Efficient Random Walk Inference with Knowledge Bases

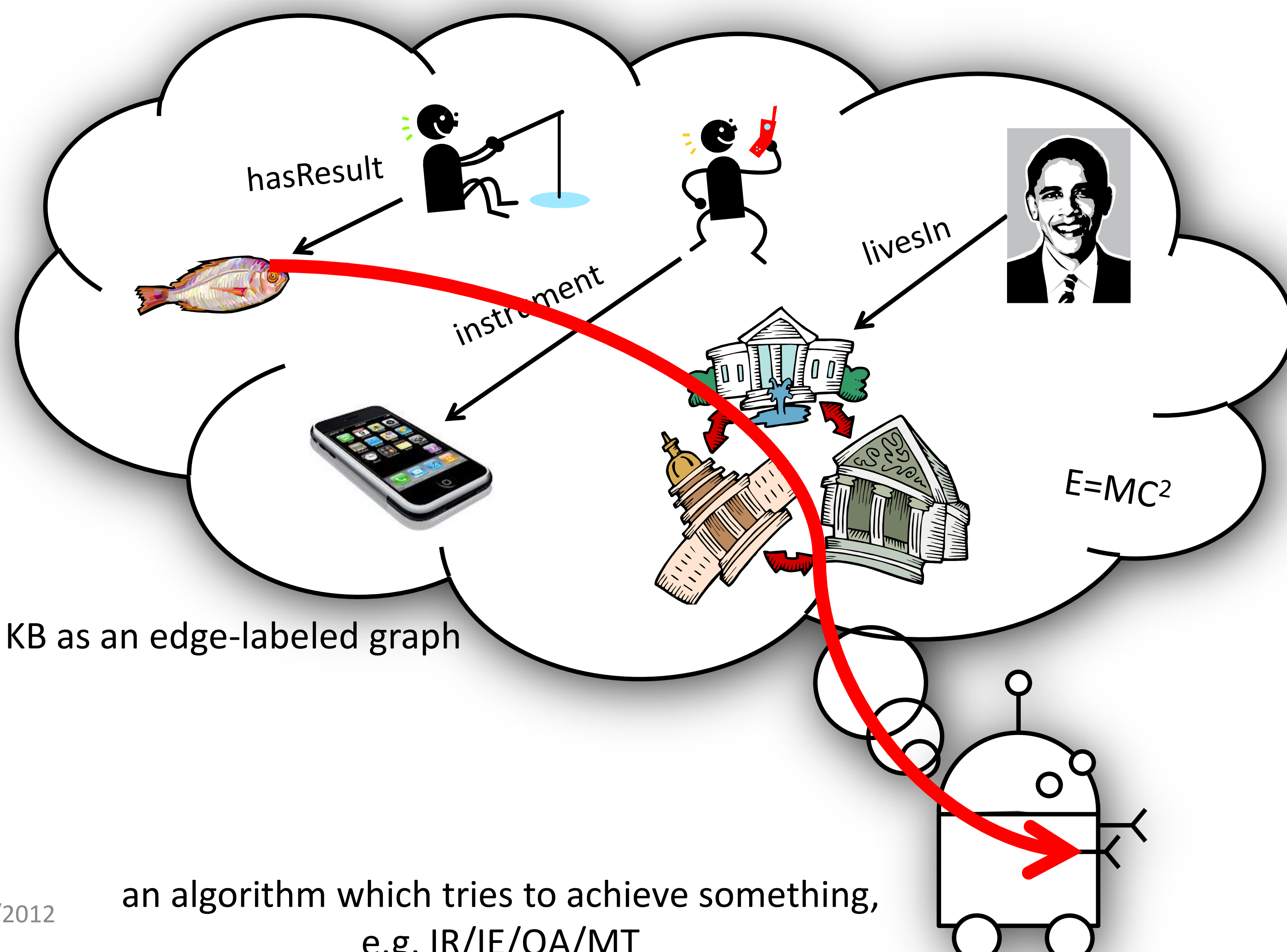
Ni Lao

Carnegie Mellon University

2012-7-11

Committee: William W. Cohen (Chair),
Teruko Mitamura, Tom Mitchell,
C. Lee Giles (Pennsylvania State University)

Knowledge itself is power.
--Francis Bacon



KB as an edge-labeled graph

an algorithm which tries to achieve something,
e.g. IR/IE/QA/MT

Link Prediction

-- a generic relational learning task

Given

a directed edge-labeled graph

a relation type r

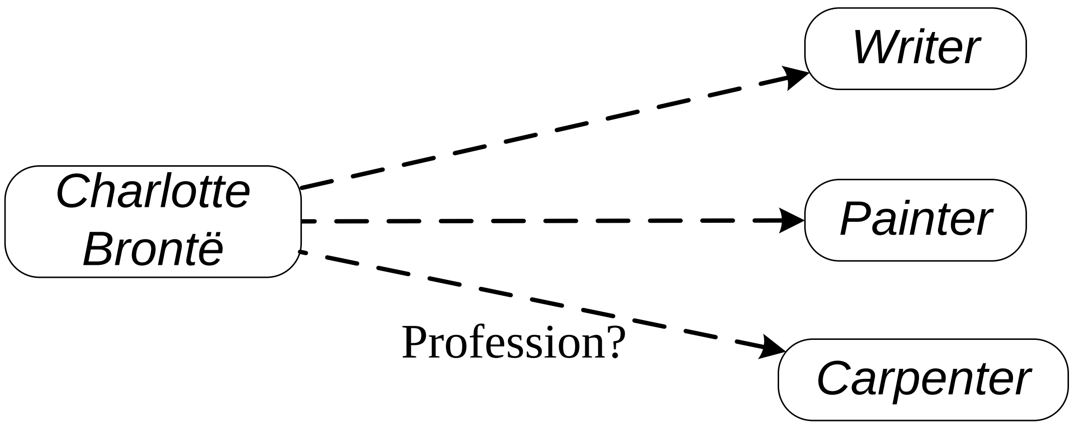
a source node s (also called a query)

Find

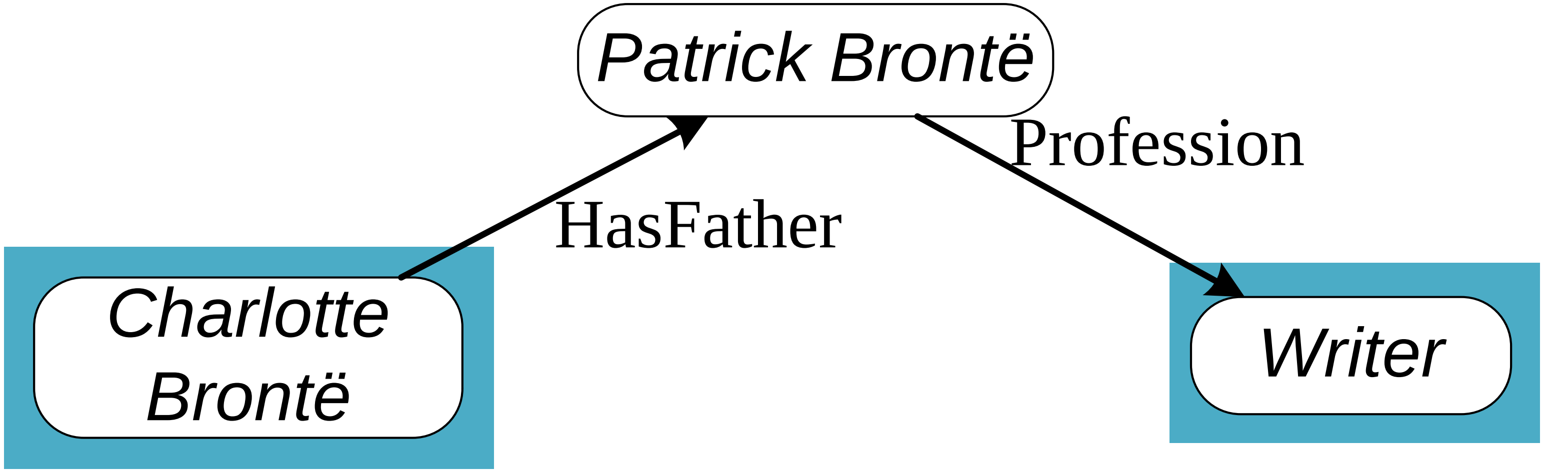
the set of nodes G, so that r(s,t) for each t in G

Infer New Knowledge

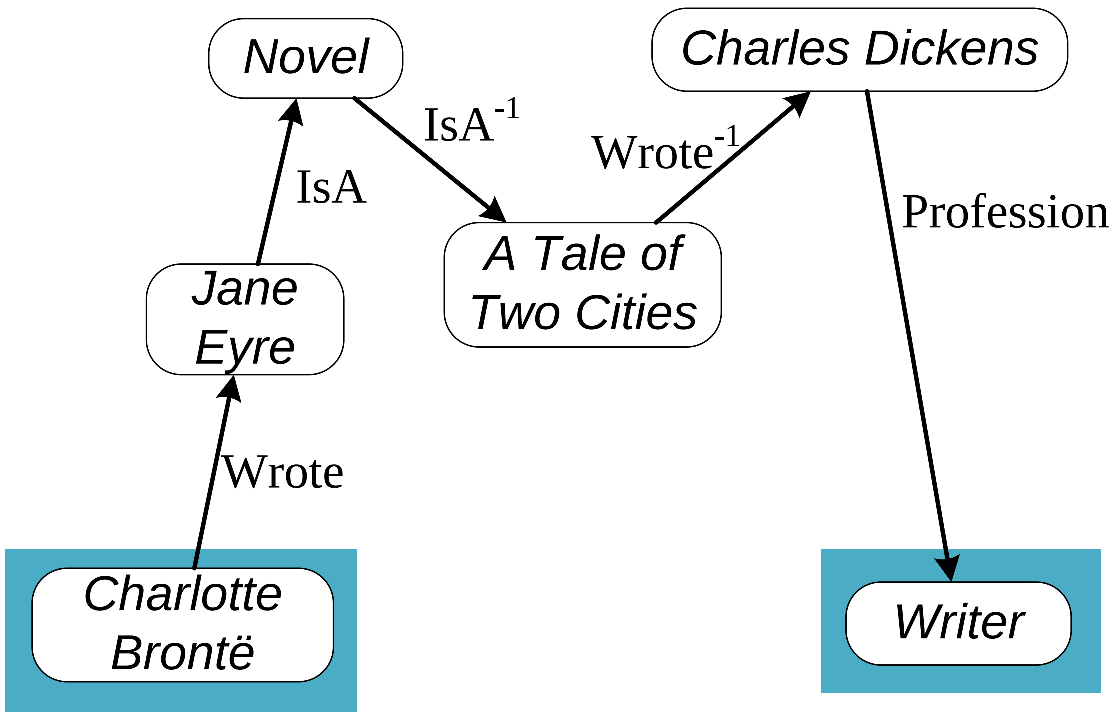
What is the profession of Charlotte Brontë?



Consider Friends/Family

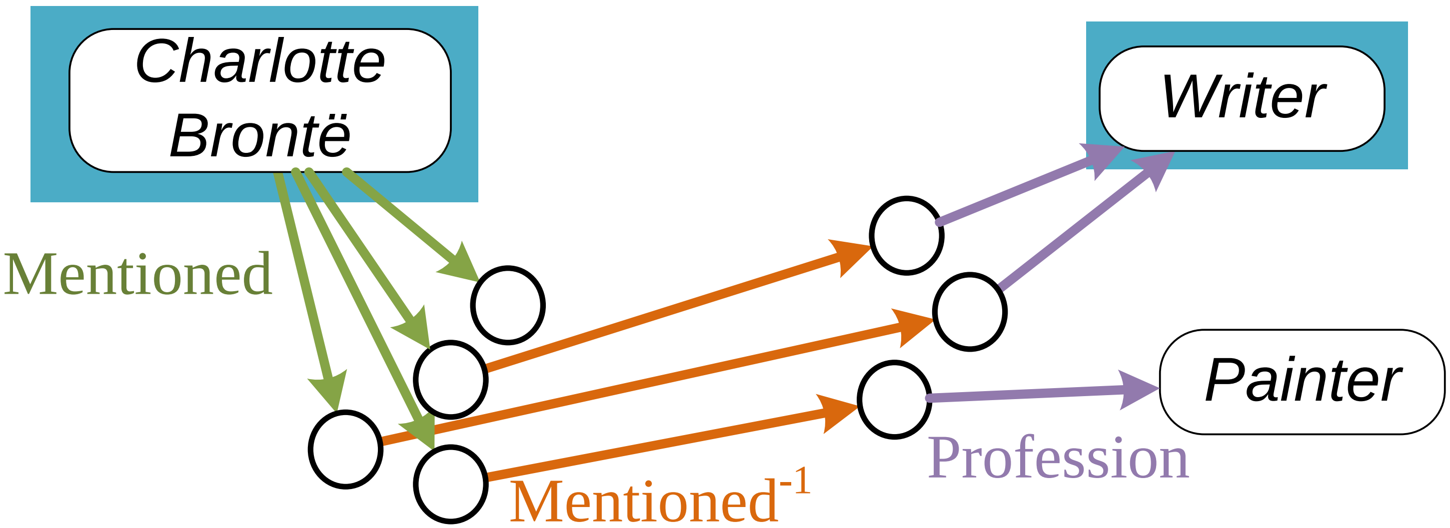


Consider Behaviors



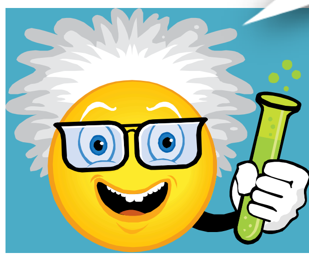
IsA^{-1} is the inverse of IsA
Wrote^{-1} is the inverse of Wrote

