Uncertain, Structured, Intensional Data

Pierre Senellart

Télécom ParisTech & National University of Singapore

16 March 2016, *École normale supérieure*

Brief CV

Brief CV

2000–2005

Élève ENS, promotion MPI 2000
(creator and maintainer of the Web site of A-Ulm)

Brief CV

2000–2005 

Élève ENS, promotion MPI 2000
(creator and maintainer of the Web site of A-Ulm)

2003–2004 
R&D Engineer at Systran
(machine translation)

Brief CV

2000–2005 

Élève ENS, promotion MPI 2000
(creator and maintainer of the Web site of A-Ulm)

2003–2004 
R&D Engineer at Systran
(machine translation)

2004–2007 

PhD student at Inria Saclay,
under Serge Abiteboul

Brief CV

2000–2005 

Élève ENS, promotion MPI 2000
(creator and maintainer of the Web site of A-Ulm)

2003–2004 
R&D Engineer at Systran
(machine translation)

2004–2007 

PhD student at Inria Saclay,
under Serge Abiteboul

2008–2008 

Post-Doc
at Max-Planck Institut für Informatik

Brief CV

2000–2005 

Élève ENS, promotion MPI 2000
(creator and maintainer of the Web site of A-Ulm)

2003–2004 
R&D Engineer at Systran
(machine translation)

2004–2007 

PhD student at Inria Saclay,
under Serge Abiteboul

2008–2008 

Post-Doc
at Max-Planck Institut für Informatik

2008–2013 

Maître de conférences
at Télécom ParisTech

Brief CV

2000–2005	Élève ENS, promotion MPI 2000 (creator and maintainer of the Web site of A-Ulm)
2003–2004	R&D Engineer at Systran (machine translation)
2004–2007	PhD student at Inria Saclay, under Serge Abiteboul
2008–2008	Post-Doc at Max-Planck Institut für Informatik
2008–2013	Maître de conférences at Télécom ParisTech
2012–2013	Sabbatical at Hong Kong University

Brief CV

2000–2005	Élève ENS, promotion MPI 2000 (creator and maintainer of the Web site of A-Ulm)
2003–2004	R&D Engineer at Systran (machine translation)
2004–2007	PhD student at Inria Saclay, under Serge Abiteboul
2008–2008	Post-Doc at Max-Planck Institut für Informatik
2008–2013	Maître de conférences at Télécom ParisTech
2012–2013	Sabbatical at Hong Kong University
2013–. . .	Professor at Télécom ParisTech

Brief CV

2000–2005		Élève ENS, promotion MPI 2000 (creator and maintainer of the Web site of A-Ulm)
2003–2004		R&D Engineer at Systran (machine translation)
2004–2007		PhD student at Inria Saclay, under Serge Abiteboul
2008–2008		Post-Doc at Max-Planck Institut für Informatik
2008–2013		Maître de conférences at Télécom ParisTech
2012–2013		Sabbatical at Hong Kong University
2013–. . .		Professor at Télécom ParisTech
2014–. . .		Senior Research Fellow at National University of Singapore

Research Interests

- Foundational and practical aspects of Web data management

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management
- On the systems side:

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management
- On the systems side:
  - Graph mining, social network analysis

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management
- On the systems side:
  - Graph mining, social network analysis
  - Web crawling and archiving

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management
- On the systems side:
  - Graph mining, social network analysis
  - Web crawling and archiving
  - Web information extraction

Research Interests

- Foundational and practical aspects of Web data management
- How to properly manage the wealth of resources that is the World Wide Web (regular Web sites, Web databases, Web services, Social Web, Semantic Web, data exchange on the Web...)?
- On the theory side:
  - Database theory (connections to logic, finite model theory, complexity theory, verification, automata)
  - Especially, probabilistic and intensional data management
- On the systems side:
  - Graph mining, social network analysis
  - Web crawling and archiving
  - Web information extraction
  - Data management applications of reinforcement learning

Teaching Activities

- $\approx$ 300 hours of teaching per year over the past few years, at Télécom ParisTech, MPRI, Hong Kong University, National University of Singapore...

Teaching Activities

- $\approx$ 300 hours of teaching per year over the past few years, at Télécom ParisTech, MPRI, Hong Kong University, National University of Singapore...
- Topics taught:
  - Formal languages
  - Information theory
  - Databases
  - Data science
  - Web development and Web technologies
  - Web information retrieval
  - Web data management
  - Introduction to programming (C++, Java)
  - Competitive programming
  - LaTeX
- Not exhaustive!

Outline

Uncertain data is everywhere

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemas
- Imprecise automatic processes (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Uncertain data is everywhere

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemas
- Imprecise automatic processes (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

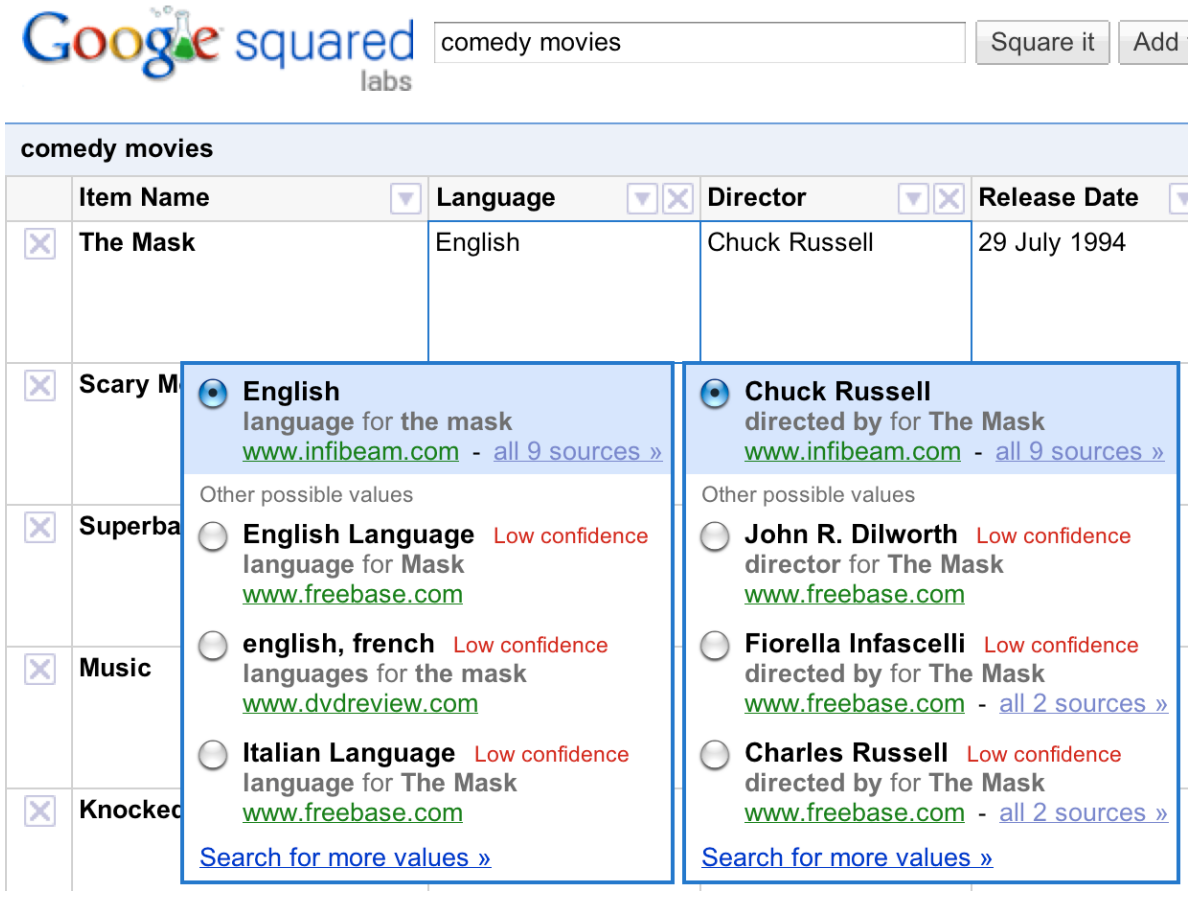
Uncertainty in Web information extraction

instance	iteration	date learned	confidence
arabic, egypt	406	08-sep-2011	(Seed) 100.0
chinese, republic_of_china	439	24-oct-2011	100.0
chinese, singapore	421	21-sep-2011	(Seed) 100.0
english, britain	439	24-oct-2011	100.0
english, canada	439	24-oct-2011	(Seed) 100.0
english, england001	439	24-oct-2011	100.0
arabic, morocco	422	23-sep-2011	100.0
cantonese, hong_kong	406	08-sep-2011	100.0
english, uk	436	19-oct-2011	100.0
english, south_vietnam	427	27-sep-2011	99.9
french, morocco	422	23-sep-2011	99.9
greek, turkey	430	07-oct-2011	99.9

