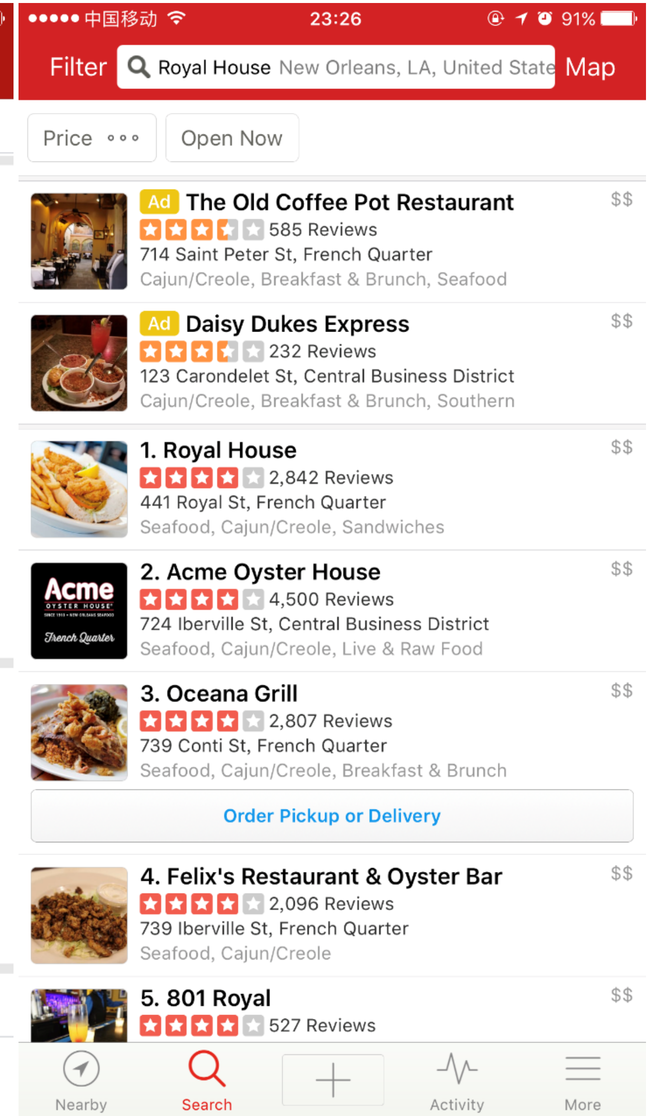
Representative Applications

- Text Classification
  – Unsupervised fusion for many meta-paths
- Recommender System
  – Feature based instead of similarity based fusion for heterogeneous linking
- Malware Detection
  – Supervised fusion using multi-kernel learning