Consider Literatures/Publications



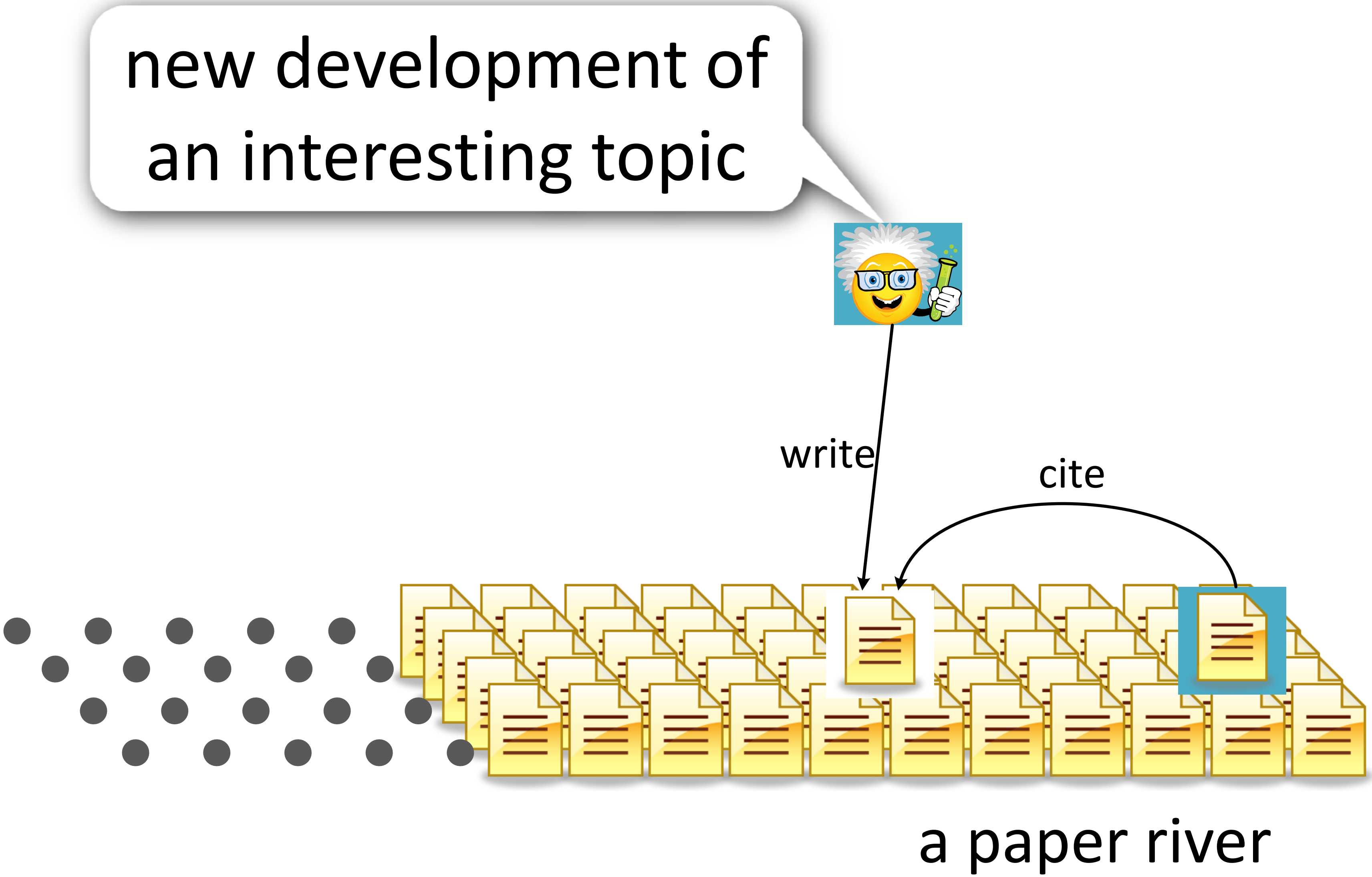
Reading Recommendation

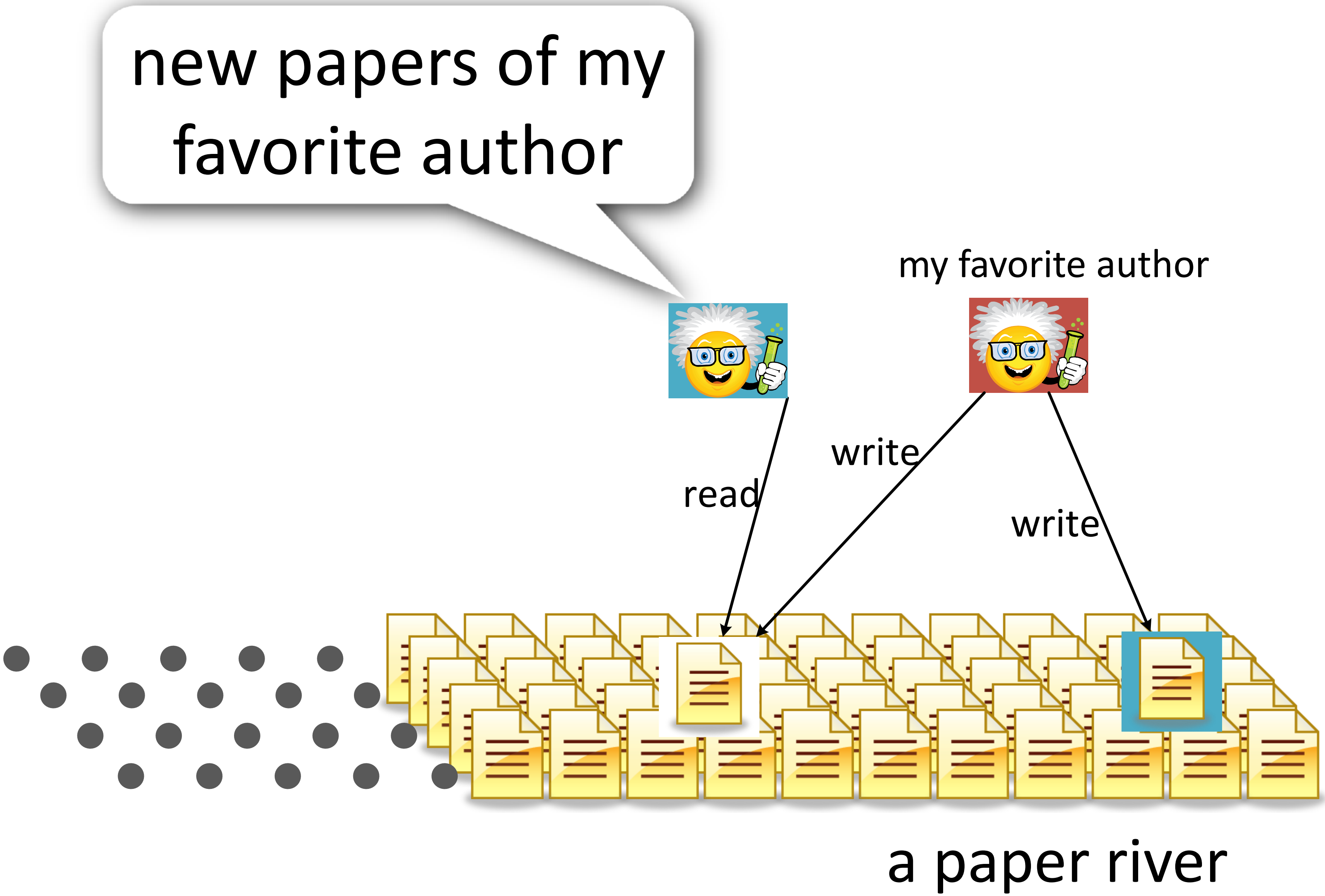
these are interesting papers



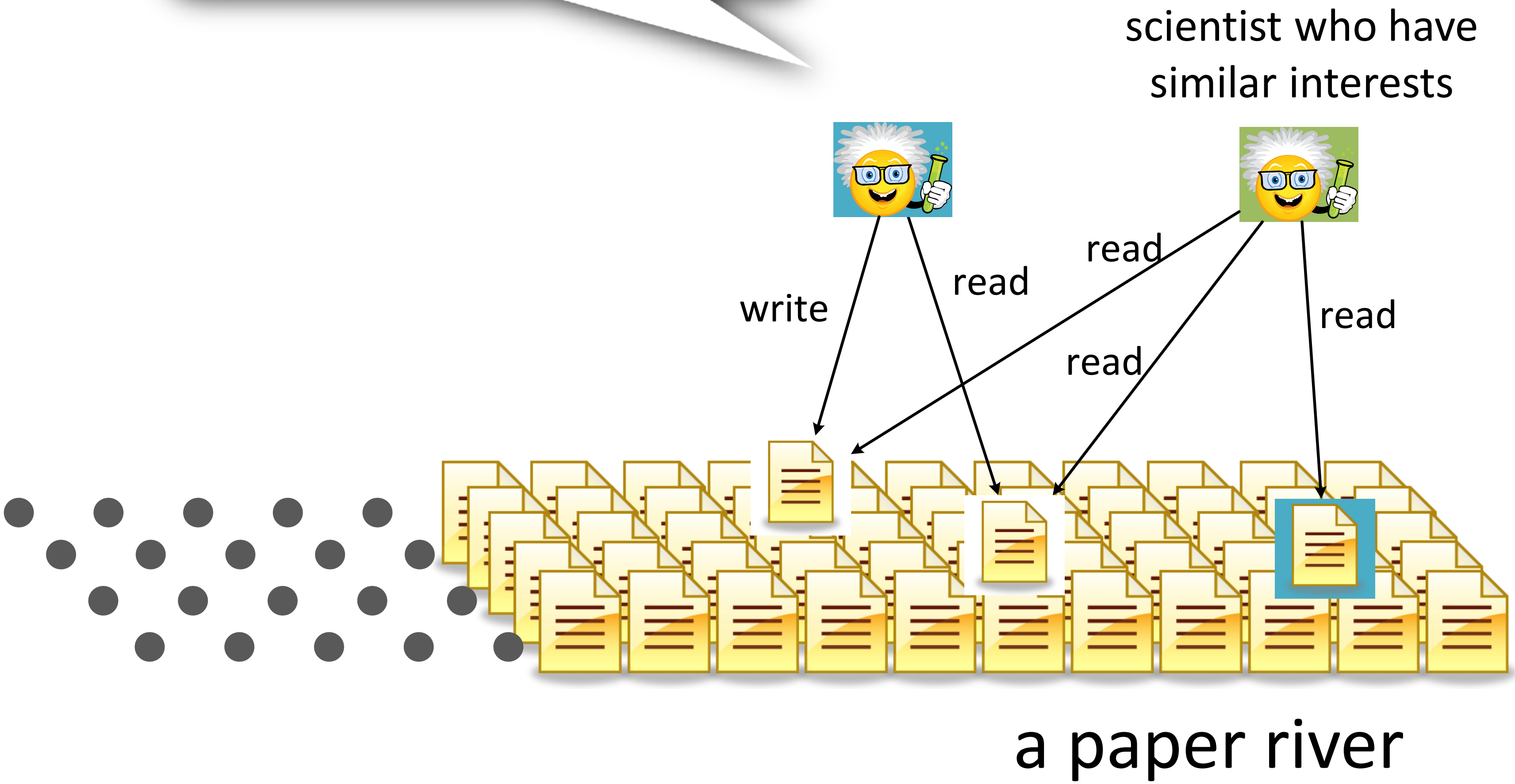
a paper stream

a paper river


new development of
an interesting topic
write
cite
a paper river


new papers of my favorite author
my favorite author
read
write
write
a paper river

social recommendation

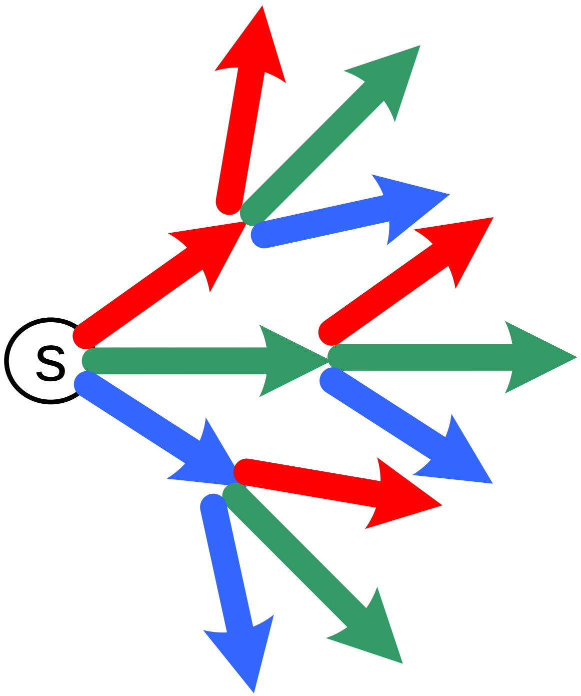


Relational Learning Goals

expressive	define features expressing sequences of relations on graph
robust	combine many such features when making decisions
scalable	efficiently discover and calculate such features

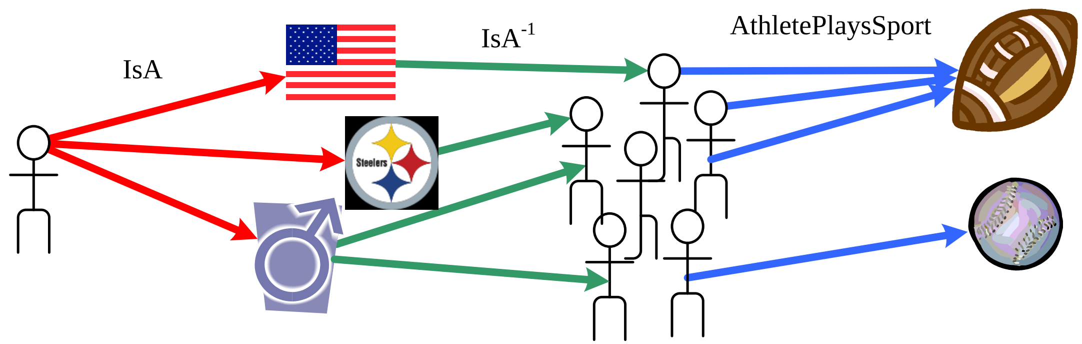
Why is relational learning computationally challenging?

Exponentially many path types



Our solution: feature metrics

Exponentially many path instantiations



Our solution: sampling

Thesis Outline

	Algorithms	Applications
Ch. 2: motivation	**Ch. 2**: Path Ranking Algorithm (Lao & Cohen, MLJ 2010)	**Ch. 3**: knowledge base inference (Lao+, EMNLP 2011)
		Ch. 4: literature recommendation (Lao & Cohen, DILS 2012)
	Ch. 5: efficient RW (Lao & Cohen, KDD 2010)	
	Ch. 6: distributed computing	**Ch. 6**: relation extraction from parsed text (Lao+, EMNLP 2012)
	Ch. 7: more expressive features (submitted)	**Ch. 7**: coordinate term extraction
Ch. 8: future work		

Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Inductive Logic Programming

e.g.

First Order Inductive Learner--FOIL (Quinlan, ECML'93)

High precision Horn clauses

HasFather(a,b) ^ Profession(b,y) → Profession(a,y)

expressive

not robust

not scalable

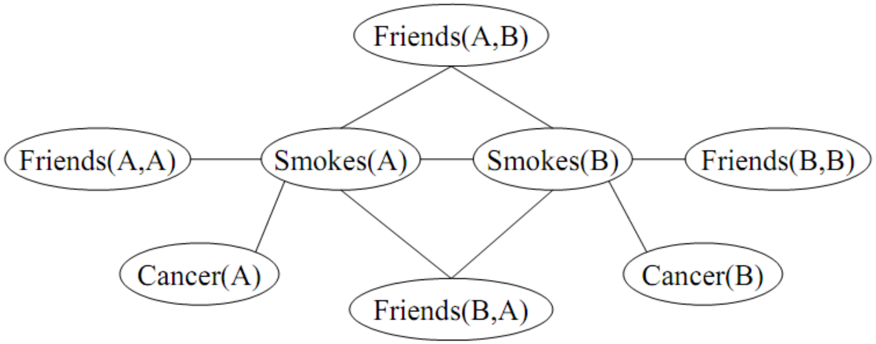
experimental comparison later

Undirected Graphical Models
-- combine logics with GM

e.g.