Never-ending Language Learning (NELL, CMU),

`http://rtw.ml.cmu.edu/rtw/kbbrowser/`

Uncertainty in Web information extraction



Google Squared (terminated),

screenshot from [Fink, Hogue, Olteanu, and Rath 2011]

Uncertainty in Web information extraction

Subject	Predicate	Object	Confidence
Elvis Presley	diedOnDate	1977-08-16	97.91%
Elvis Presley	isMarriedTo	Priscilla Presley	97.29%
Elvis Presley	influences	Carlo Wolff	96.25%

YAGO, `http://www.mpi-inf.mpg.de/yago-naga/yago`

[Suchanek, Kasneci, and Weikum 2007]

Structured data is everywhere

Data is structured, not flat:

- Variety of representation formats of data in the wild:
  - relational tables
  - trees, semi-structured documents
  - graphs, e.g., social networks or semantic graphs
  - data streams
  - complex views aggregating individual information
- Heterogeneous schemas
- Additional structural constraints: keys, inclusion dependencies

Intensional data is everywhere

Lots of data sources can be seen as intensional: accessing all the data in the source (in extension) is impossible or very costly, but it is possible to access the data through views, with some access constraints, associated with some access cost.

- Indexes over regular data sources
- Deep Web sources: Web forms, Web services
- The Web or social networks as partial graphs that can be expanded by crawling
- Outcome of complex automated processes: information extraction, natural language analysis, machine learning, ontology matching
- Crowd data: (very) partial views of the world
- Logical consequences of facts, costly to compute

Interactions between uncertainty, structure, intensionality

- If the data has complex structure, uncertain models should represent possible worlds over these structures (e.g., probability distributions over graph completions of a known subgraph in Web crawling).
- If the data is intensional, we can use uncertainty to represent prior distributions about what may happen if we access the data. Sometimes good enough to reach a decision without having to make the access!
- If the data is a RDF graph accessed by semantic Web services, each intensional data access will not give a single data point, but a complex subgraph.

State of the art and opportunities

Probabilistic databases cover limited structure variations, do not consider intensionality

Active and reinforcement learning deal with uncertainty and intensionality, but assumes trivial structures and simple goals

Crowdsourcing, focused crawling, deep Web crawling focus on specific applications of the uncertainty/structure/intensionality problem

Answering queries using views assumes simplistic cost models

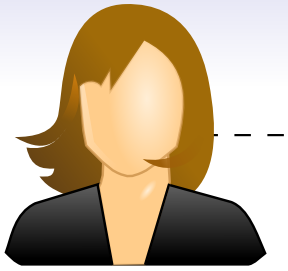
Opportunities for data management systems that take all dimensions into account

Research Framework

[SIGWEB Newletter'15: Amarilli, Maniu, and Senellart 2015]

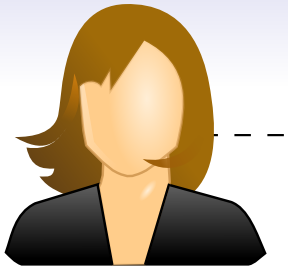
- Jointly deal with Uncertainty, Structure, and the fact that access to data is limited and has a cost, to solve a user's knowledge need
- Lazy evaluation whenever possible
- Evolving probabilistic, structured view of the current knowledge of the world
- Solve at each step the problem: What is the best access to do next given my current knowledge of the world and the knowledge need
- Knowledge acquisition plan (recursive, dynamic, adaptive) that minimizes access cost, and provides probabilistic guarantees