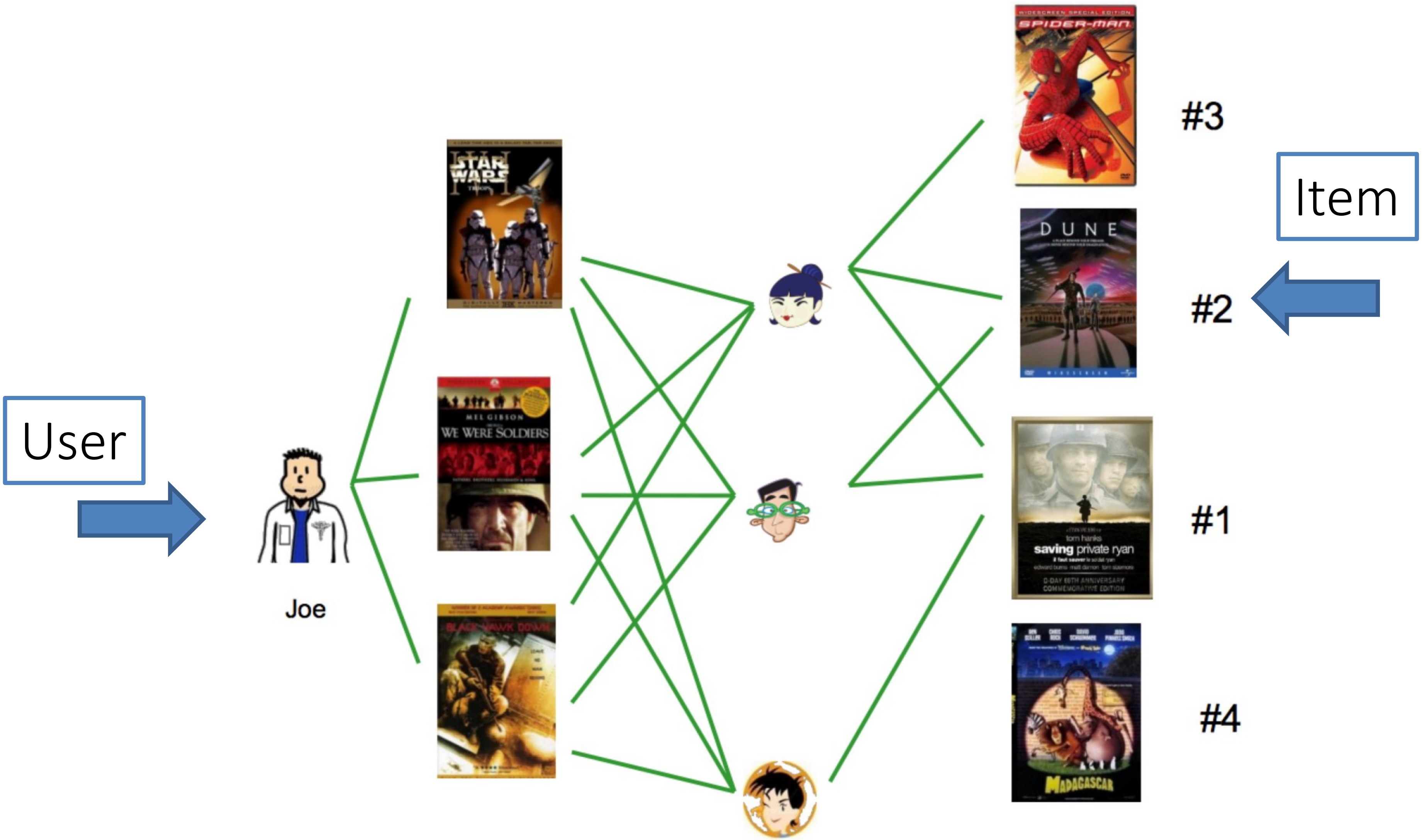
RS is Everywhere Nowadays







Typical Recommendation Problem

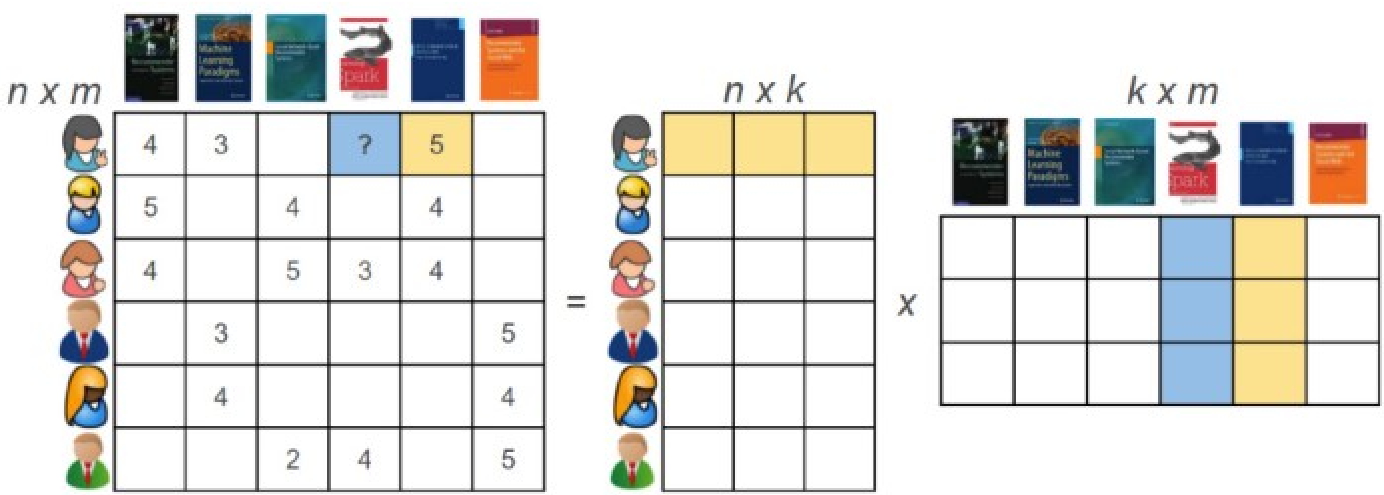


Matrix Factorization

- 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*

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