Markov Logic Networks (Kok & Domingos, ICML'05)

Relational CRFs (Lao+, NIPS'10)



Horn clauses

smokes(A) & Friends(A,B)→smokes(B)

as CRFs features

smokes(A) & Friends(A,B) & !smokes(B)

expressive

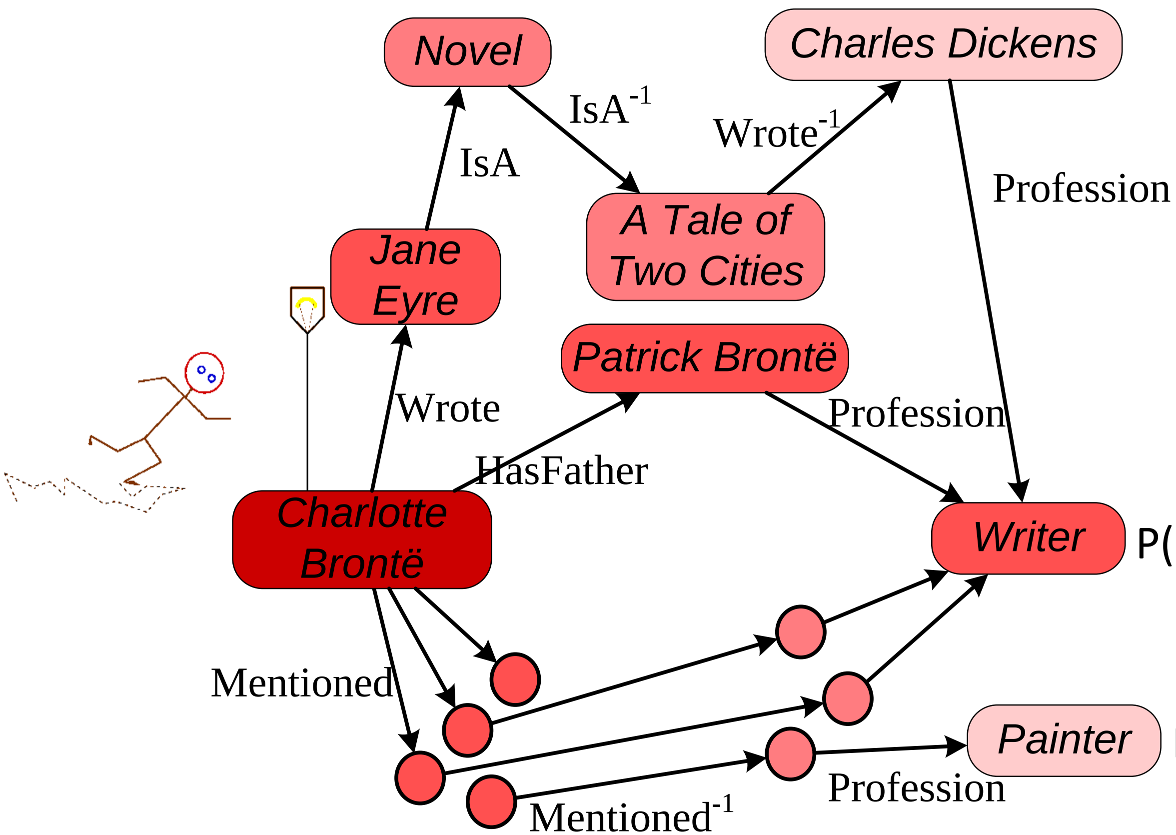
robust

not scalable

Random Walk with Restart

-- ignore logic

e.g. Tong+, ICDM'06



P(Charlotte → Writer)

P(Charlotte → Painter)

not expressive

robust

scalable

experimental comparison later

Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

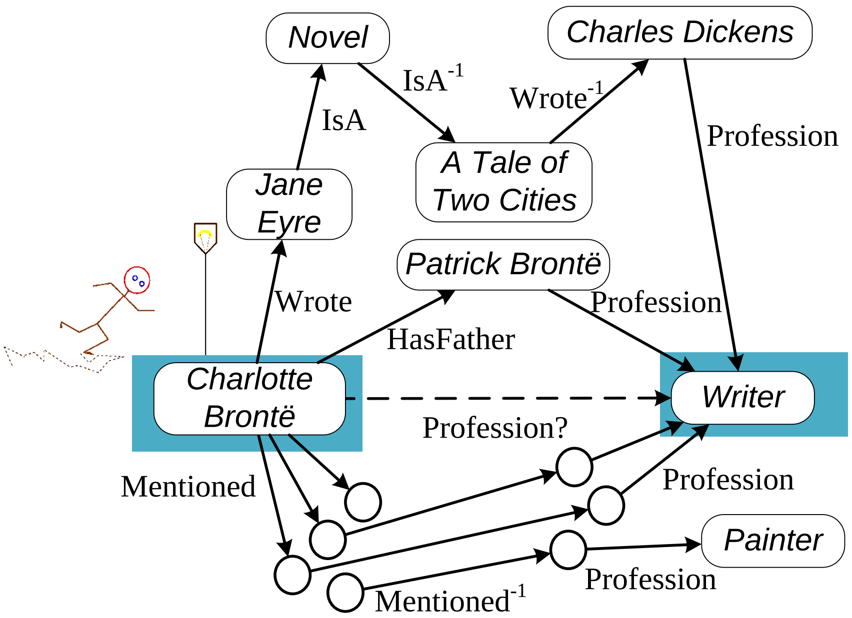
Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Relational Classification

-- combine logics with RWs

e.g. Path Ranking algorithm (Lao & Cohen, MLJ'10)



expressive

robust

scalable

$P(Charlotte \rightarrow Writer; <HasFather,IsA>)$

$P(Charlotte \rightarrow Writer; <Mention,Mention^{-1},IsA>)$

…

$P(Charlotte \rightarrow Painter; <HasFather,IsA>)$

$P(Charlotte \rightarrow Painter; <Mention,Mention^{-1},IsA>)$

…

Contribution

Apply relational learning
at scales not possible before

made possible by

- a family of easy-to-learn features
- fast random walk
- distributed computing

Outline

Motivation	Algorithms	Applications
the problem	Path Ranking Algorithm (PRA)	knowledge base inference
previous work		literature recommendation
idea	efficient RW	
contribution	distributed computing	relation extraction from parsed text
	more expressive features	coordinate term extraction

Path Ranking Algorithm (PRA)

(Lao & Cohen, MLJ 2010)

$$score(s,t) = \sum_{\pi \in \boxed{B}} \boxed{P(s \rightarrow t; \pi)}\boxed{\theta_\pi}$$

a weight

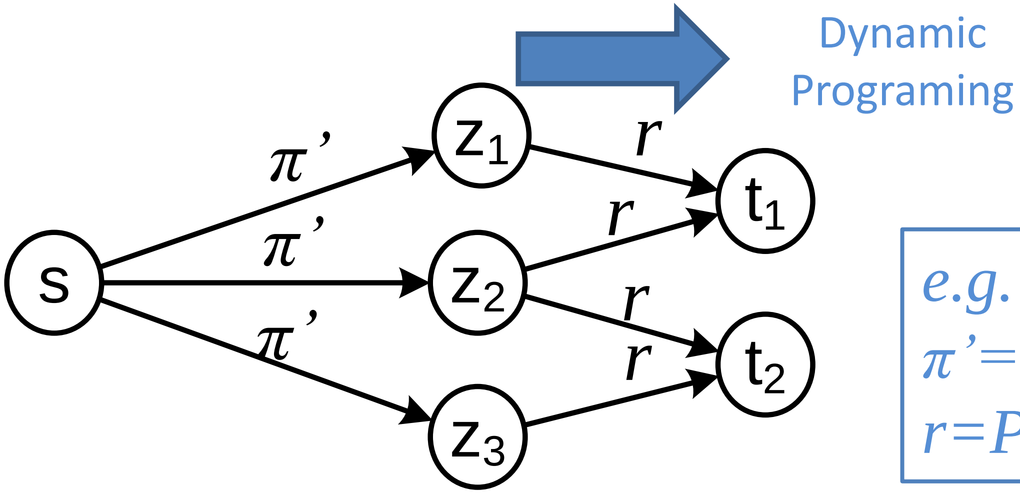
e.g. π=<*Mention*,*Mention*$^{-1}$,*IsA*>

expressive

robust

Random Walk Calculation

$$score(s,t) = \sum_{\pi \in B} \boxed{P(s \to t; \pi)}\, \theta_\pi$$



e.g.
π'=<Mention,Mention^{-1}>
r=Profession

$$P(s \to t; \pi) = \sum_z P(s \to z; \pi')P(z \to t; r)$$

scalable

later about how to do it x100 more efficiently using sampling

Feature Selection with Labeled Data

$$score(s,t) = \sum_{\pi \in \boxed{B}} P(s \rightarrow t; \pi)\theta_\pi$$

given training query set $\{(s_i, G_i)\}$

$$hits(f) = \sum_i I\left[\sum_{j \in G_i} f(s_i, t_j)\right] \geq h$$

$$accuracy(f) = \frac{1}{N}\sum_i \left[\frac{\sum_{j \in G_i} f(s_i, t_j)}{\sum_j f(s_i, t_j)}\right] \geq a$$

I(): the indicator function

N: total number of queries

Estimating θ

$$score(s,t) = \sum_{\pi \in B} P(s \to t; \pi) \boxed{\theta_\pi}$$

for a relation *r*

generate positive and negative node pairs $\{(s_i, t_i)\}$

for each (s_i, t_i) generate (x_i, y_i)

x_i is a vector of RW features of different paths π

y_i is a binary label $r(s_i, t_i)$

estimate θ by L1/L2 regularized (elastic-net) logistic regression

Outline

Motivation	Algorithms	Applications
the problem	Path Ranking Algorithm (PRA)	knowledge base inference
previous work		literature recommendation
idea	efficient RW	
contribution	distributed computing	relation extraction from parsed text
	more expressive features	coordinate term extraction

Knowledge Base Inference

(Lao, Mitchell, Cohen, EMNLP 2010)

Example NELL relations

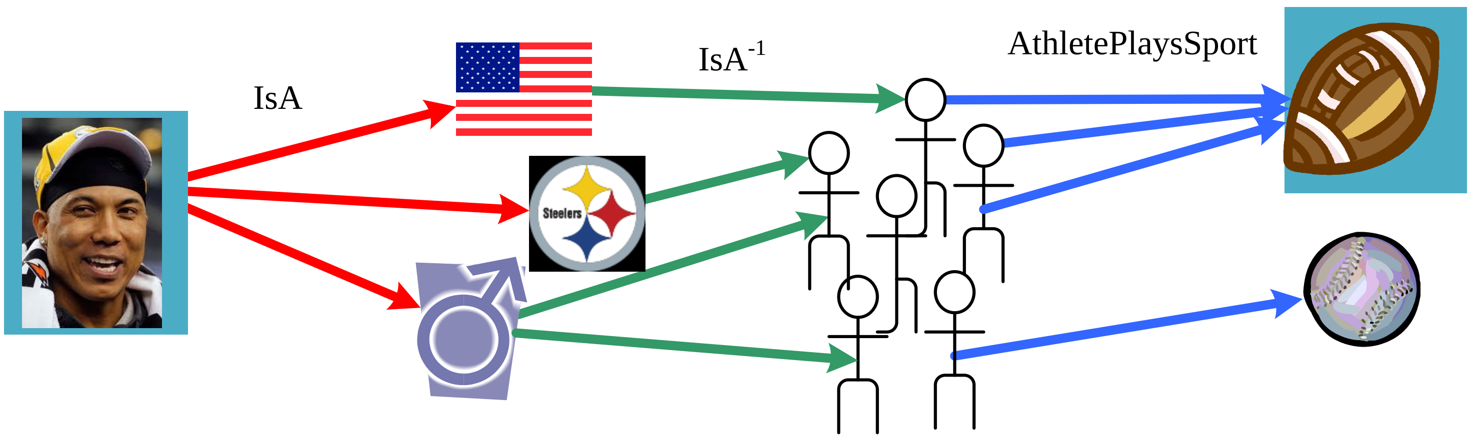
Target Relation	N_Q
athletePlaysForTeam	498
athletePlaysInLeague	892
athletePlaysSport	1119
stadiumLocatedInCity	254
teamHomeStadium	186
teamPlaysInCity	135
teamPlaysInLeague	341
teamPlaysSport	339
teamMember	142
companiesHeadquarteredIn	393
publicationJournalist	68
producedBy	134
competesWith	226
hasOfficeInCity	398
teamWonTrophy	149
worksFor	363

NELL (Never Ending Language Learner) v165

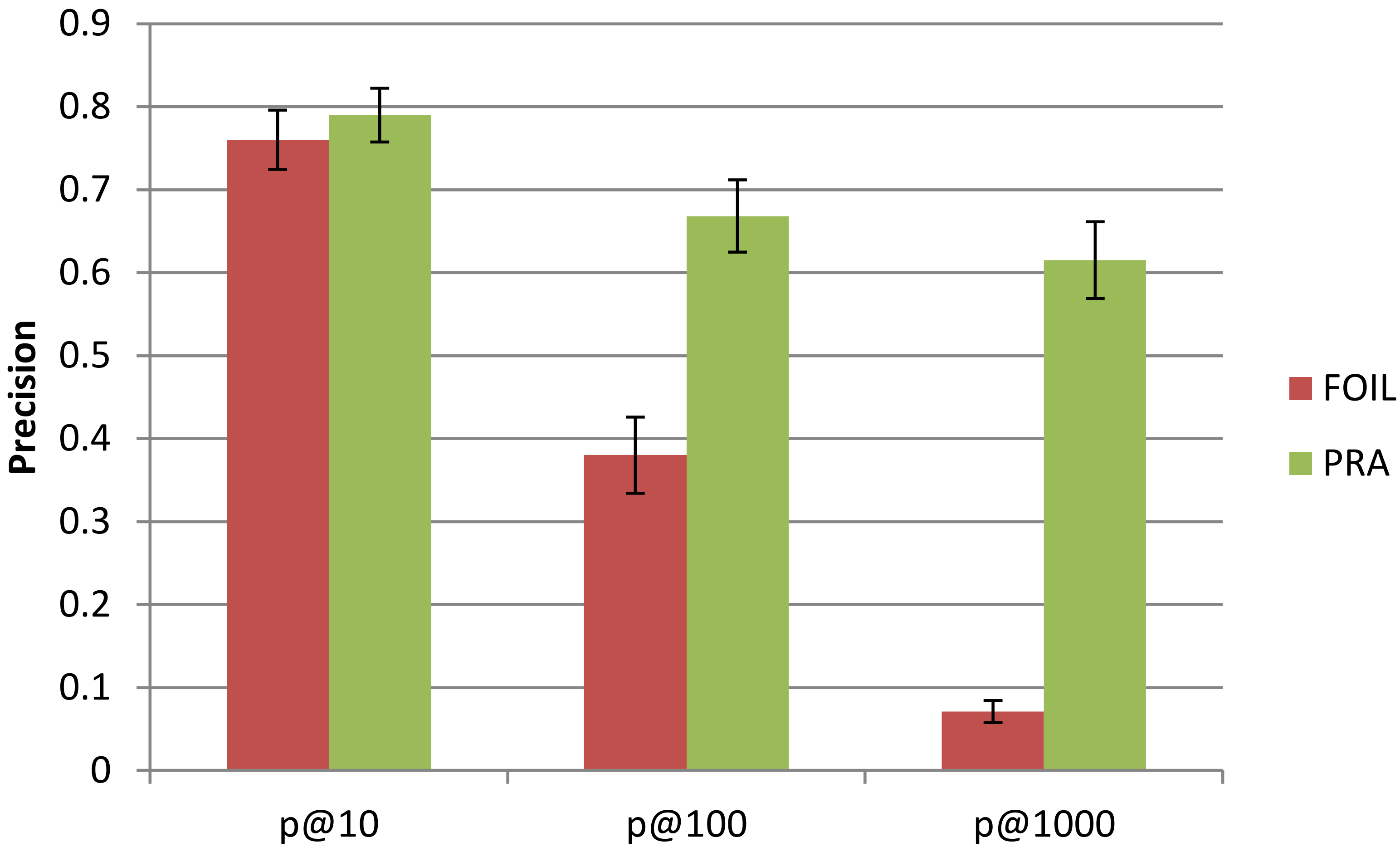
- 353 relations
- 0.7M nodes (concepts)
- 1.7M edges

PRA Uses Broad Coverage Features

AthletePlaysSport(HinesWard, ?)