with 



formulation

Knowledge need

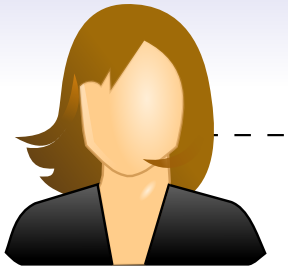


formulation

Knowledge need

Structured source profiles

modeling



formulation

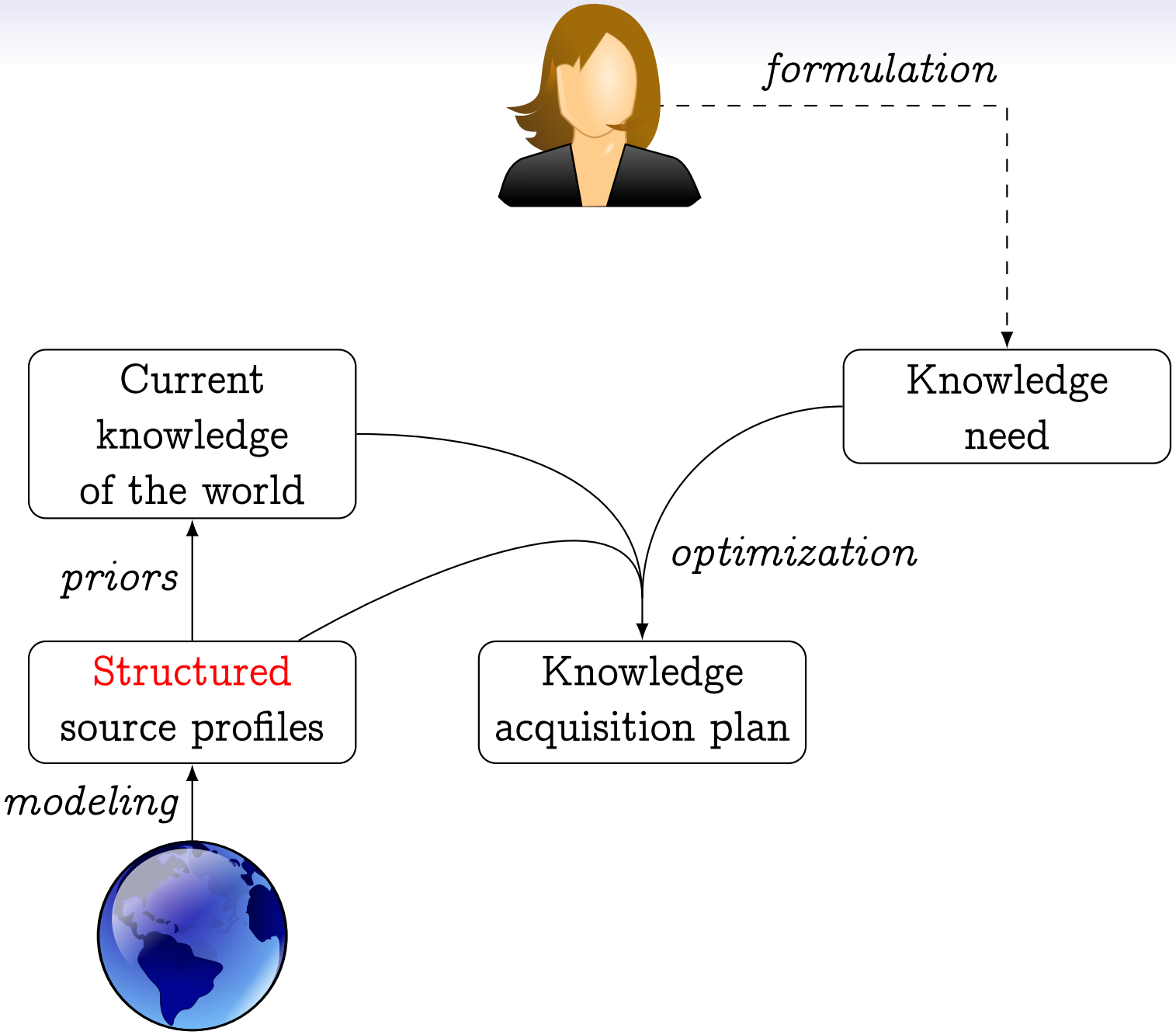
Knowledge need

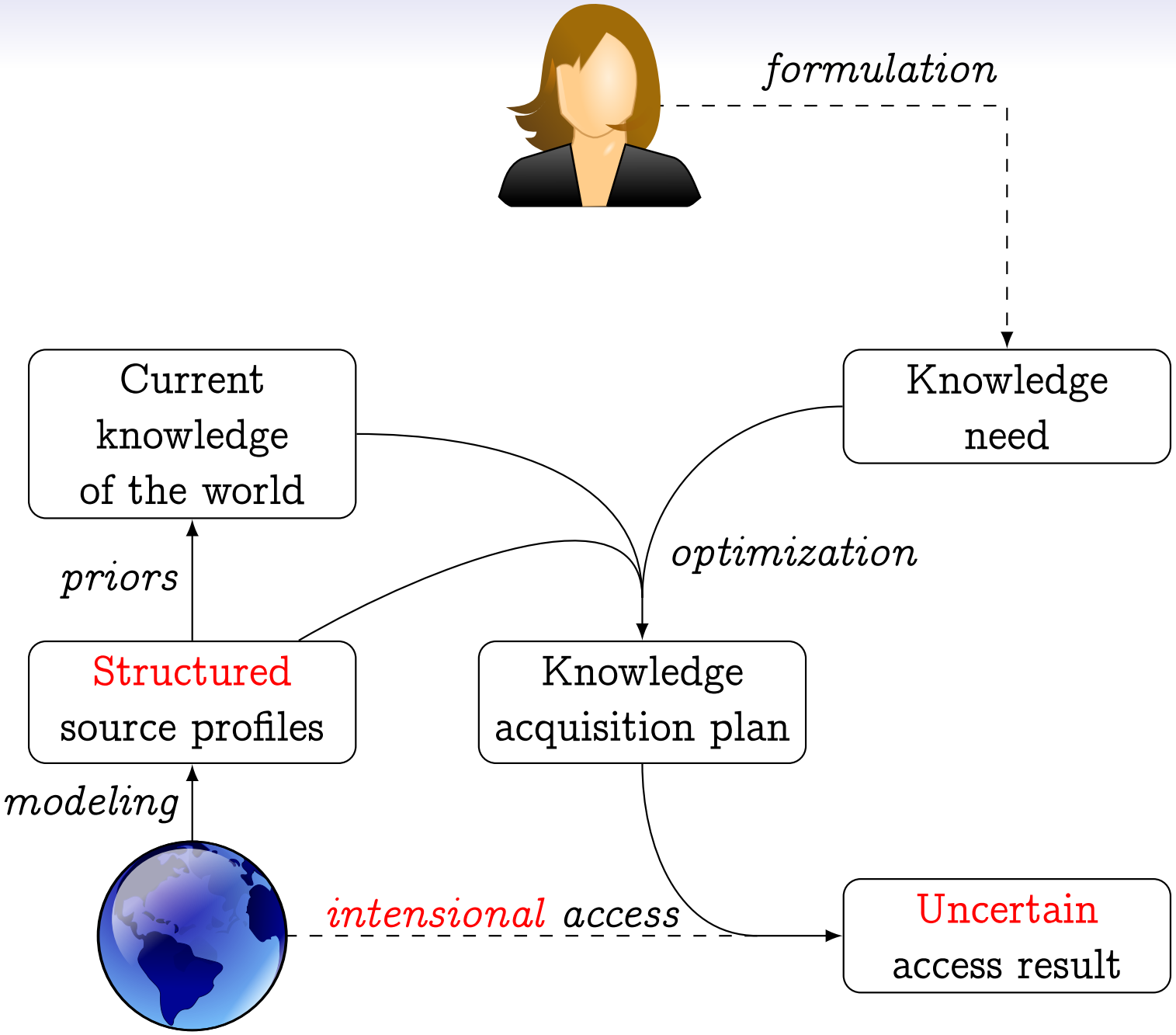
Current knowledge of the world

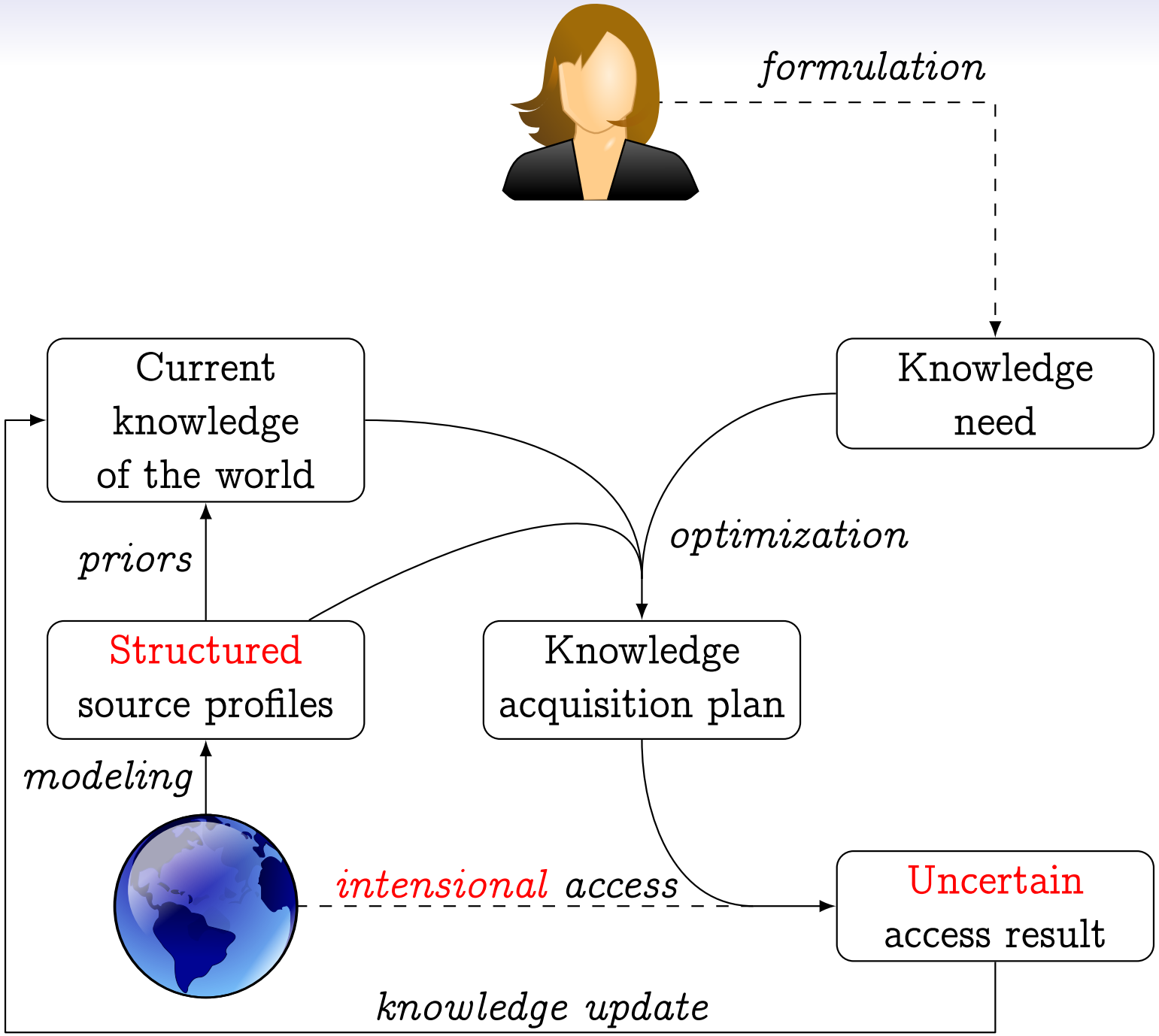
priors

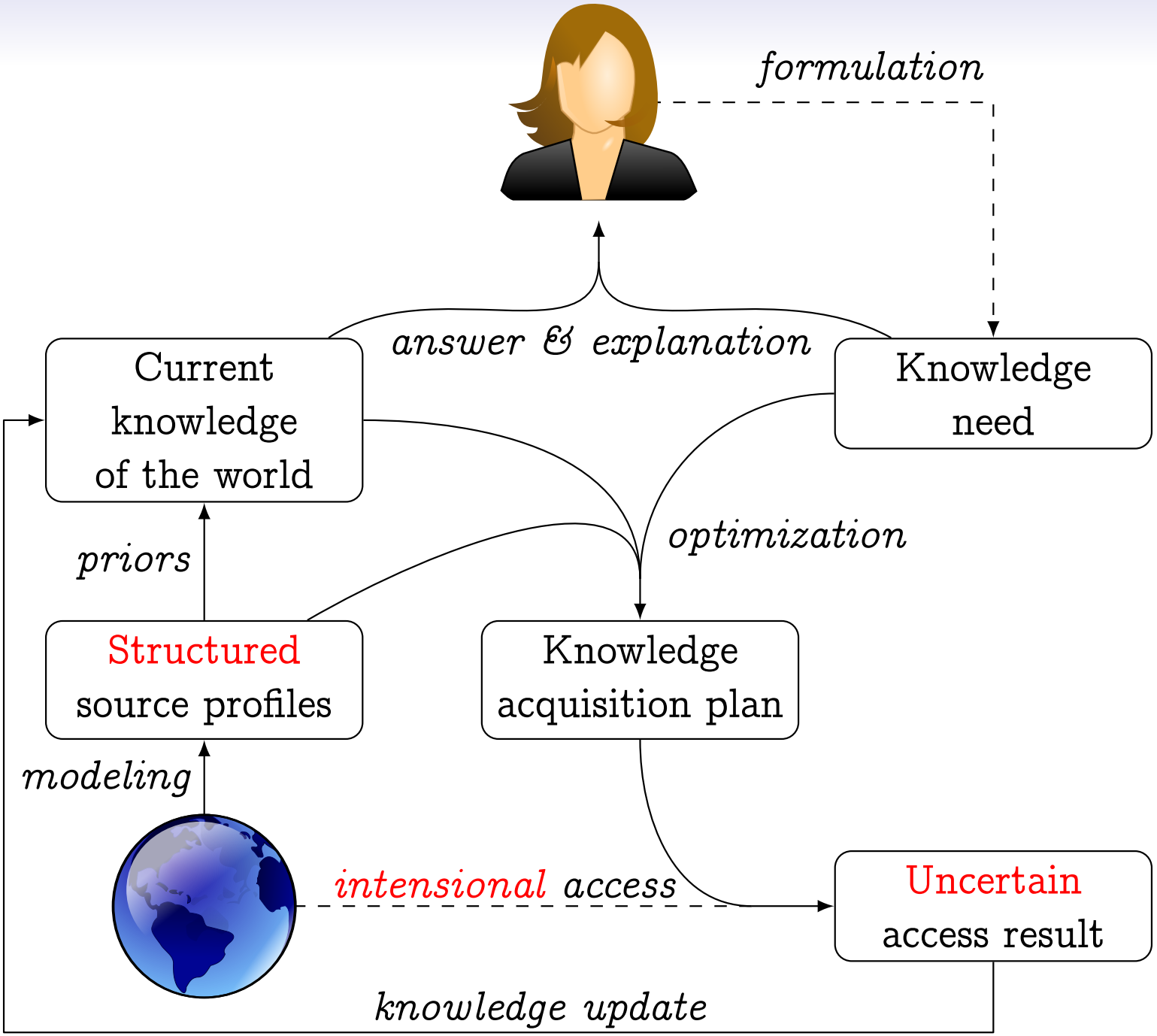
Structured source profiles

modeling


formulation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling


formulation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result

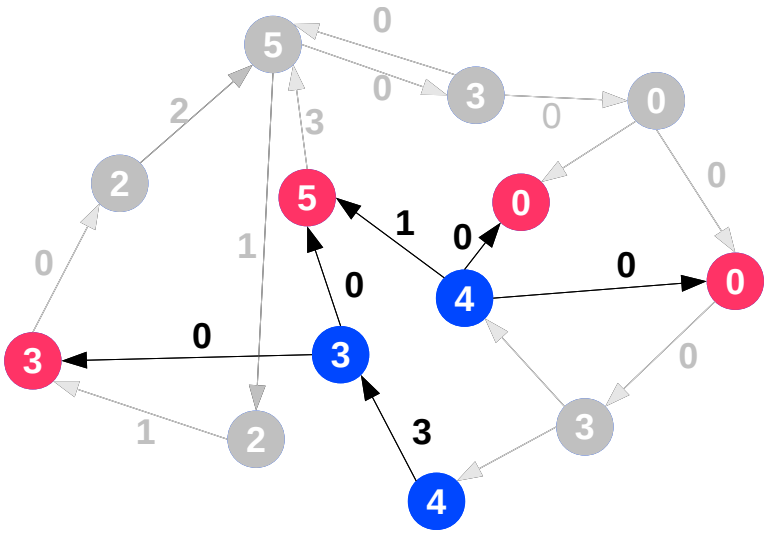

formulation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update


formulation
answer & explanation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update