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

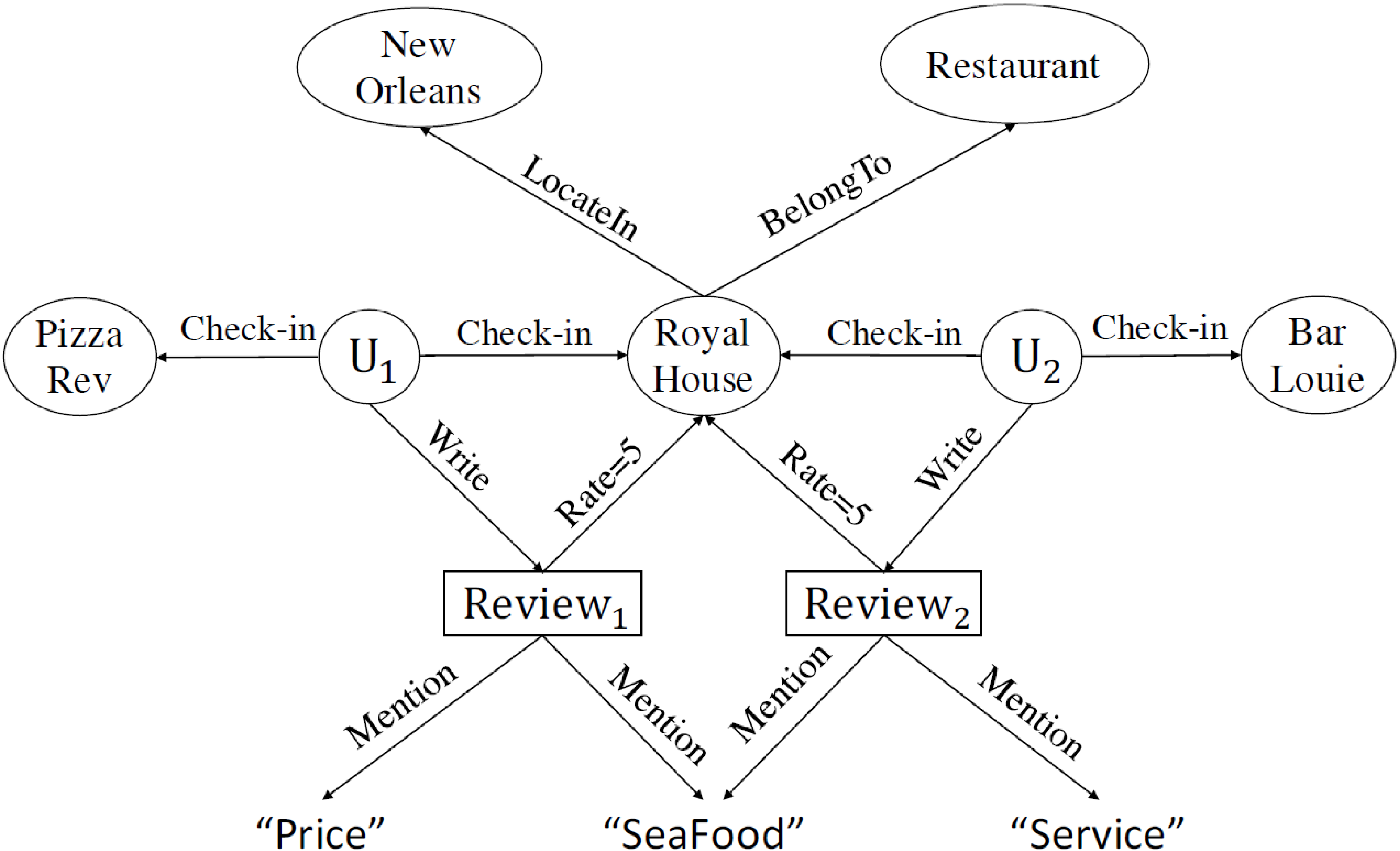


$$\min_{\mathbf{U},\mathbf{B}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{u}_i\mathbf{b}_j)^2 + \frac{\lambda_1}{2}||\mathbf{U}||_F^2 + \frac{\lambda_2}{2}||\mathbf{B}||_F^2$$