PRA Has Much Higher Recall and Is Much Faster



Mechanical Turk evaluate new beliefs of 8 functional relations

PRA trains in an hour vs. FOIL trains in a few days

Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications

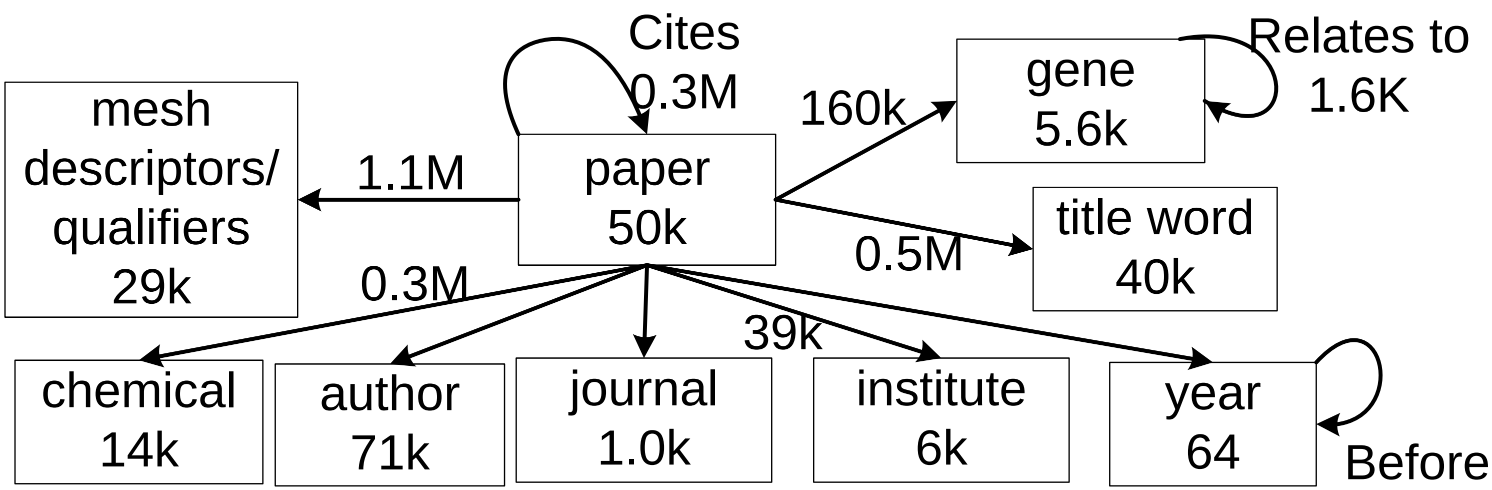
- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Biology Literatures

Databases

Yeast: 0.8M nodes, 3.5M edges

Fly: 0.7M nodes, 16.9M edges



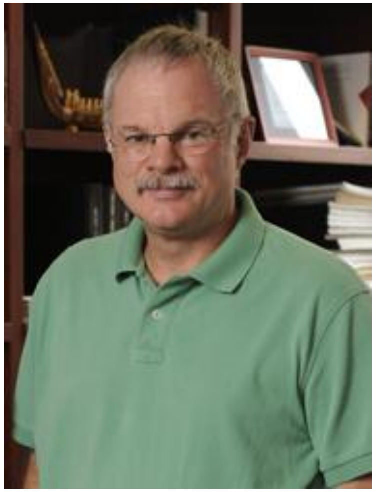
Recommendation Tasks

Literature Recommendation

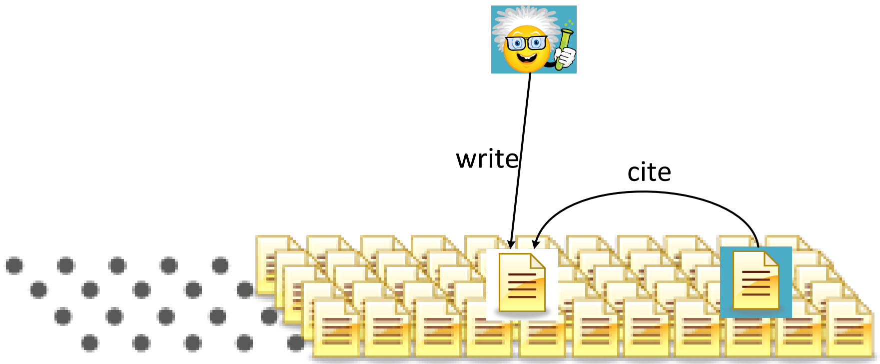
year, author → papers a user is going to read

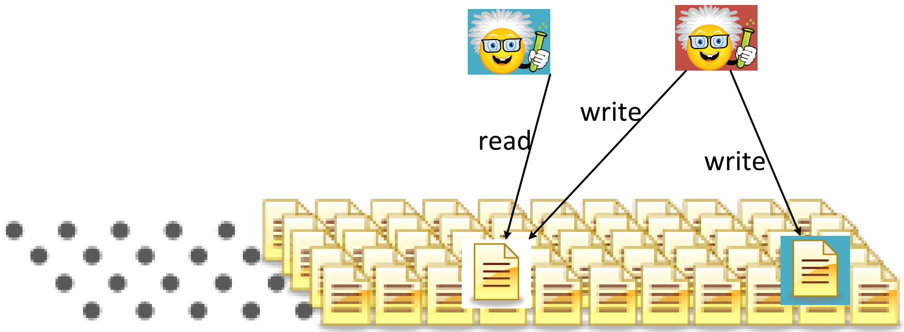
training data --- 1 user over 20 years

(collected from Dr. John Woolford's computer)

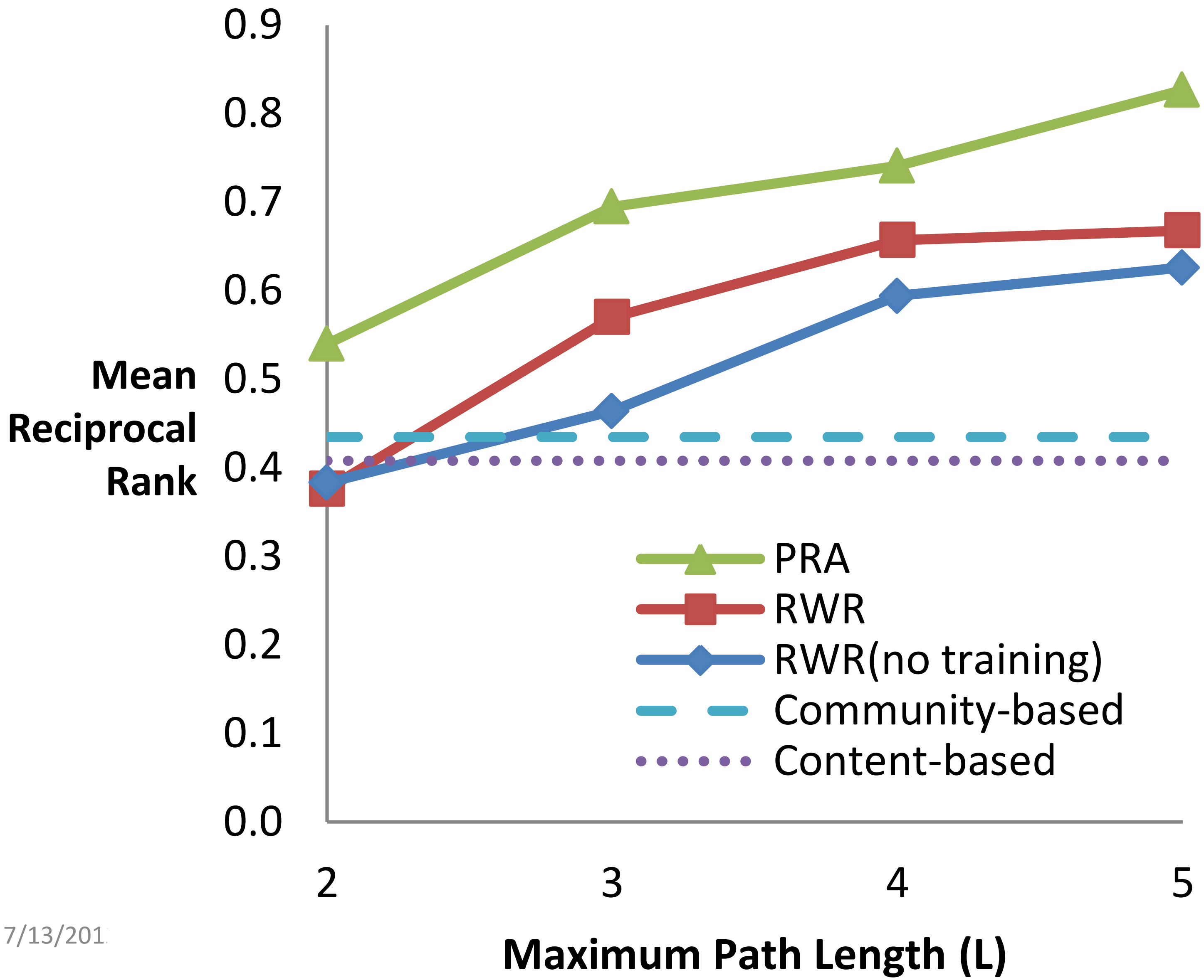


PRA Combines Dozens of Recommendation Strategies





Reading Recommendation

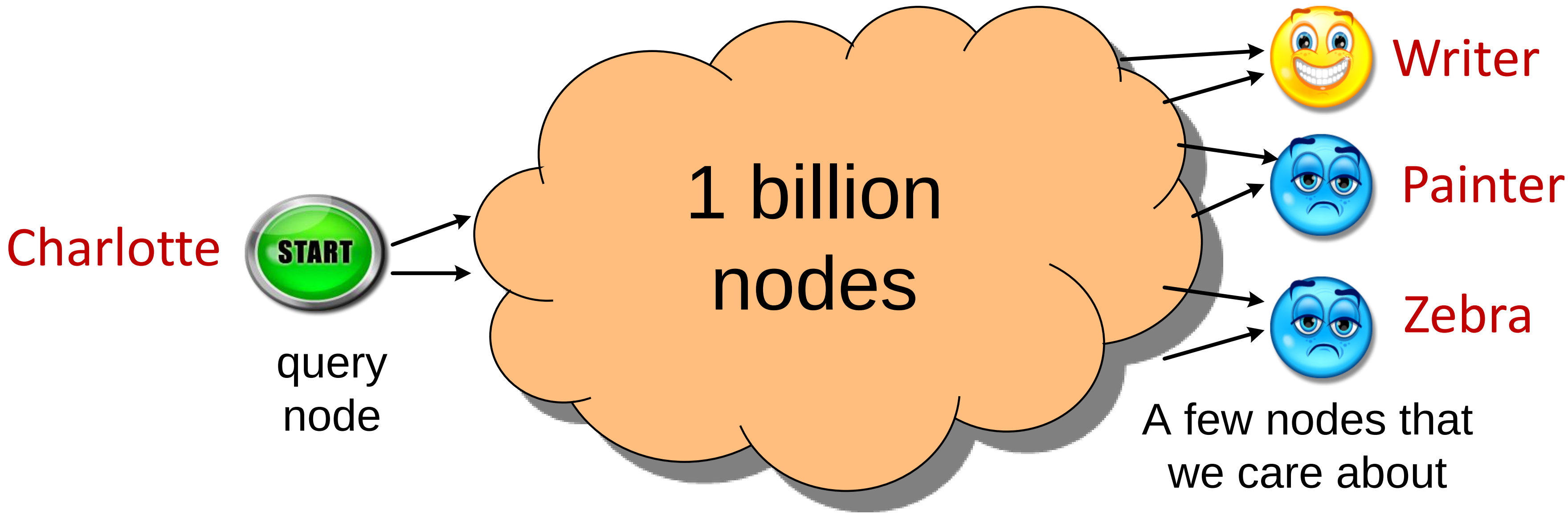


Outline

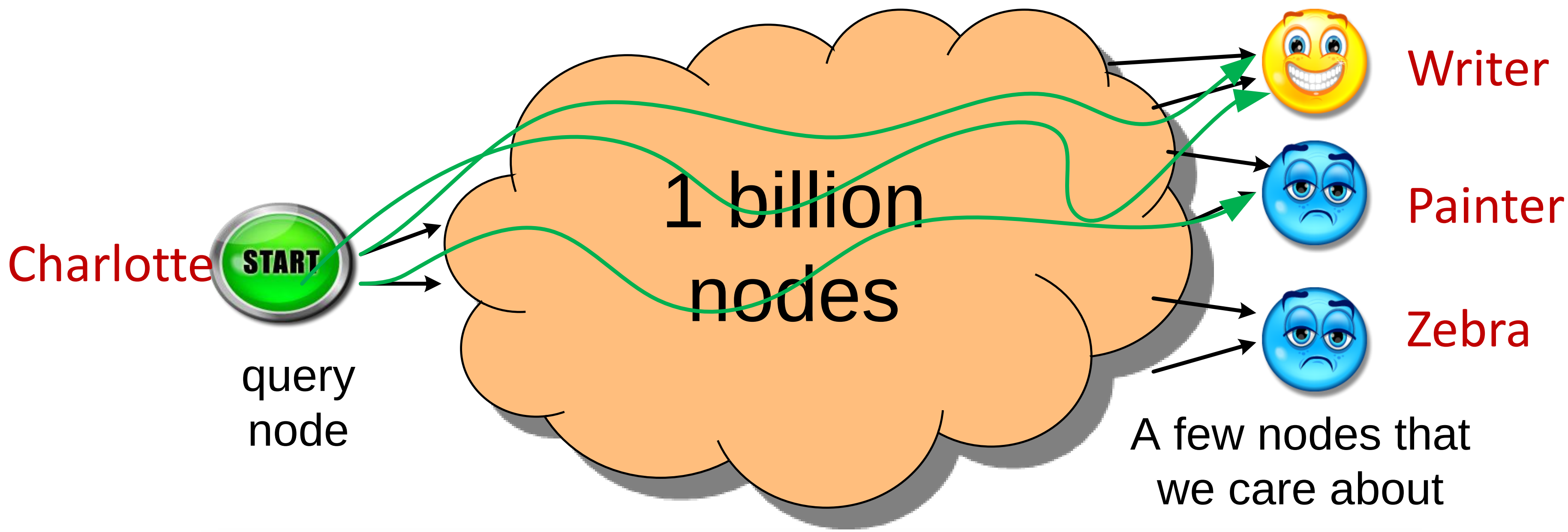
Motivation	Algorithms	Applications
the problem	Path Ranking Algorithm (PRA)	knowledge base inference
previous work		literature recommendation
idea	efficient RW	
contribution	distributed computing	relation extraction from parsed text
	more expressive features	coordinate term extraction

Efficient Random Walks

(Lao & Cohen, KDD 2010)



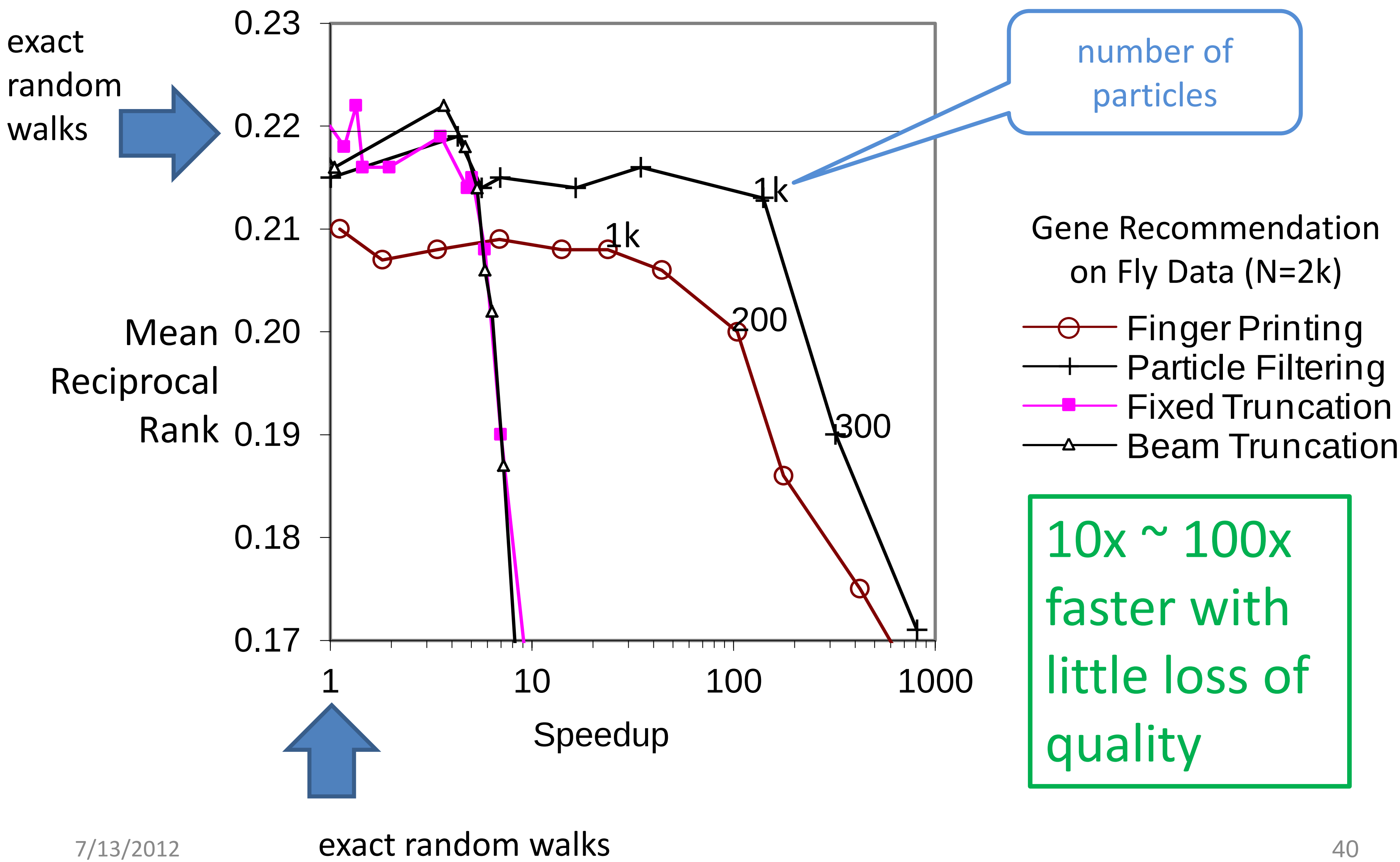
Exact calculation of random walks results in non-zero probabilities for many internal nodes