Outline

Adaptive focused crawling

[HyperText'14: Gouriten, Maniu, and Senellart 2014] (best paper award)

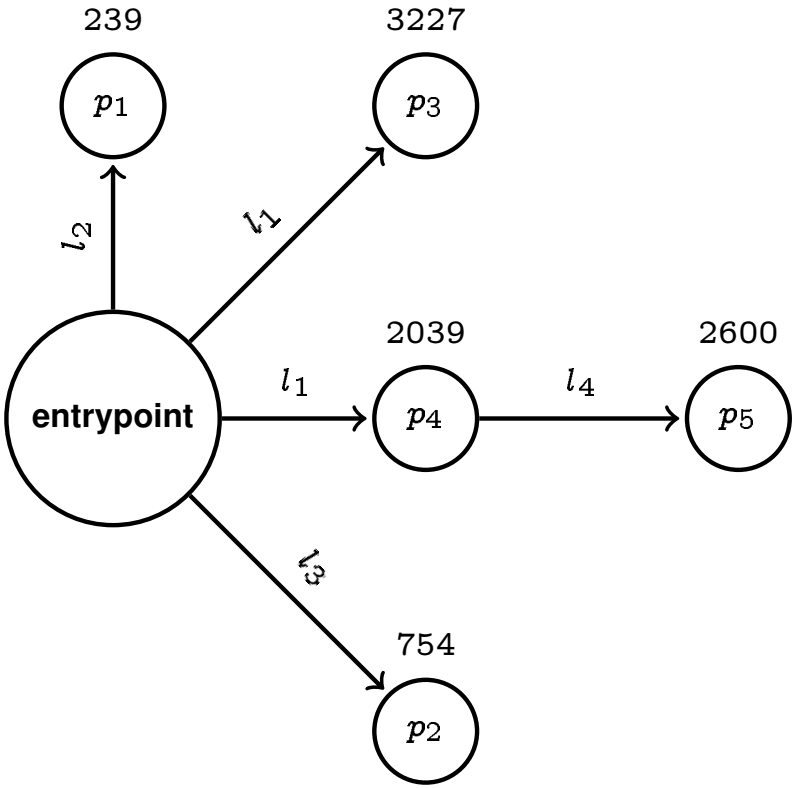


- Problem: Efficiently crawl nodes in a graph such that total score is high
- Challenge: The score of a node is unknown till it is crawled
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with 

Adaptive Web application crawling

[ICADL'15: Faheem and Senellart 2015]

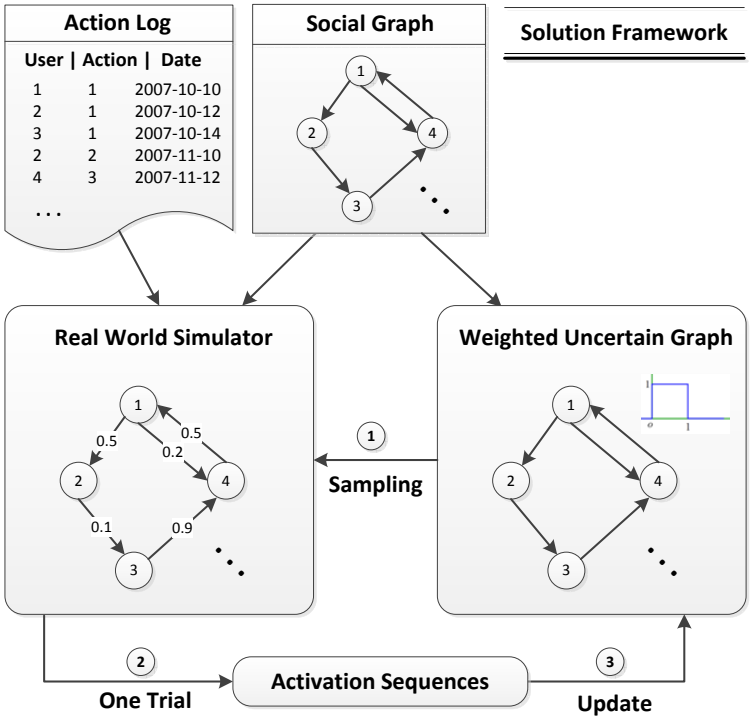


- Problem: Optimize the amount of distinct content retrieved from a Web site w.r.t. the number of HTTP requests
- Challenge: No way to know a priori where the content lies on the Web site
- Methodology: Sample a small part of the Web site and discover optimal crawling patterns from it