Other Existing Approaches

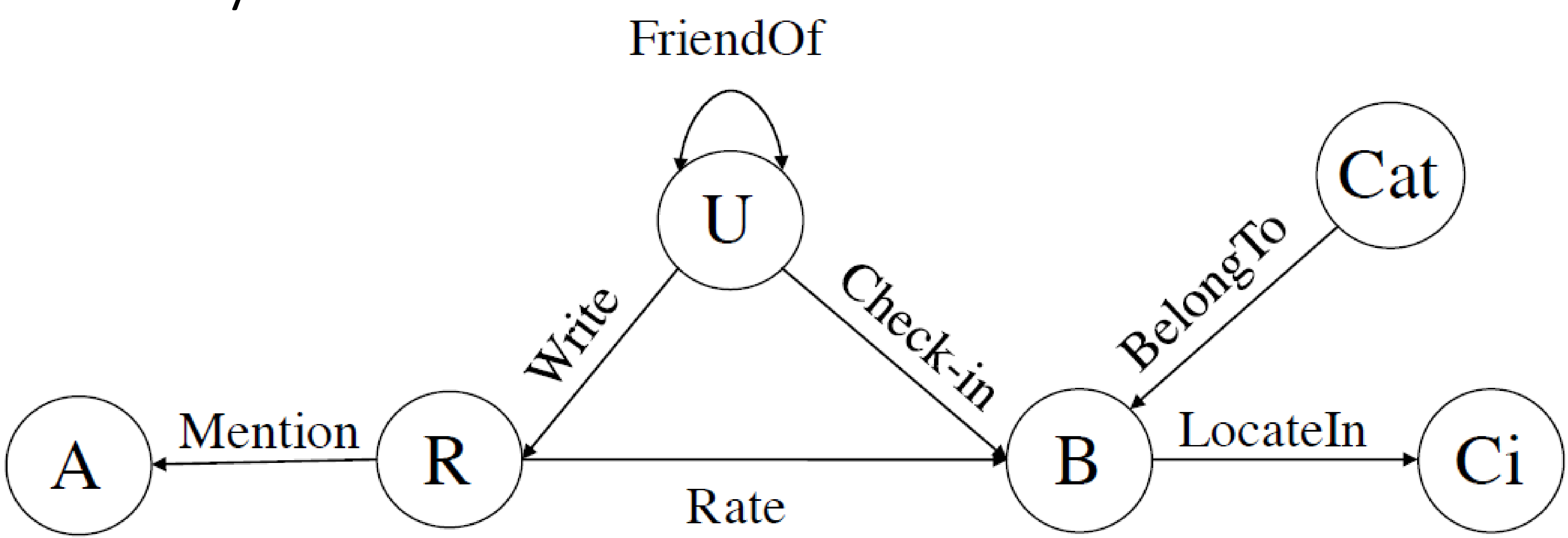
- Collaborative Filtering: Recommend items based only on the users past behavior
  - User based: find similar users for what they liked
  - Item based: find similar items which I have liked
- Content based: extract features for items
- Personalized learning to rank
- Demographic: user profiling
- Social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

It’s a Heterogeneous Information Network!