Idea: a few random walkers (particles) are enough to distinguish good target nodes from bad ones

details

Compare Speedup Approaches



10x ~ 100x faster with little loss of quality

Outline

Motivation

- the problem
- previous work
- idea
- contribution

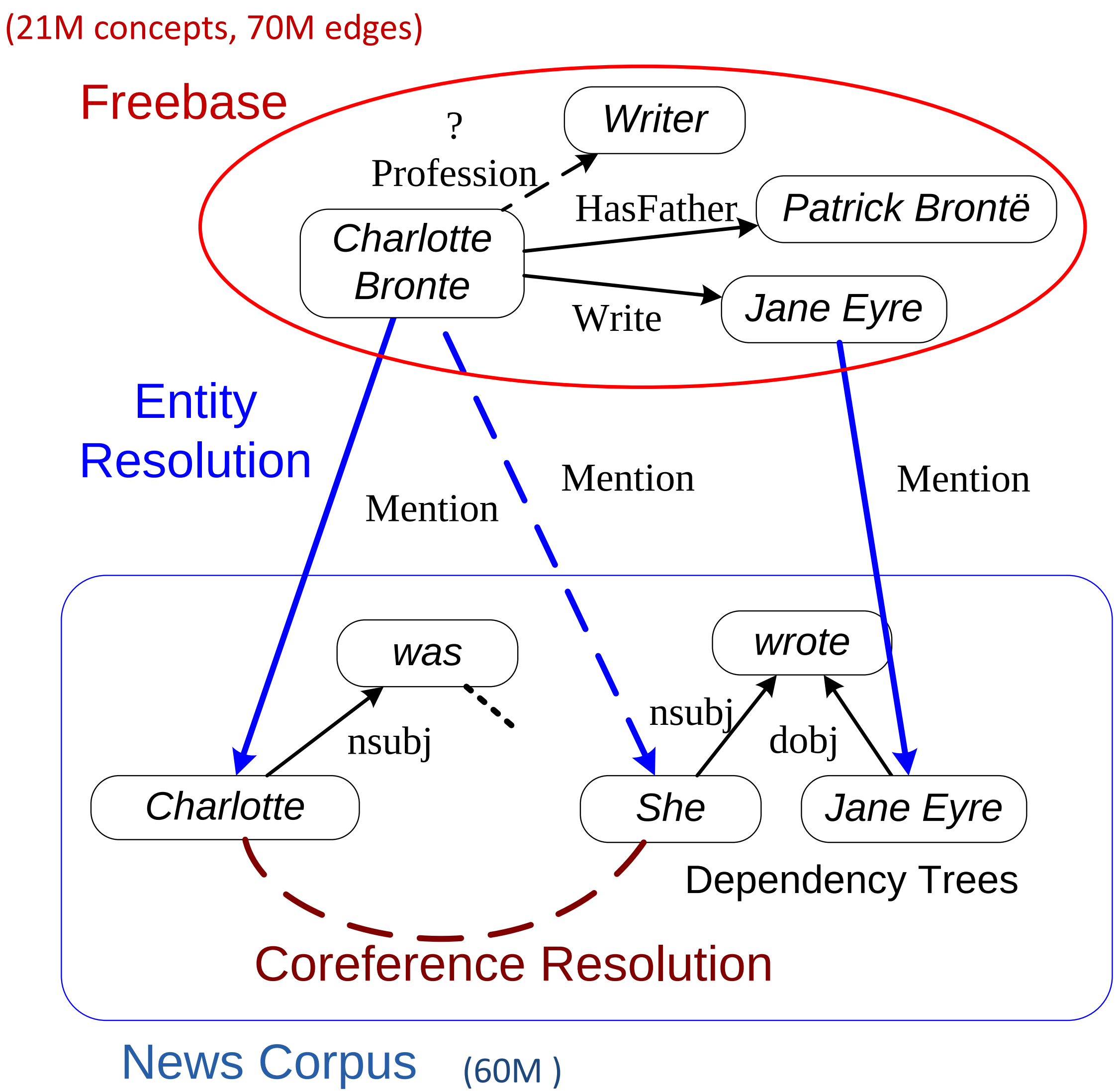
Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Relation Extraction



Can PRA scale?

Can PRA learn syntactic-semantic rules?

Distributed Computing

Large number of queries

e.g. 0.3M/2M persons have known profession

Solution: map/reduce to explore path, generate training samples, calculate gradient, and do predictions for each query

Large text graph

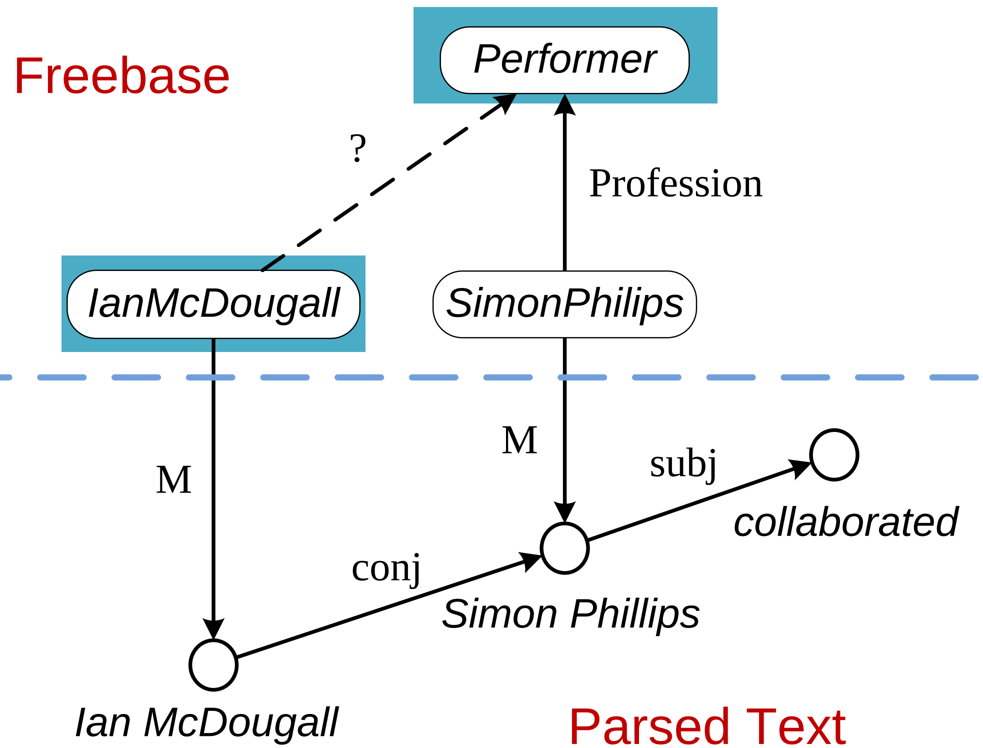
e.g. 60M documents

Solution: each node keeps the Freebase graph in memory

relevant sentences are loaded/unloaded for each query

Combine Syntax with Semantics

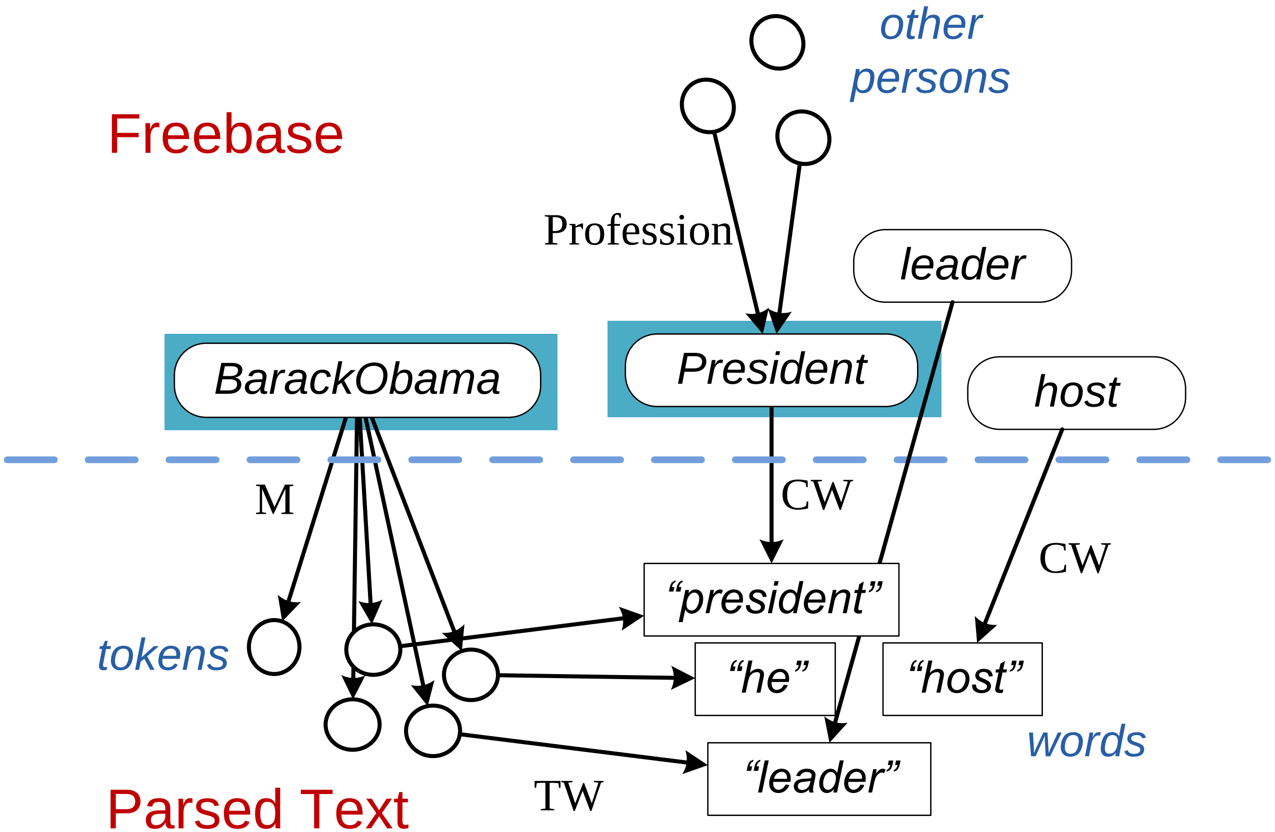
<M, conj, M^{-1}, Profession>



"McDougall and Simon Phillips collaborated..."