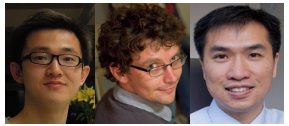
with 

Online influence maximization

[KDD'15: Lei, Maniu, Mo, Cheng, and Senellart 2015]

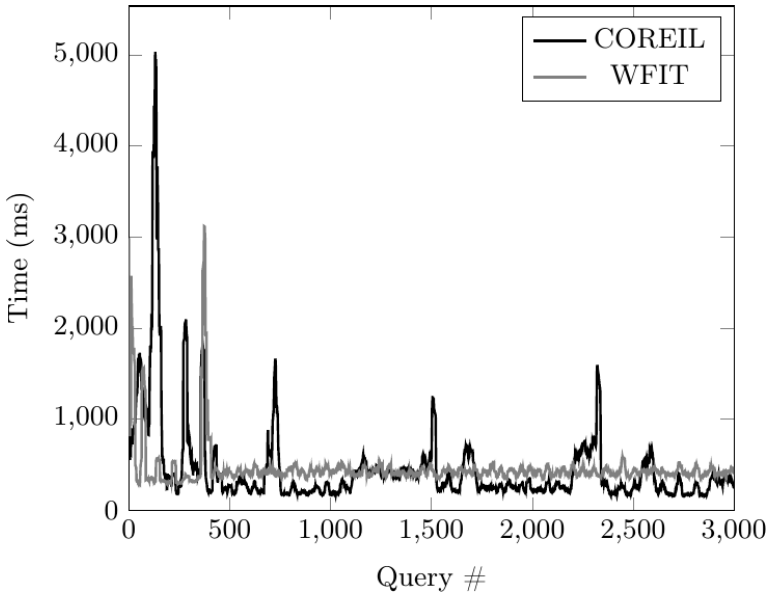


- Problem: Run influence campaigns in social networks, optimizing the amount of influenced nodes
- Challenge: Influence probabilities are unknown
- Methodology: Build a model of influence probabilities and focus on influent nodes, with an exploration/exploitation trade-off

with 

Cost-Model–Free Database Tuning

[TLDKS'16: Basu, Lin, Chen, Vo, Yuan, Senellart, and Bressan 2016]



- Problem: Automatically find which indexes to create in a database for optimal performance
- Challenge: The workload and cost model may be unknown
- Methodology: Use reinforcement learning techniques to iteratively learn a cost model and workload characteristics

with 

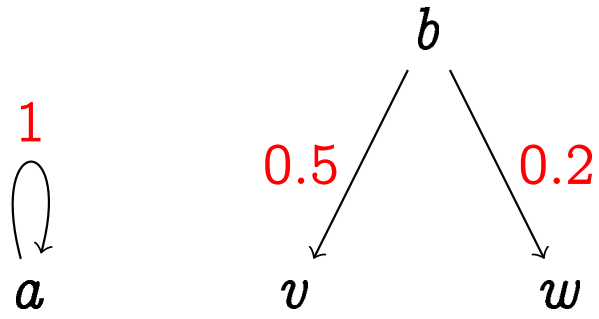
Outline

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2

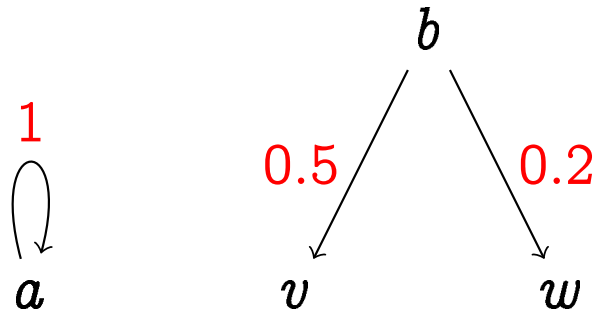
Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



Tuple-independent databases (TID)

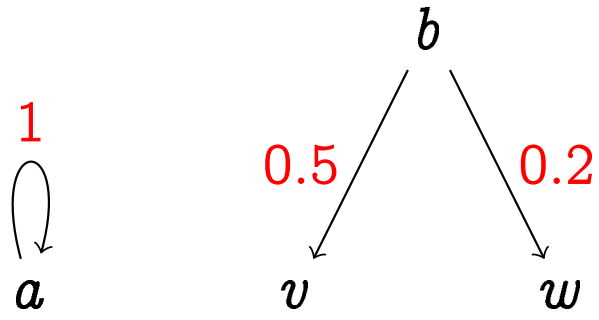
S		
a	a	1
b	v	0.5
b	w	0.2



This TID instance represents the following probability distribution:

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



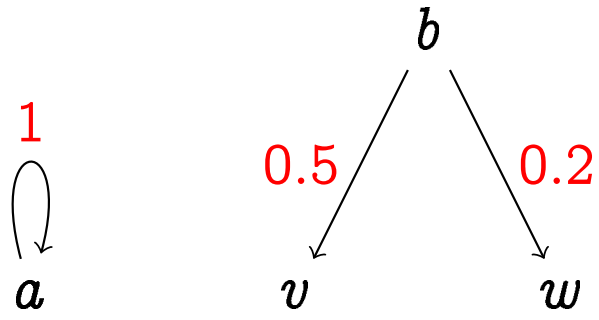
This TID instance represents the following probability distribution:

0.5×0.2

S	
a	a
b	v
b	w

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



This TID instance represents the following probability distribution:

0.5×0.2

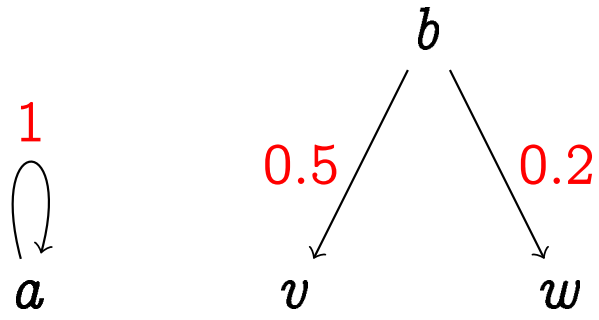
S	
a	a
b	v
b	w

$0.5 \times (1 - 0.2)$

S	
a	a
b	v

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



This TID instance represents the following probability distribution:

0.5×0.2

S	
a	a
b	v
b	w

$0.5 \times (1 - 0.2)$

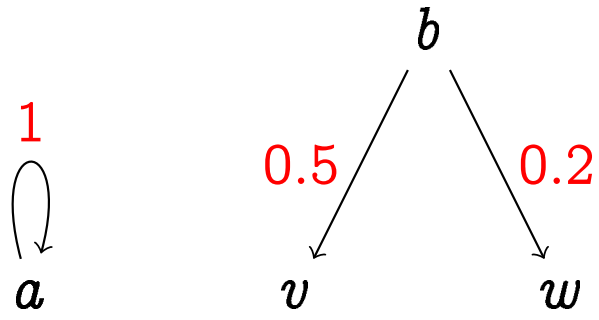
S	
a	a
b	v

$(1 - 0.5) \times 0.2$

S	
a	a
b	w

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



This TID instance represents the following probability distribution:

0.5×0.2

S	
a	a
b	v
b	w

$0.5 \times (1 - 0.2)$

S	
a	a
b	v

$(1 - 0.5) \times 0.2$

S	
a	a
b	w

$(1 - 0.5) \times (1 - 0.2)$

S	
a	a

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

$0.4 \times$

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

$0.4 \times (1 -$

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

$0.4 \times (1 - (1 - 0.5 \times 0.3)$

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

$0.4 \times (1 - (1 - 0.5 \times 0.3) \times (1 - 0.2 \times 0.7))$

Query evaluation on probabilistic instances

We want to evaluate the probability of a query on a TID instance

$$q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$$

R	
a	1
b	0.4
c	0.6

S		
a	a	1
b	v	0.5
b	w	0.2

T	
v	0.3
w	0.7
b	1

- The query is true iff $R(b)$ is here and one of:
  - $S(b, v)$ and $T(v)$ are here
  - $S(b, w)$ and $T(w)$ are here