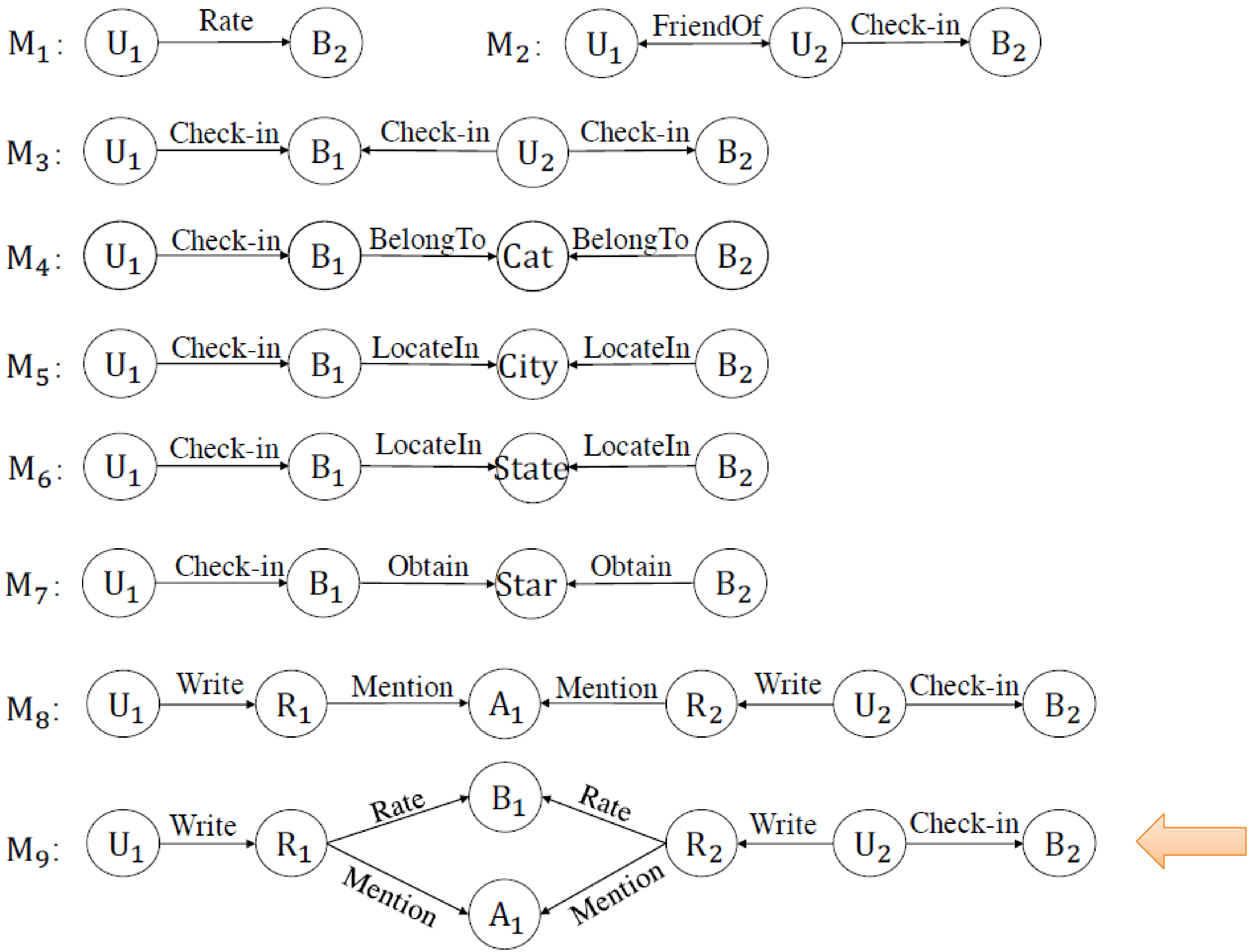


A Typical Network Schema of Yelp

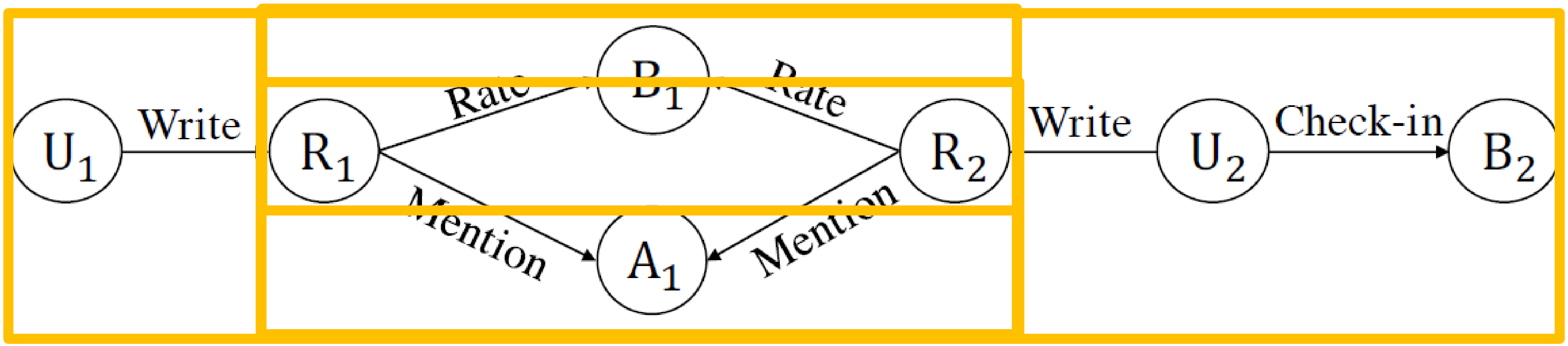
- R: reviews;
- U: users;
- B: business;
- Cat: category of item;
- Ci: city



Meta-graphs Extracted From Yelp



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?

- Existing works still work on similarities

- 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

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**

	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]

Comparison Results

Traditional Approaches

HIN Based Approaches

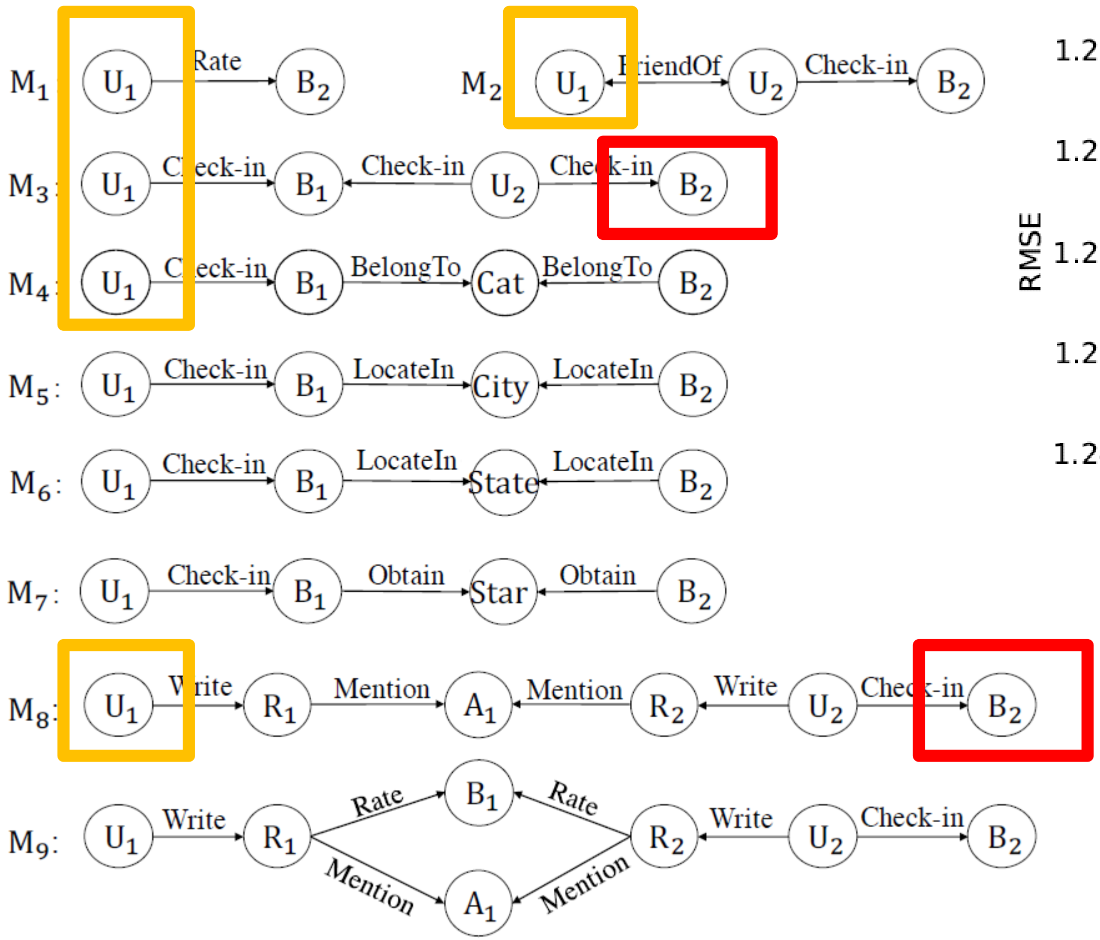
	Amazon-200k	Yelp-200k	CIKM-Yelp	CIKM-Douban
RegSVD	2.9656 (+60.0%)	2.5141 (+49.9%)	1.5323 (+27.7%)	0.7673 (+9.0%)
FMR	1.3462 (+11.9%)	1.7637 (+28.6%)	1.4342 (+22.8%)	0.7524 (+7.2%)
HeteRec	2.5368 (+53.2%)	2.3475 (+47.0%)	1.4891 (+25.6%)	0.7671 (+9.0%)
SemRec	- -	1.4603 (+13.8%)	1.1559 (+4.2%)	0.7216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