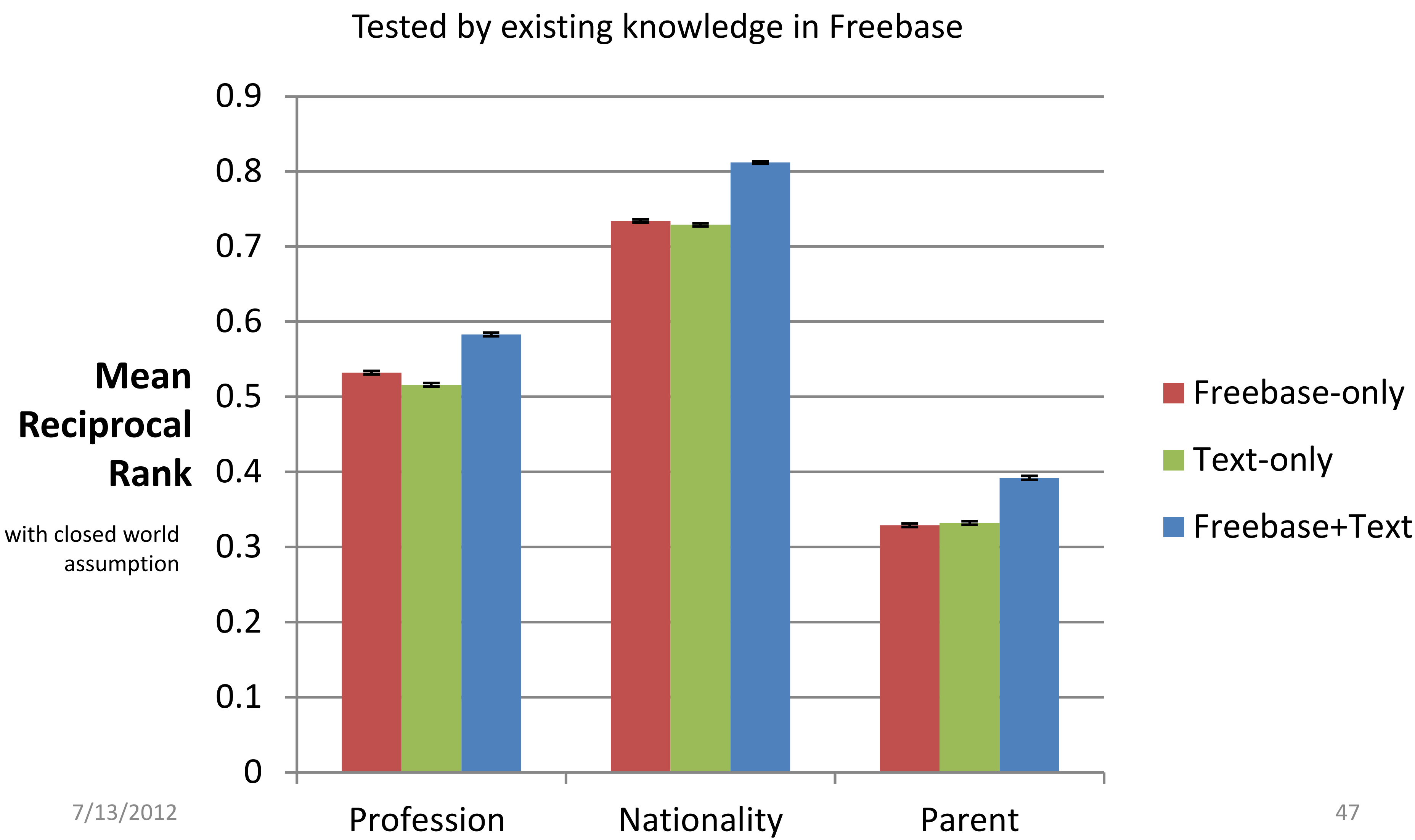
Combine Text with Semantics

<M, WORD, CW^{-1}, $profession^{-1}$, profession>

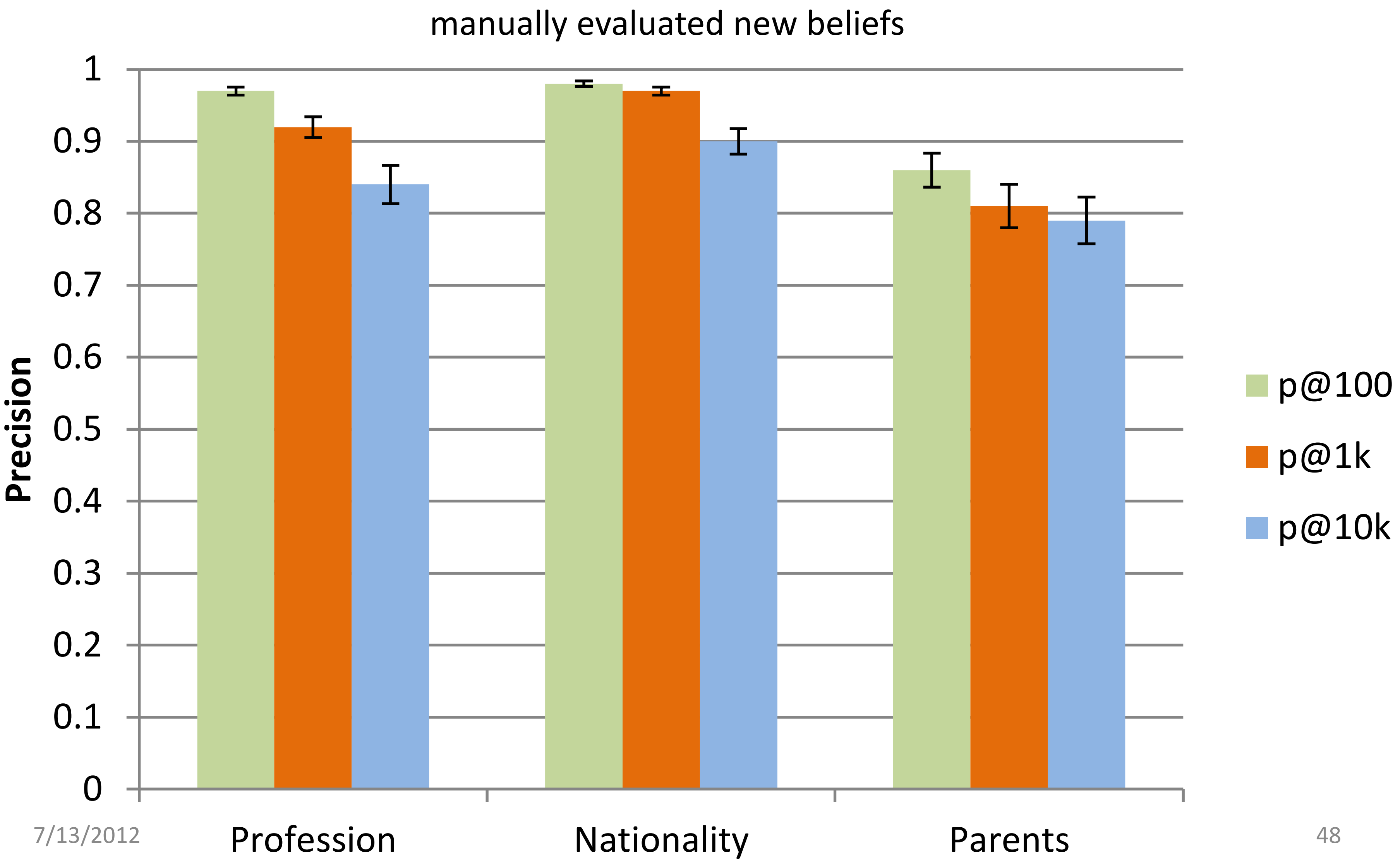


e.g. "The president said …"

Text and KB Work Better Together



Highly Accurate New Beliefs



Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

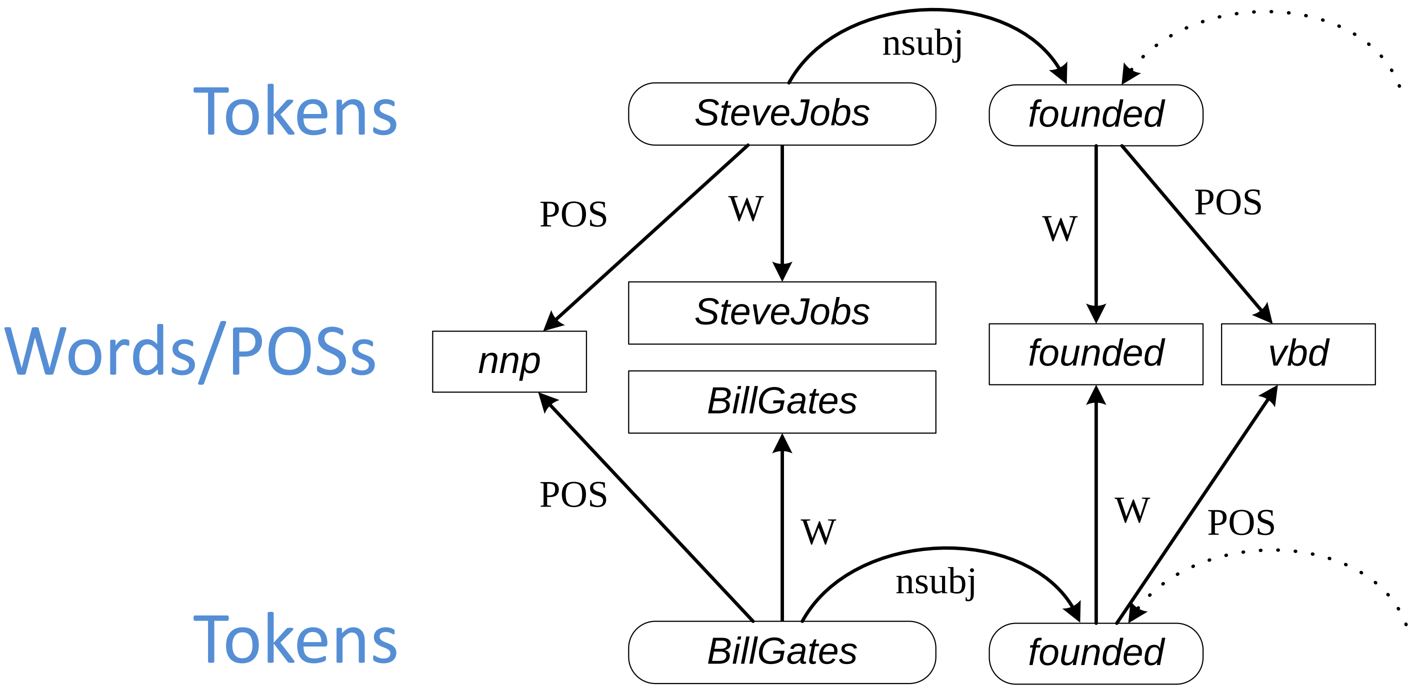
- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Coordinate Term Extraction Task

(Minkov & Cohen, ECML 2010)



W: word
POS: part of speech
nnp: singular proper nouns
vbd: verb, past tense
nsubj: subject of a verb

parsed MUC-6 corpus

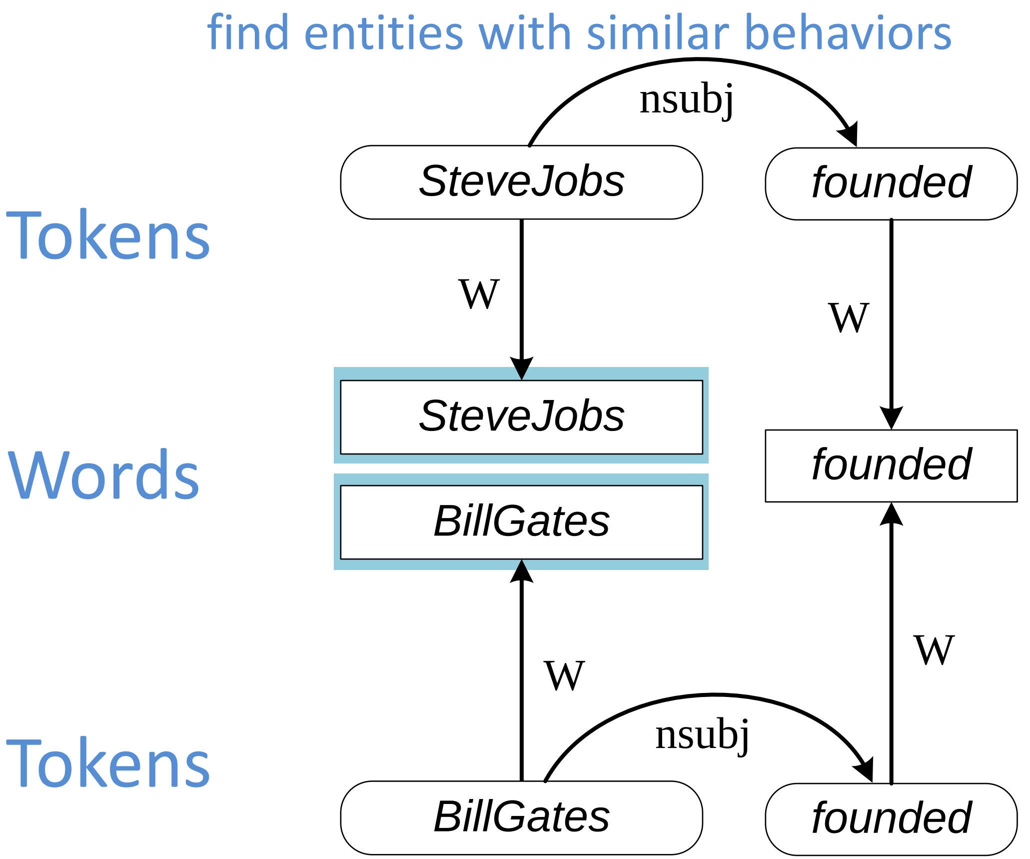
- 153k nodes, 748K edges

30 queries

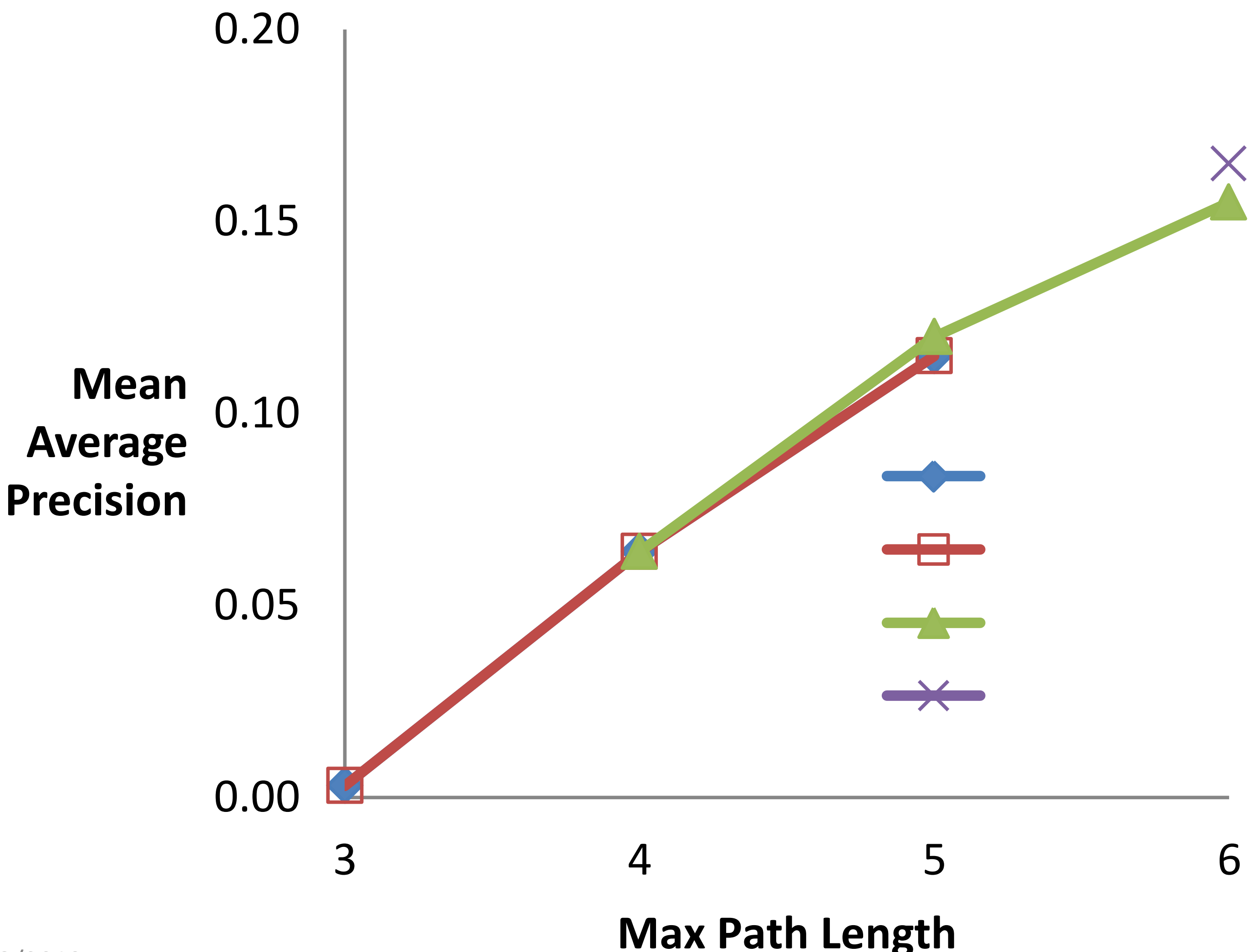
- given 4 person names as seeds, find other persons