$\rightarrow$ Probability:

$0.4 \times (1 - (1 - 0.5 \times 0.3) \times (1 - 0.2 \times 0.7)) = 0.1076$

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

- Existing dichotomy result: [Dalvi and Suciu 2012]
  - $\mathcal{Q}$ are (unions of) conjunctive queries, $\mathcal{I}$ is all TID instances
  - There is a class $\mathcal{S} \subseteq \mathcal{Q}$ of safe queries

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

- Existing dichotomy result: [Dalvi and Suciu 2012]
  - $\mathcal{Q}$ are (unions of) conjunctive queries, $\mathcal{I}$ is all TID instances
  - There is a class $\mathcal{S} \subseteq \mathcal{Q}$ of safe queries
  - PQE is PTIME for any $q \in \mathcal{S}$ on all instances

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

- Existing dichotomy result: [Dalvi and Suciu 2012]
  - $\mathcal{Q}$ are (unions of) conjunctive queries, $\mathcal{I}$ is all TID instances
  - There is a class $\mathcal{S} \subseteq \mathcal{Q}$ of safe queries
  - PQE is PTIME for any $q \in \mathcal{S}$ on all instances
  - PQE is #P-hard for any $q \in \mathcal{Q} \backslash \mathcal{S}$ on all instances

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

- Existing dichotomy result: [Dalvi and Suciu 2012]
  - $\mathcal{Q}$ are (unions of) conjunctive queries, $\mathcal{I}$ is all TID instances
  - There is a class $\mathcal{S} \subseteq \mathcal{Q}$ of safe queries
  - PQE is PTIME for any $q \in \mathcal{S}$ on all instances
  - PQE is #P-hard for any $q \in \mathcal{Q} \backslash \mathcal{S}$ on all instances
  - $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$ is unsafe!

Complexity of probabilistic query evaluation (PQE)

What is the data complexity of probabilistic query evaluation on TID depending on the class $\mathcal{Q}$ of queries and class $\mathcal{I}$ of instances?

- Existing dichotomy result: [Dalvi and Suciu 2012]
  - $\mathcal{Q}$ are (unions of) conjunctive queries, $\mathcal{I}$ is all TID instances
  - There is a class $\mathcal{S} \subseteq \mathcal{Q}$ of safe queries
  - PQE is PTIME for any $q \in \mathcal{S}$ on all instances
  - PQE is #P-hard for any $q \in \mathcal{Q} \backslash \mathcal{S}$ on all instances
  - $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$ is unsafe!

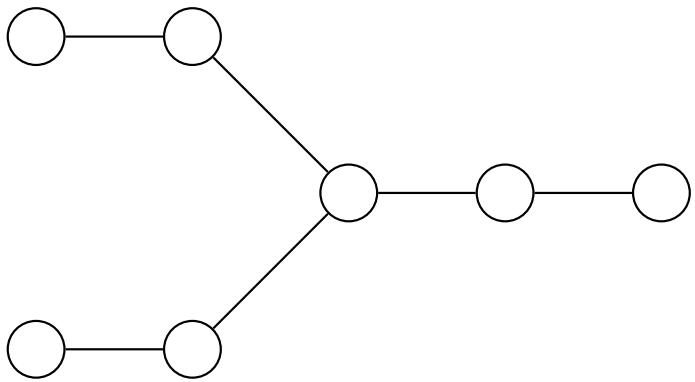
Is there a smaller class $\mathcal{I}$ such that PQE is tractable for a larger $\mathcal{Q}$?

Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

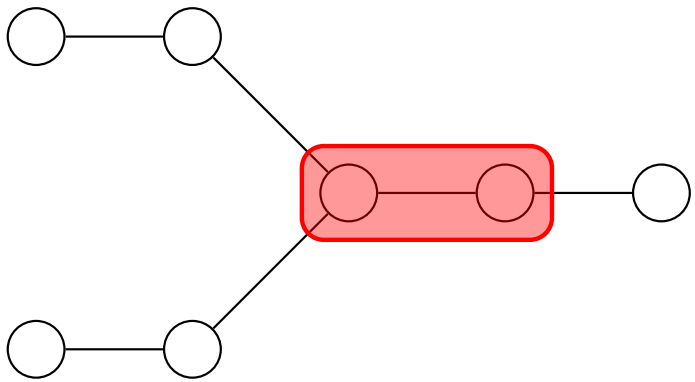
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



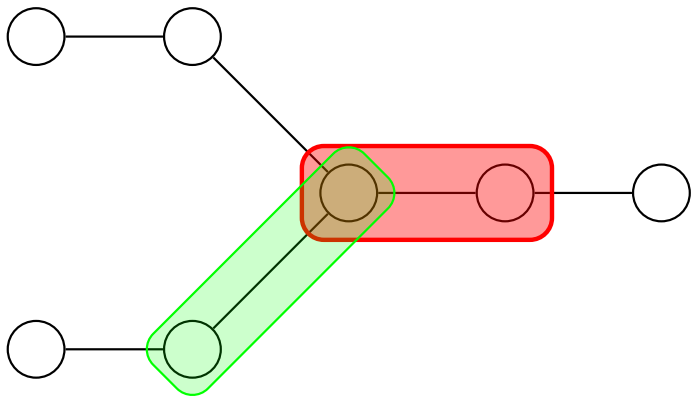
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



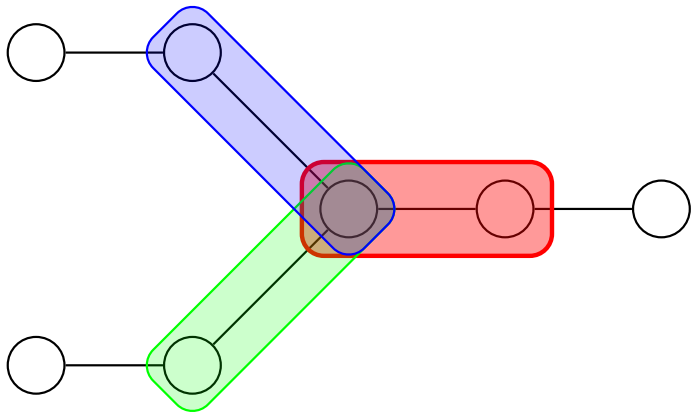
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



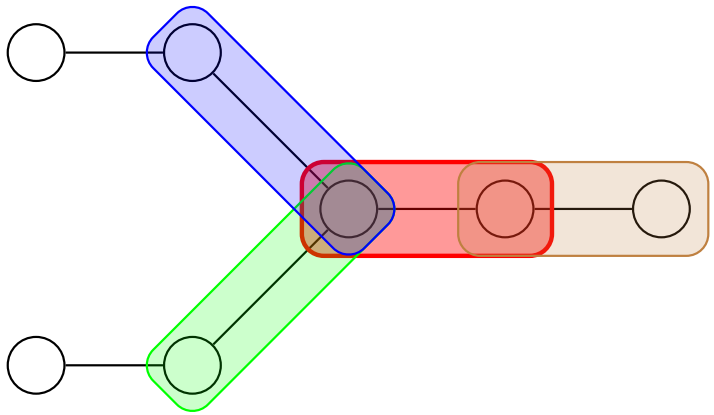
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



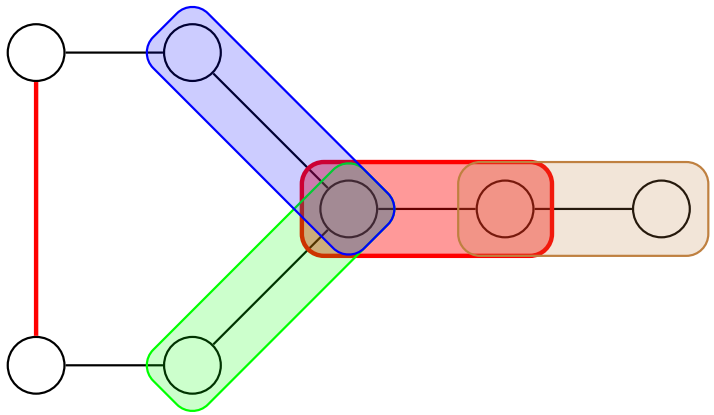
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