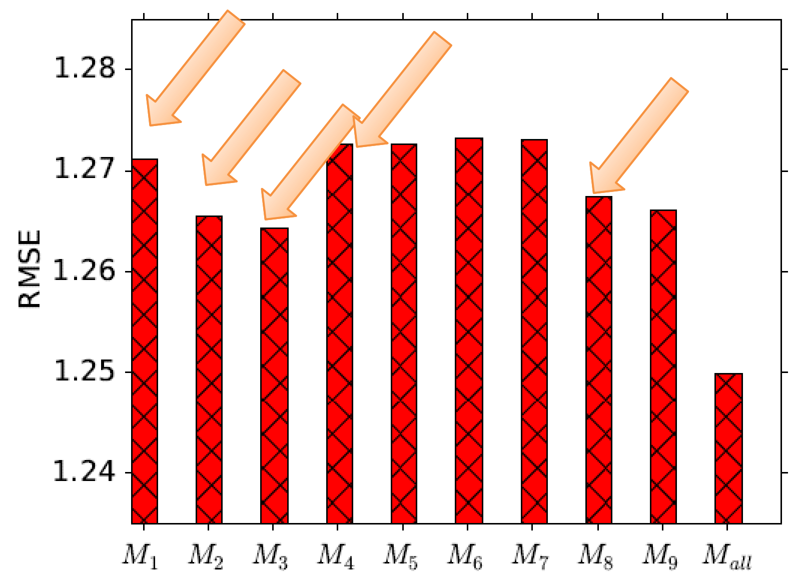
- HeteRec [Yu et al., WSDM'14]:
  - Factorize each meta-path
  - Ensemble using the recovered matrices
  - Item-based CF
- SemRec [Shi et al., CIKM'15]:
  - Ensemble of original similarity matrices based on different meta-paths
  - User based CF

	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	$M_1, M_2, M_4 - M_7, M_9$





Representative Applications

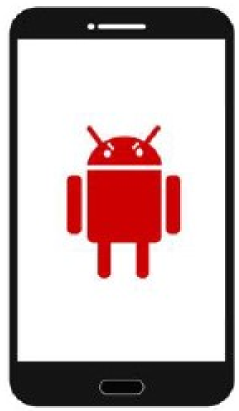
- Text Classification
  - Unsupervised fusion for many meta-paths

- Recommender System
  - Feature based instead of similarity based fusion for heterogeneous linking