Good Paths Are Quite Long

<W^{-1},nsubj,W,W^{-1},$nsubj^{-1}$,W>

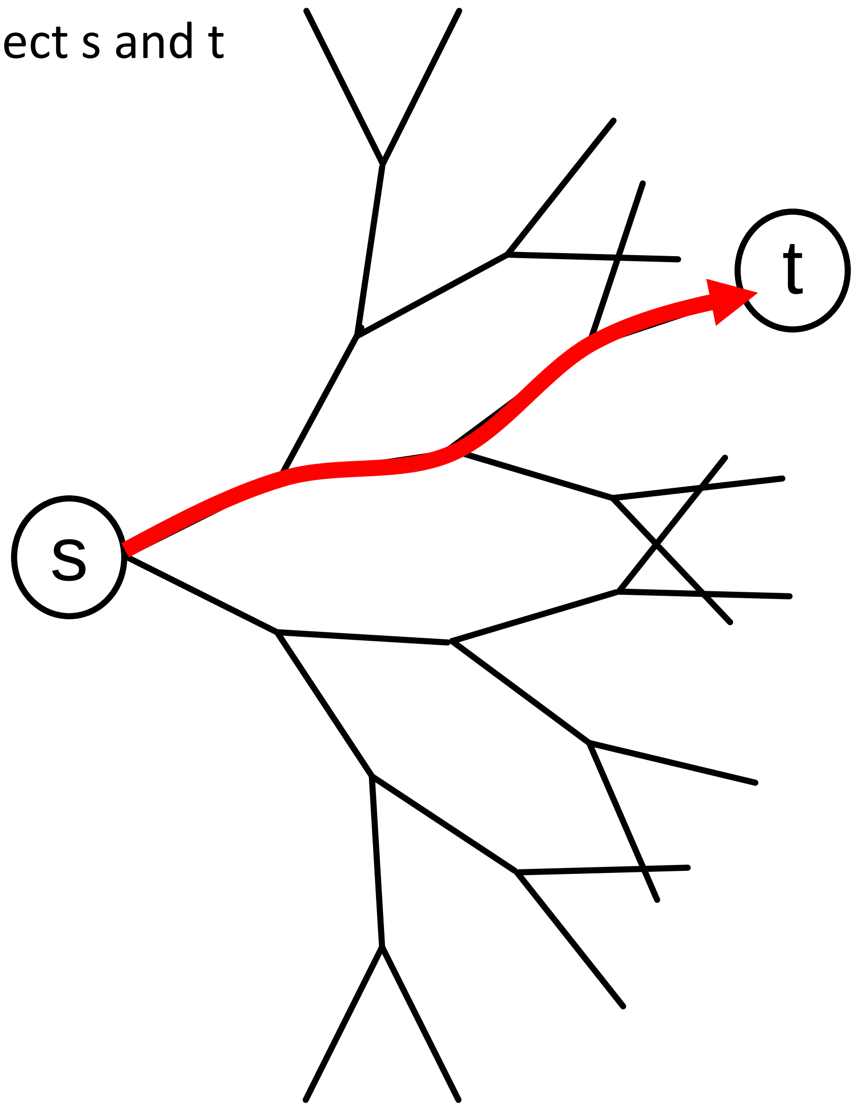


Good Paths Are Quite Long

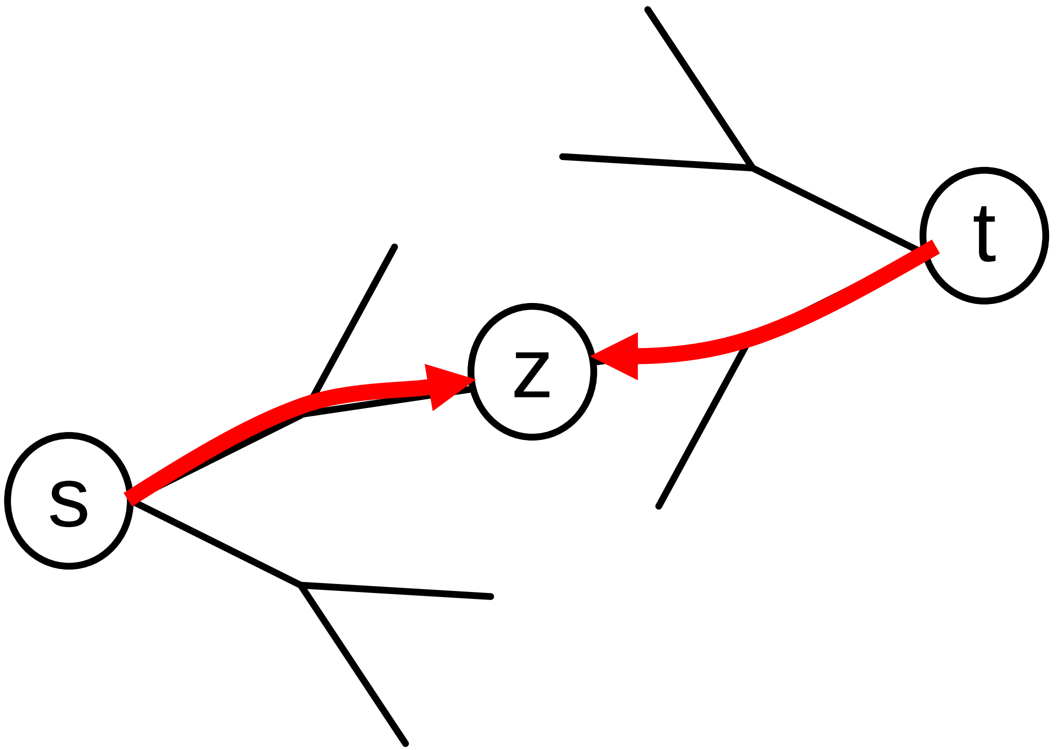


Forward Search Is Wasteful

Find paths that connect s and t



Bidirectional Search Is More Efficient

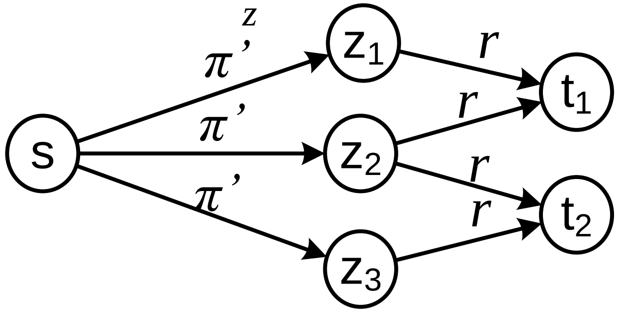


challenge is to calculate $P(s \rightarrow t; \pi)$

Forward vs. Backward RWs

Forward

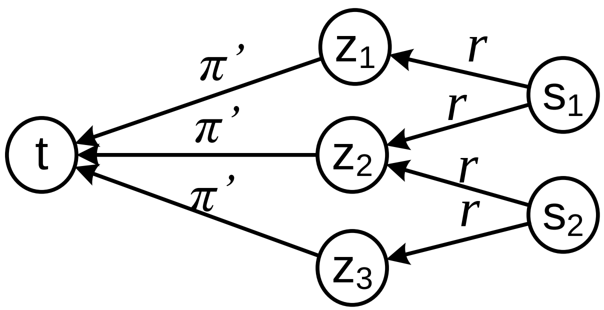
$$P(s \rightarrow t;\pi) = \sum_z P(s \rightarrow z;\pi')P(z \rightarrow t;r)$$



evaluate $P(s\rightarrow t;\pi)$
for many t

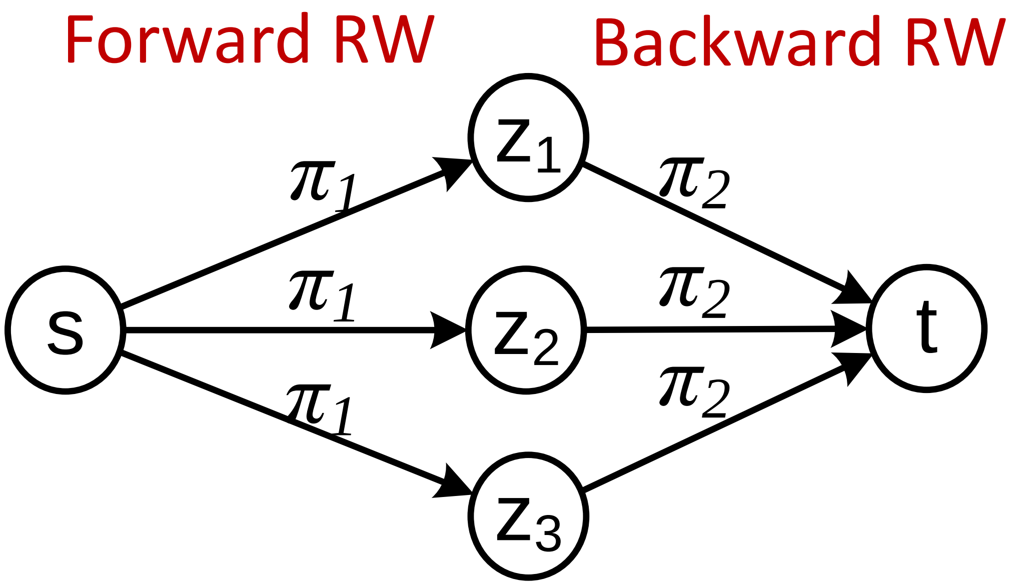
Backward

$$P(t \leftarrow s;\pi) = \sum_z P(t \leftarrow z;\pi')P(z \leftarrow s;r)$$



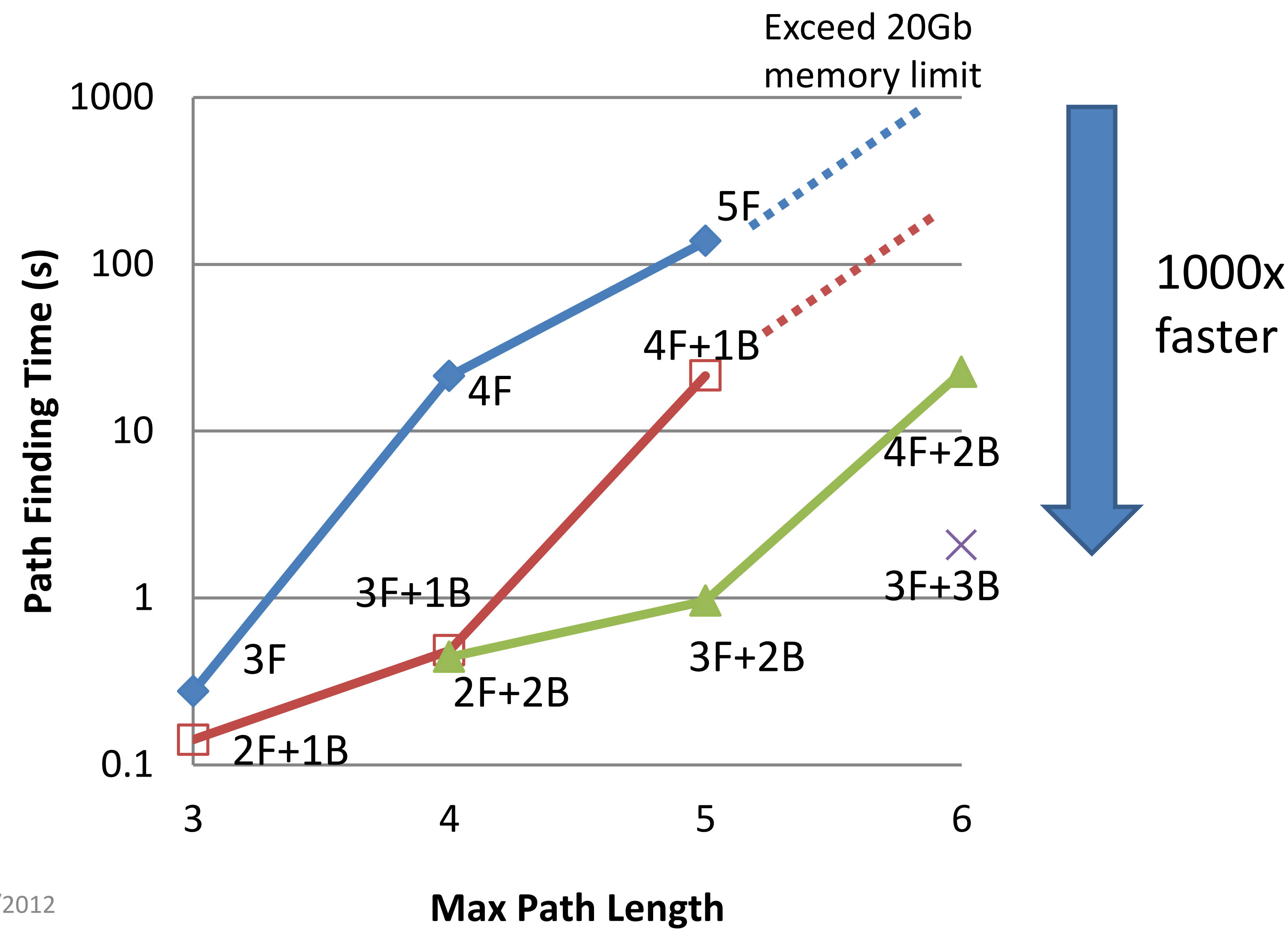
evaluate $P(s\rightarrow t;\pi)$
for many s

Bidirectional Search with RW



$$P(s \rightarrow t; \pi_1 \pi_2) = \sum_z P(s \rightarrow z; \pi_1) P(t \leftarrow z; \pi_2)$$

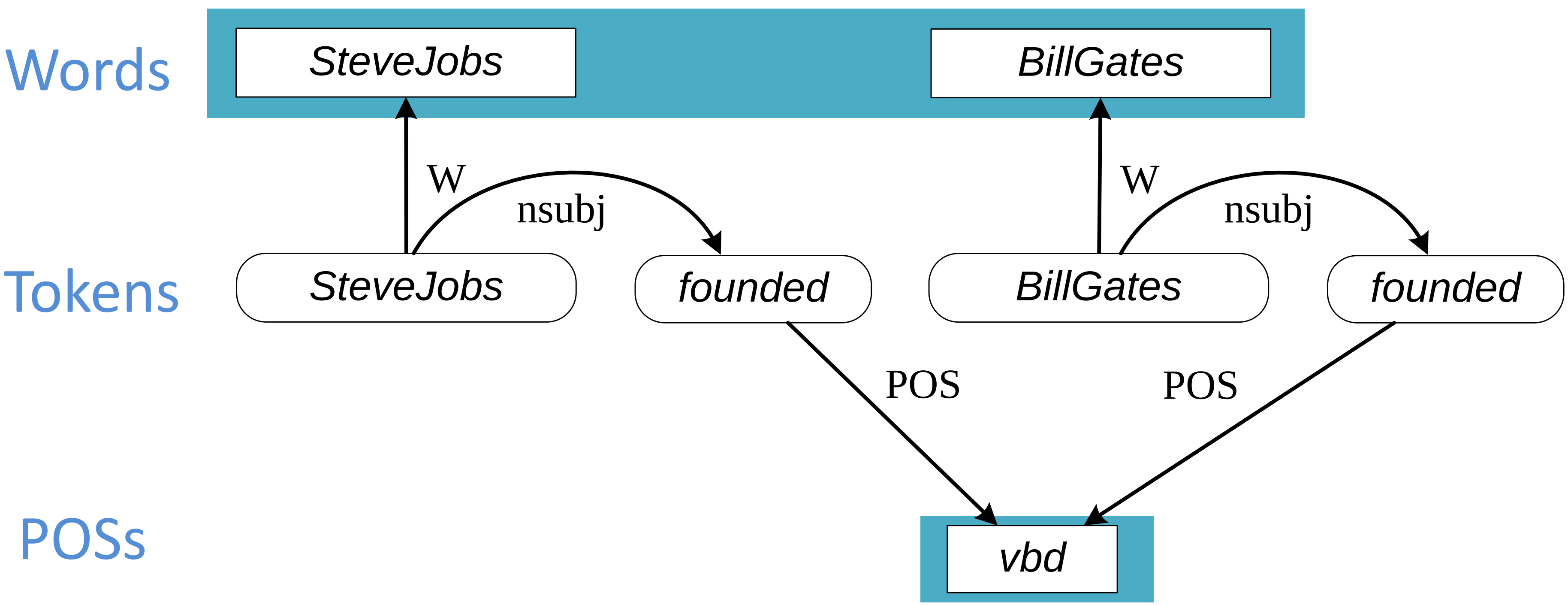
Bidirectional Search Is Much Faster



Need for Lexicalized Paths

Task: find person entities

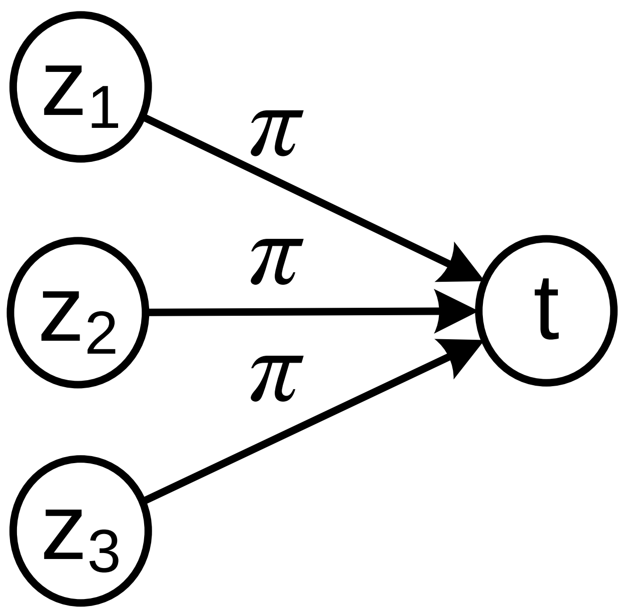
$P(vbd \rightarrow t \mid \langle POS^{-1}, nsubj^{-1}, W \rangle)$