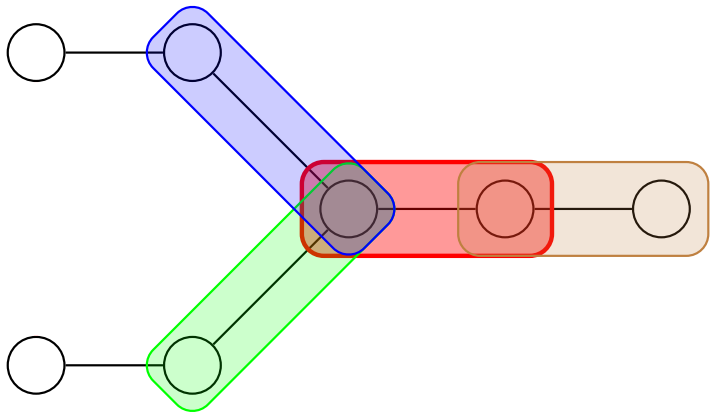
Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



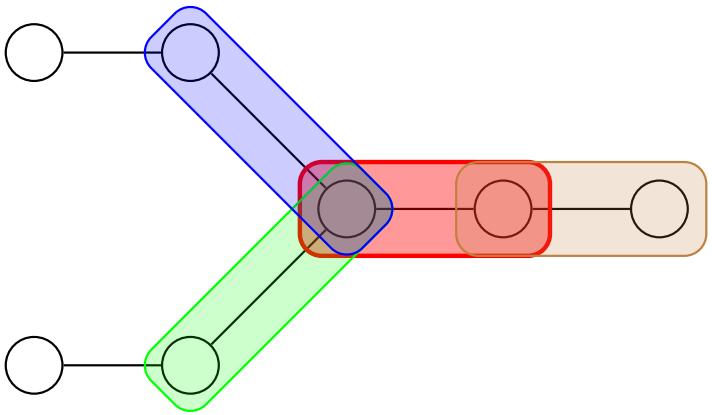
Trees and treelike instances

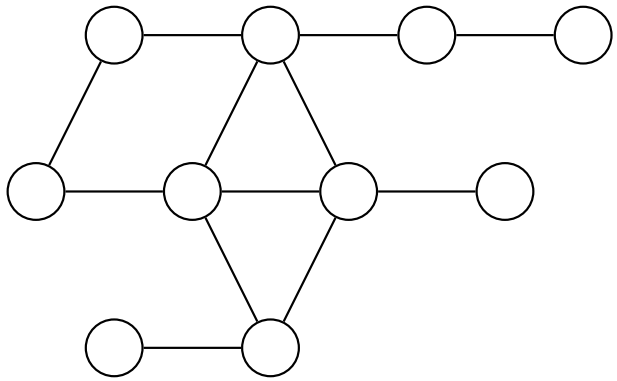
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



Trees and treelike instances

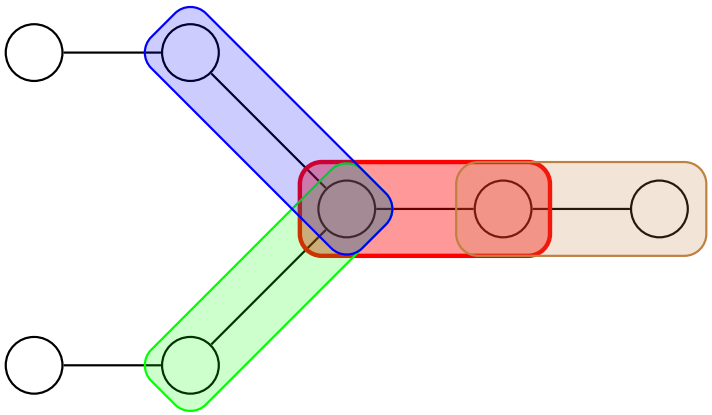
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

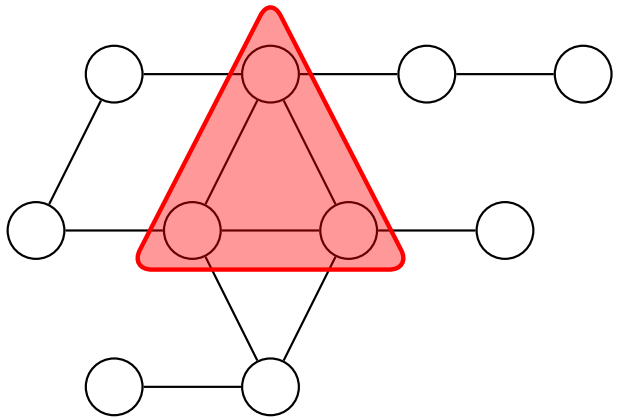




Trees and treelike instances

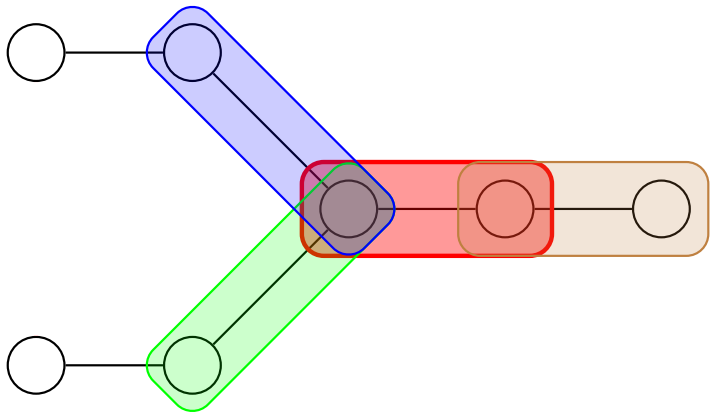
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

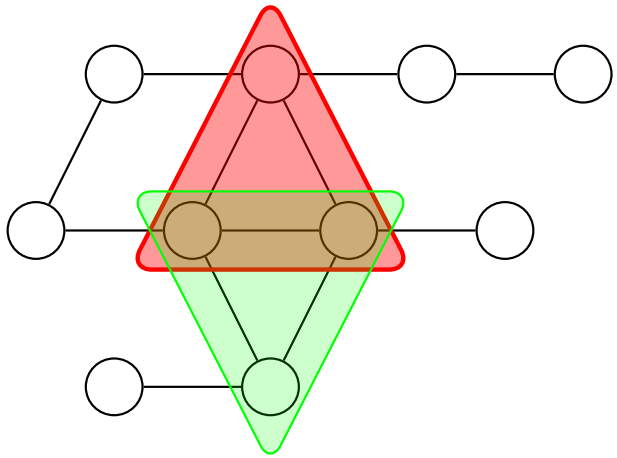




Trees and treelike instances

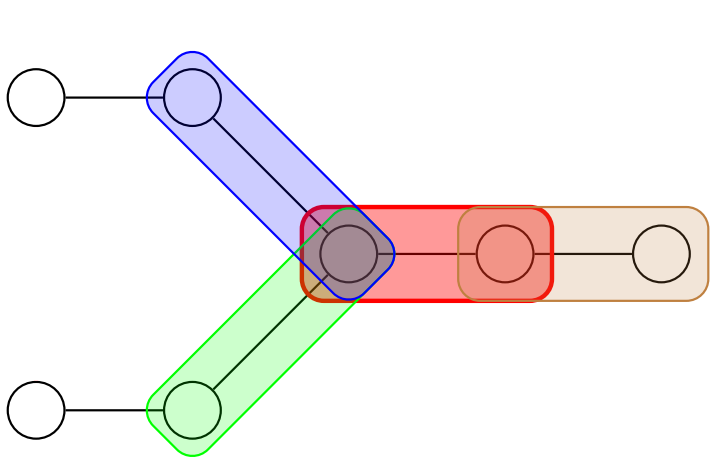
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