  - Malware Detection
    - Supervised fusion using multi-kernel learning

Malicious Software

Steal money

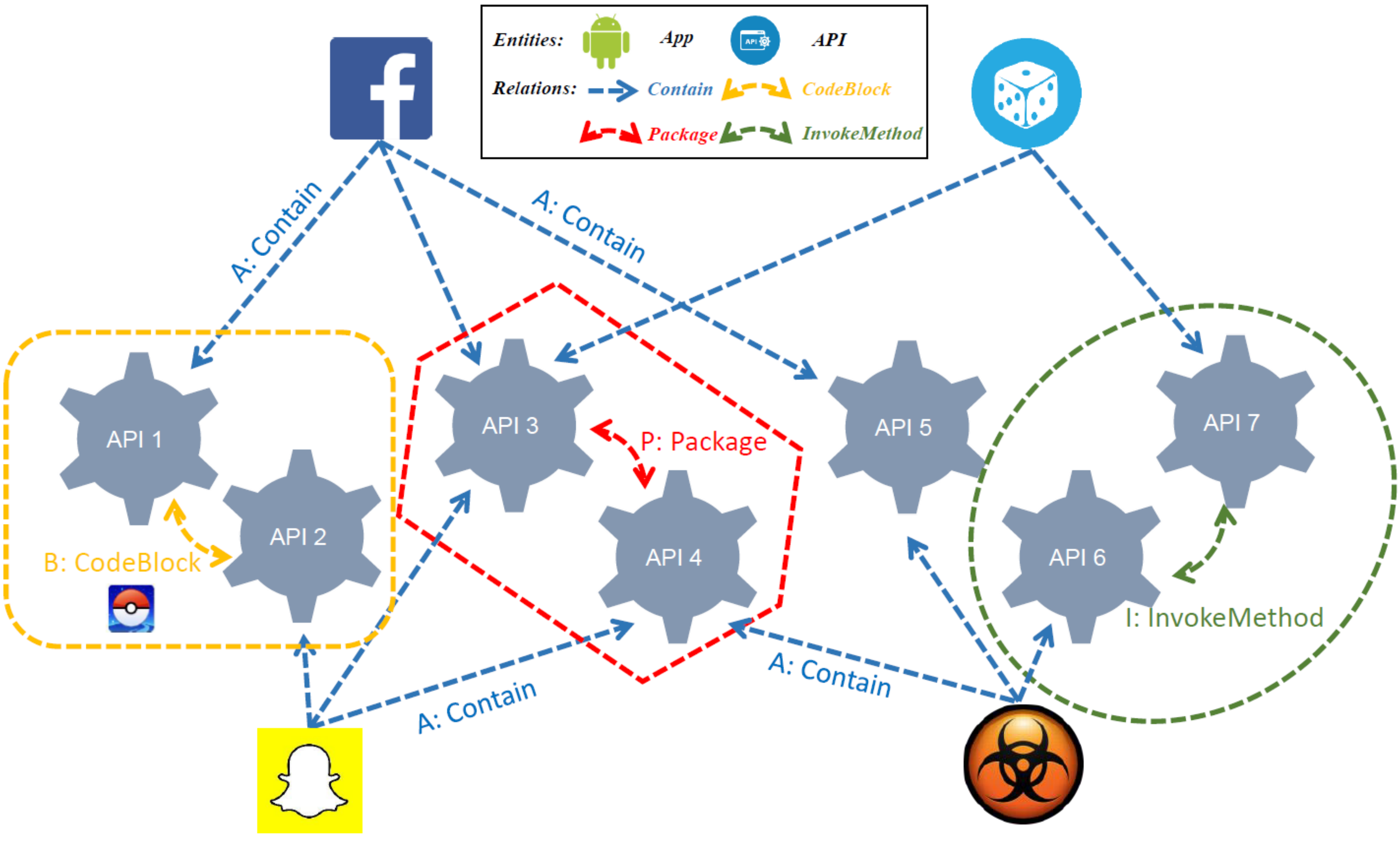


Send SMS message

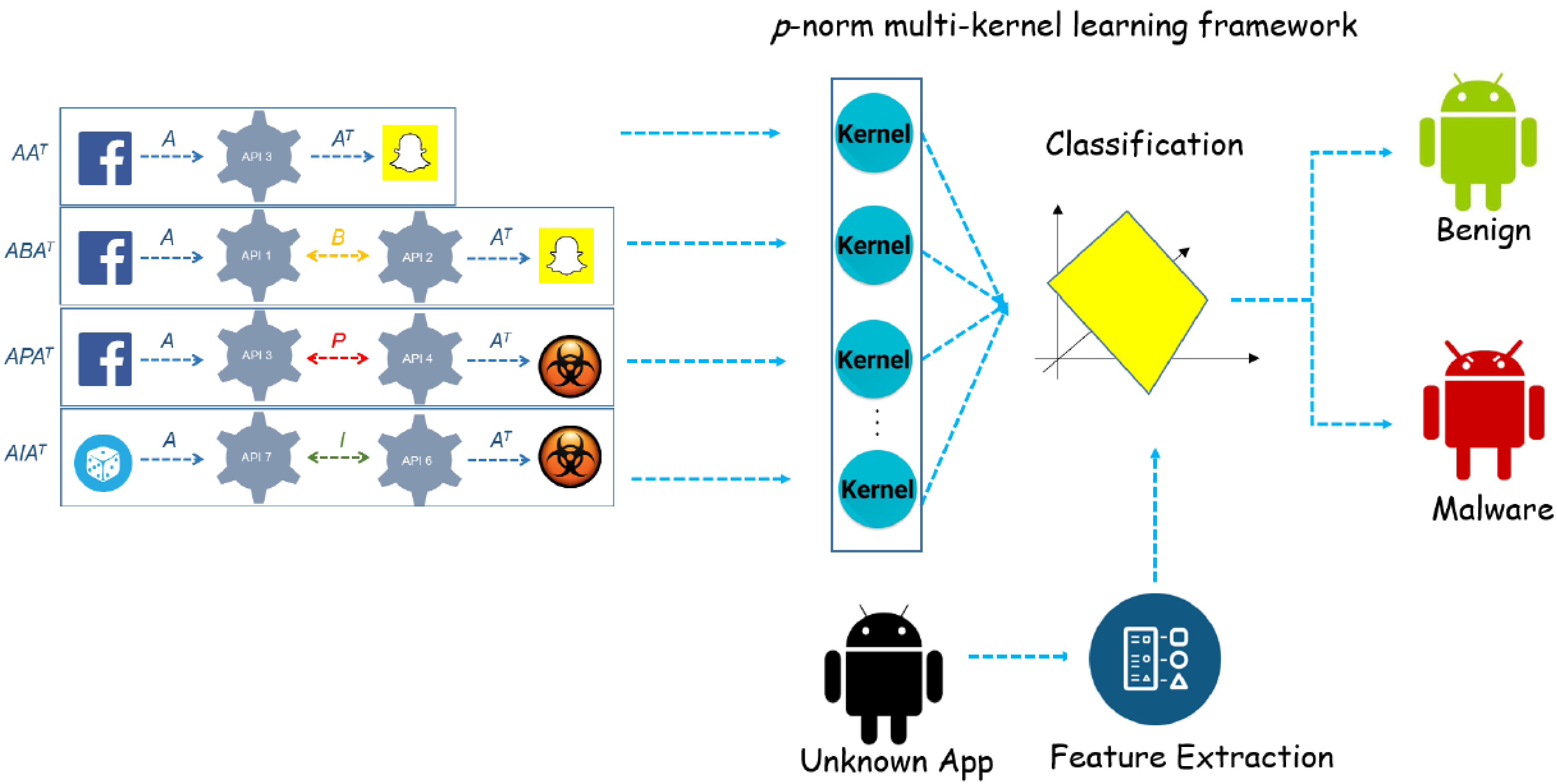
Push advertisement

Download unwanted app

Heterogeneous Information Network Representation



Multi-kernel Learning



Performance of Different Meta-paths

PID	Method	F1	β	ACC	TP	FP	TN	FN
1	AA^T	0.9529	0.1069	94.40%	283	19	189	19
2	ABA^T	0.9581	0.0900	95.00%	286	9	189	16
3	APA^T	0.9495	0.0858	94.20%	273	0	198	29
4	AIA^T	0.9183	0.0623	90.40%	270	16	182	32
5	$ABPB^TA^T$	0.9479	0.0670	94.00%	273	1	197	29
6	$APBP^TA^T$	0.9502	0.0565	94.20%	277	4	194	25
7	$ABIB^TA^T$	0.8683	0.0639	84.60%	254	29	169	48
8	$AIBI^TA^T$	0.8722	0.0639	85.00%	256	29	169	46
9	$APIP^TA^T$	0.8373	0.0445	81.20%	242	34	164	60
10	$AIPI^TA^T$	0.8761	0.0572	86.60%	237	2	196	65
11	$ABPIP^TB^TA^T$	0.9184	0.0616	90.80%	259	3	195	43
12	$APBIB^TP^TA^T$	0.8597	0.0617	84.60%	236	11	187	66
13	$ABIPI^TB^TA^T$	0.9284	0.0426	91.80%	266	5	193	36
14	$AIBPB^TI^TA^T$	0.8237	0.0426	82.60%	218	3	195	84
15	$AIPBP^TI^TA^T$	0.8597	0.0469	81.60%	215	5	193	87
16	$APIBI^TP^TA^T$	0.8597	0.0458	84.60%	236	11	187	66
17	Combined-kernel (5)	0.9214	——	91.20%	258	0	198	44
18	Combined-kernel (16)	0.9740	——	96.80%	300	14	184	2
19	Multi-kernel (5)	0.9834	——	98.00%	297	5	193	5
20	Multi-kernel (16)	**0.9884**	——	**98.60%**	299	4	194	3

Conclusion

Heterogeneous information networks as explicit semantic analysis

We worked on how to fuse different kinds of information

Many interesting ideas and results and could be applied in the context of DL

Thank You! ☺