Evaluate Lexicalized Paths

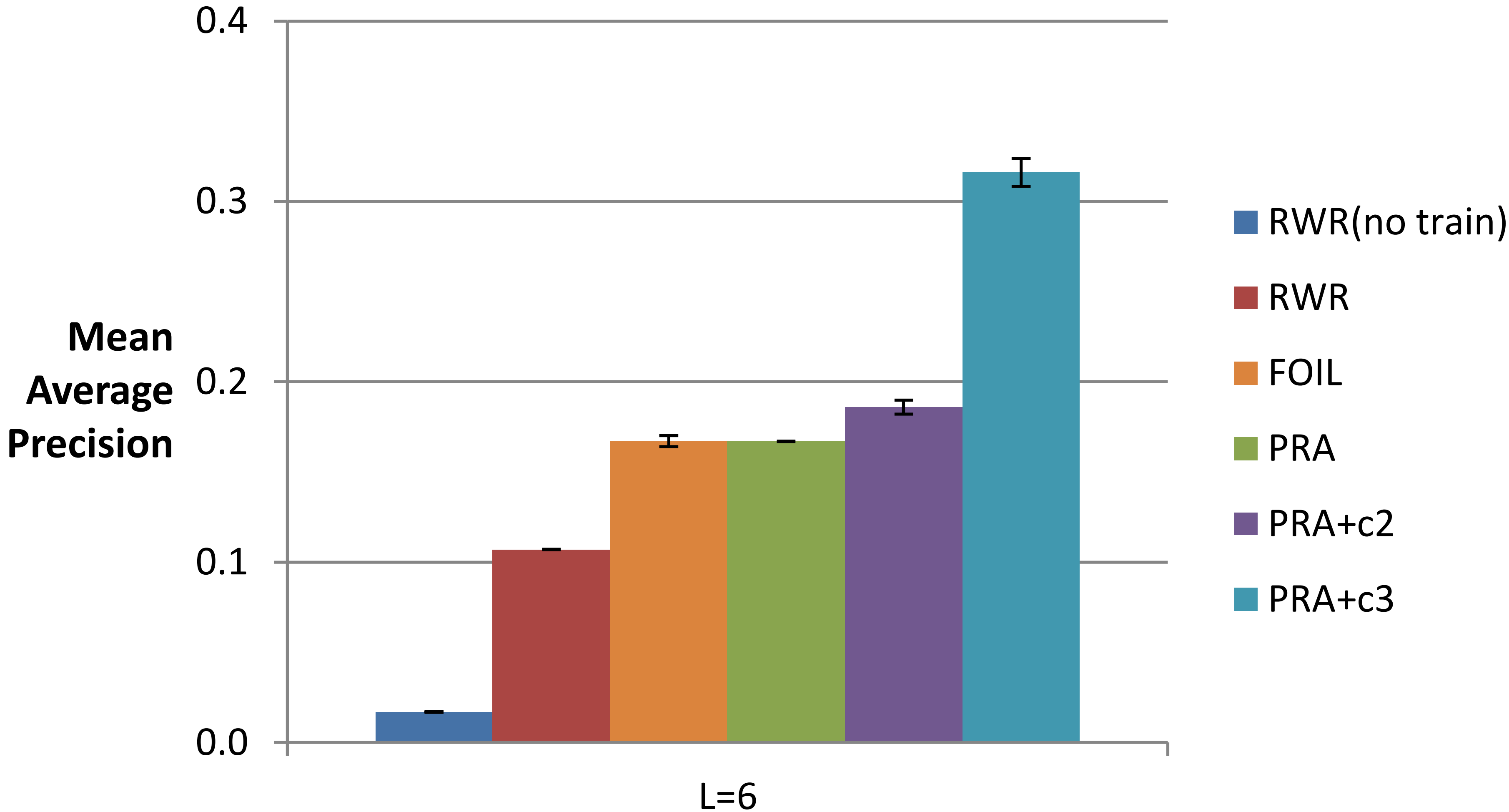
Given an example (s_i, t_i)

calculate $P(z \longrightarrow t_i; \pi)$ for many z



Person Name Extraction

~1000 correct answers



Outline

Motivation

- the problem
- previous work
- idea
- contribution

Algorithms

- Path Ranking Algorithm (PRA)
- efficient RW
- distributed computing
- more expressive features

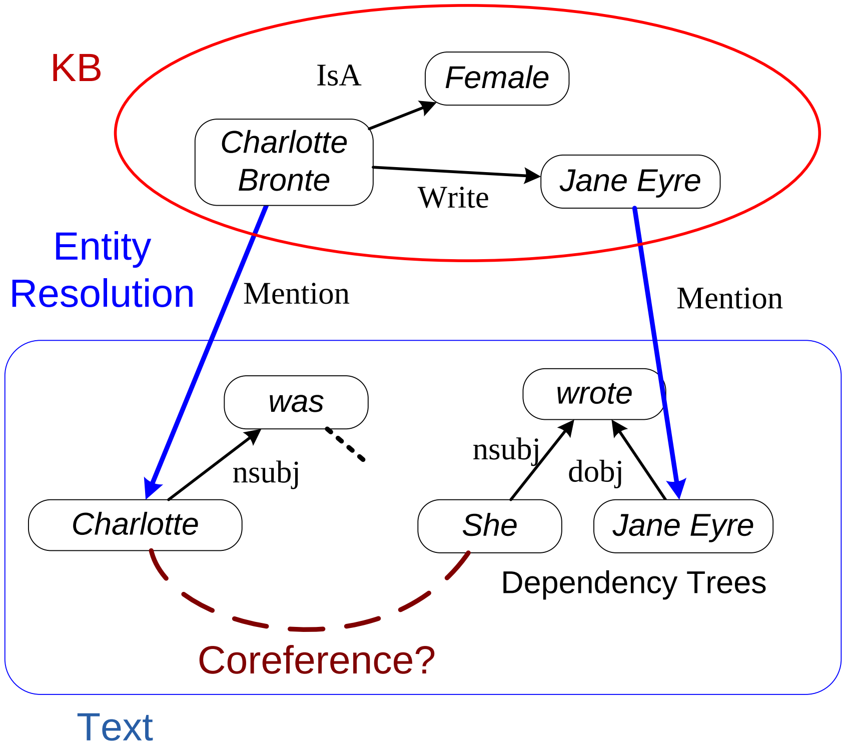
Applications

- knowledge base inference
- literature recommendation
- relation extraction from parsed text
- coordinate term extraction

Future Work

Apply knowledge to NLP/IE/IR/CV tasks

$$\arg\max_{decision} P(decision \mid context, KB)$$

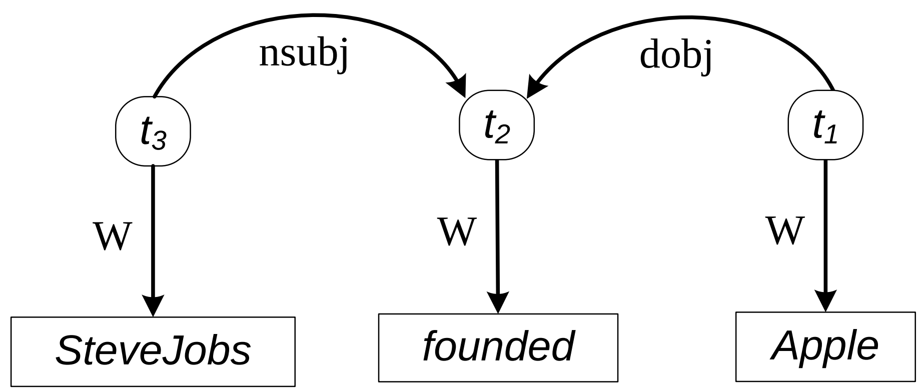


Future Work

Conjunctions of Paths

rules can have tree structures

with source/constant/target nodes as leafs

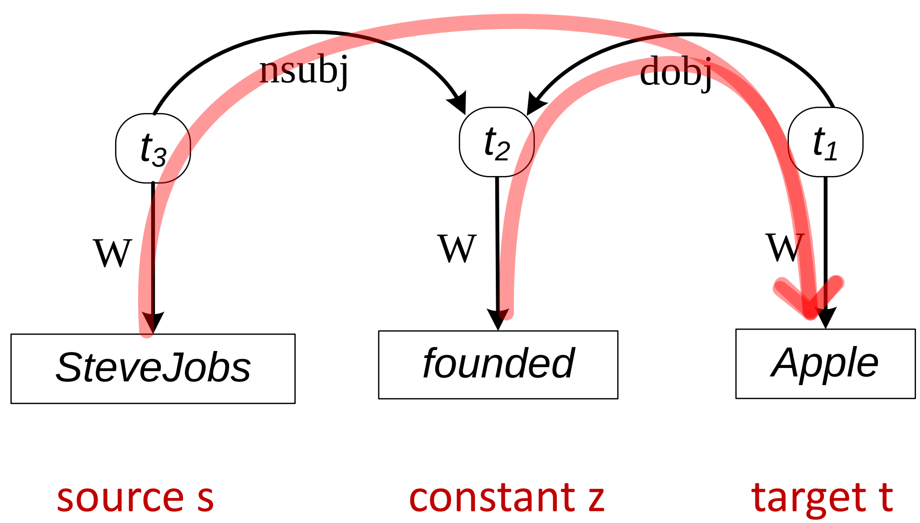


source s　　constant z　　target t

Future Work

Conjunctions of Paths

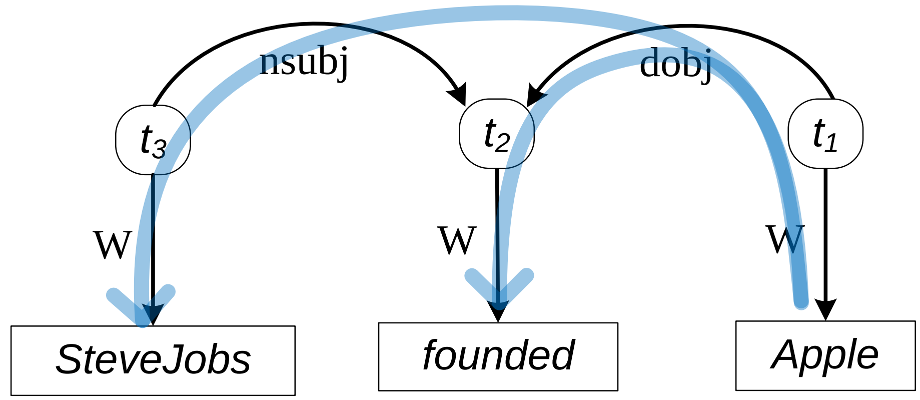
forward PCRW with multiple walkers



Future Work

Conjunctions of Paths

backward PCRW with multiple walkers



source s
constant z
target t

Contribution

Apply relational learning at scales not possible before.

Leads to new applications!

Made possible by

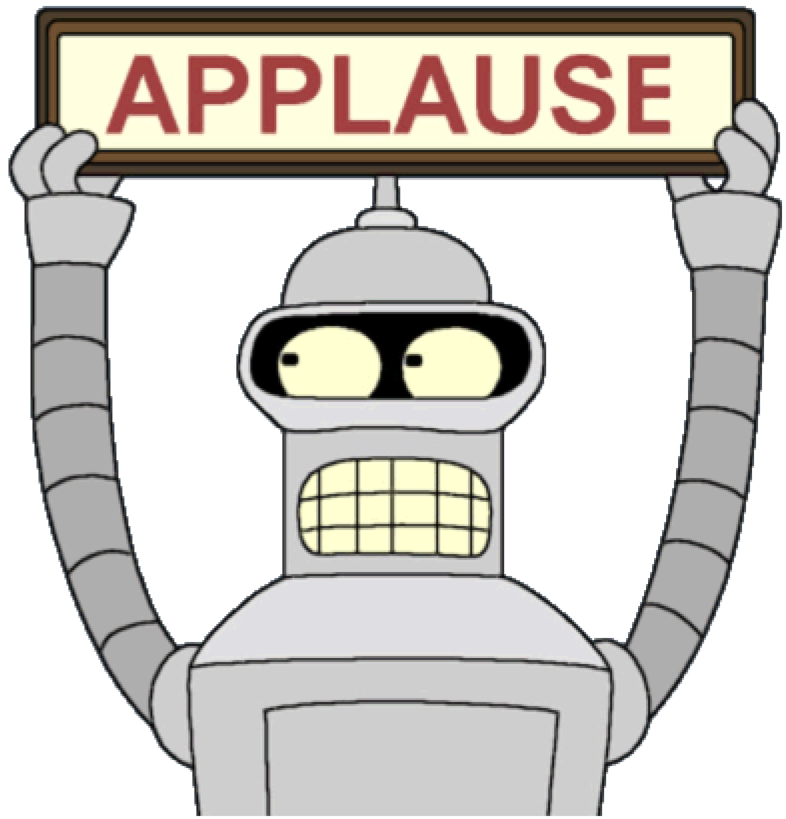
- a family of easy-to-learn features (3 types)
- fast random walk (sampling)
- distributed computing

other work I did at CMU

Relational CRFs (Lao+, NIPS’10)

Question answering (Lao+, NTCIR’08)

Utility based retrieval evaluation (Yang+, SIGIR’07)


APPLAUSE

Future Work

KB extension

new relation types, new concepts

Unsupervised

$$\arg\max_{\Delta KB} P(corpus \mid KB + \Delta KB)$$

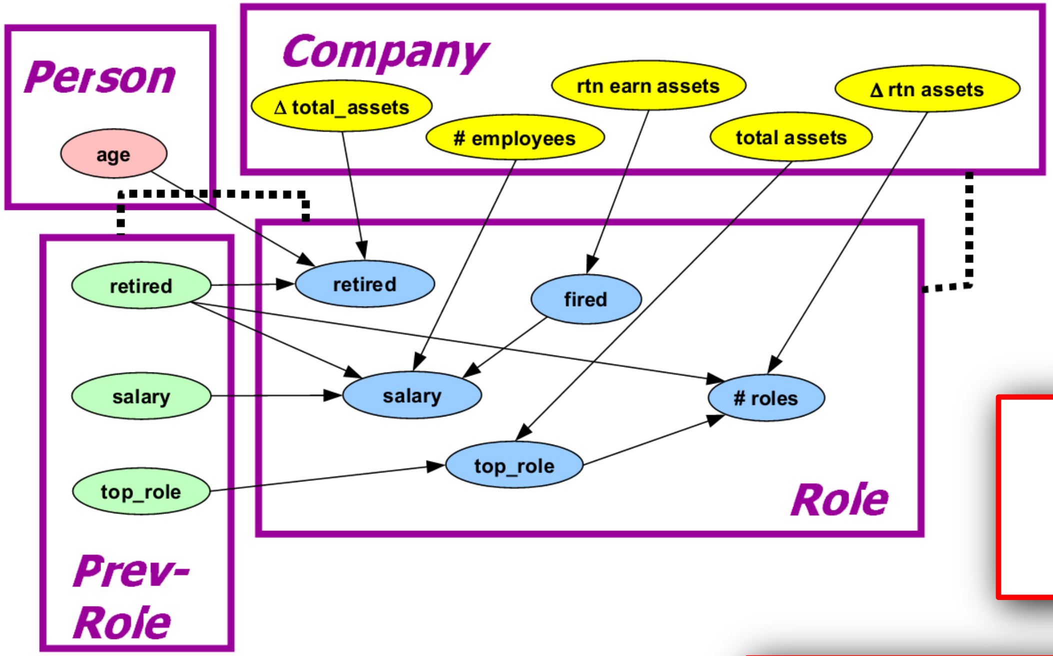
Supervised

$$\arg\max_{\Delta KB} P(decisions \mid contexts, KB + \Delta KB)$$

Directed Graphical Models

e.g.

Probabilistic Relational Models (Getoor+, ICML'01)



model structure restricted to DAG

cannot express features corresponding to chains

Coverage of top k triples

Profession Triples	Unique Persons
1k	970
10k	8,726
100k	79,885

Repeatedly Combine Forward and Backward RWs

<W^{-1},conj_and,W,

Forward search

W^{-1},conj_and,W,

W^{-1},conj_and,W>

+Backward Search

+Backward Search

Summary of PRA

Stage	Computation
Path Discovery	given $\{(s_i, G_i)\}$, find $\{f$; $acc(f)>=a$, $hits(f)>=h\}$
Generate Training Samples	generate $\{(s_i, t_i)\}$ and $\{(x_i, y_i)\}$
Logistic Regression Training	$\theta = \arg\max_{\theta} \left[\sum_i l_i(\theta) - \lambda_1 \Vert \theta \Vert_1 - \lambda_2 \Vert \theta \Vert_2^2 \right]$
Prediction	apply model to nodes *s* in *domain*(*r*)

Need for Lexicalized Paths

Task=AthletePlaysInLeague

$$P(mlb \rightarrow t; \phi)$$

Bias toward MLB

$$P\left(BostonBraves \rightarrow t; \left\langle \begin{matrix} AthletePlaysForTeam^{-1}, \\ AthletePlaysInLeagure \end{matrix} \right\rangle \right)$$

A prior over the leagues participated by Boston Braves university athletes

Need for Lexicalized Paths

Task=CompetesWith

$$P(google \to t;\phi)$$

Bias toward Google

$$P\left(Google \to t;\langle CompetesWith, CompetesWith\rangle\right)$$

Companies around Google

Knowledge Base Inference