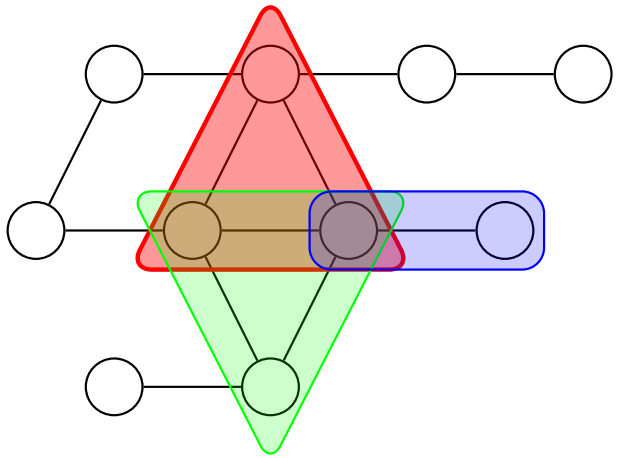




Trees and treelike instances

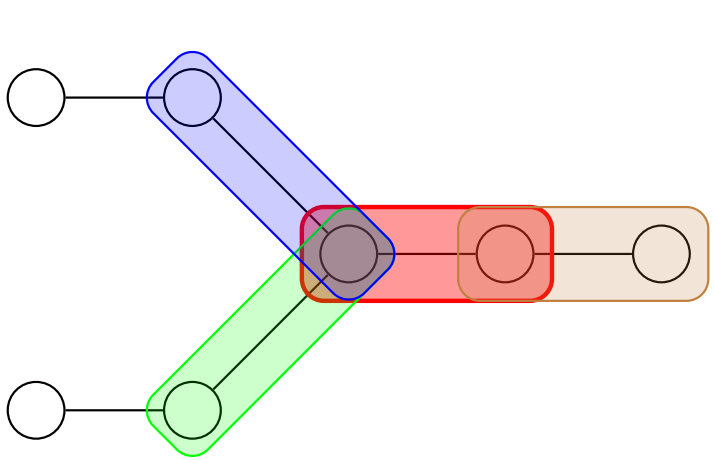
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

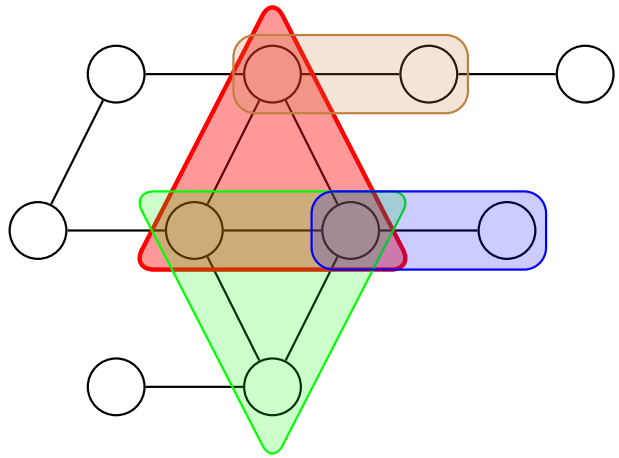




Trees and treelike instances

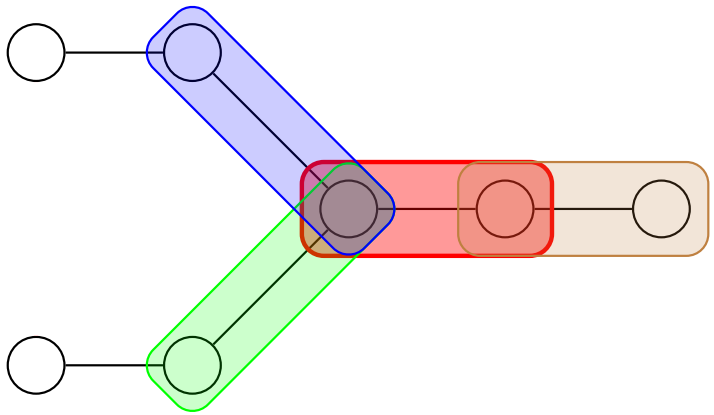
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

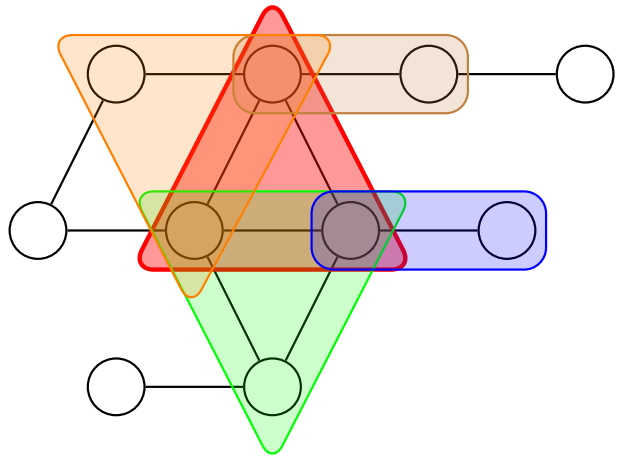




Trees and treelike instances

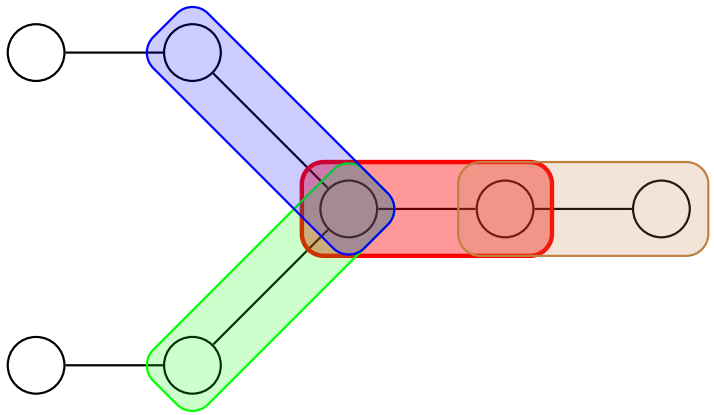
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

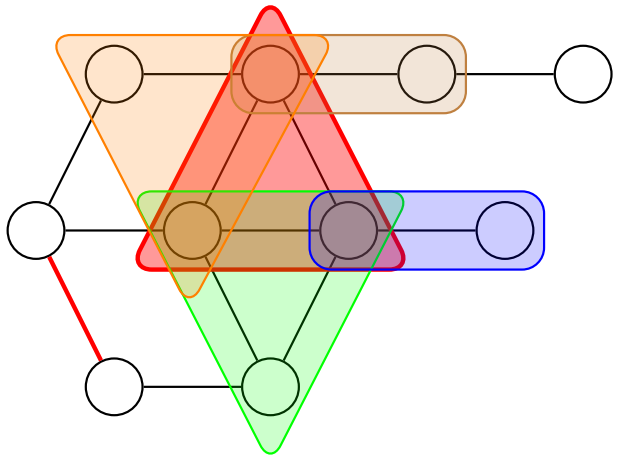




Trees and treelike instances

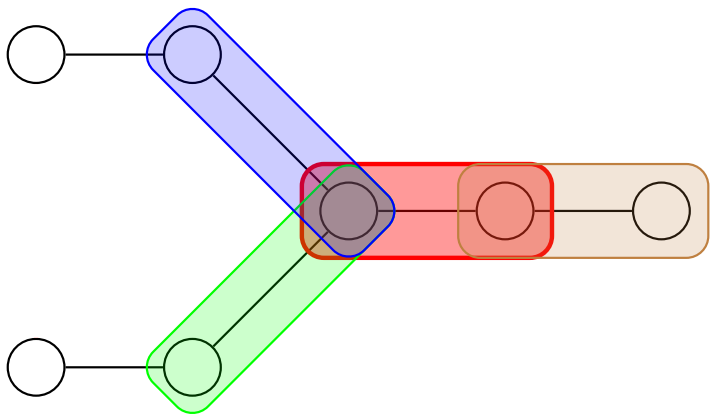
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

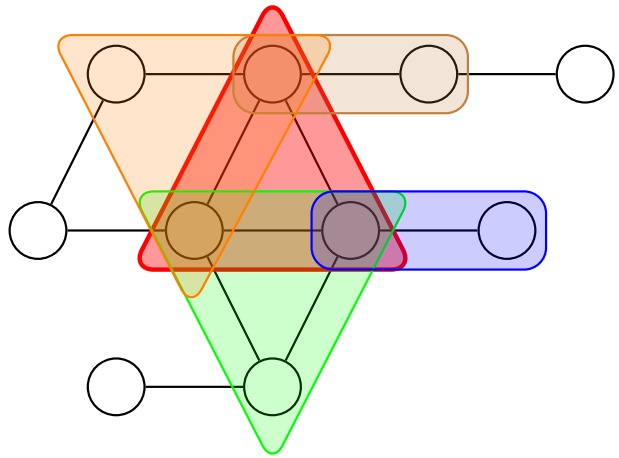




Trees and treelike instances

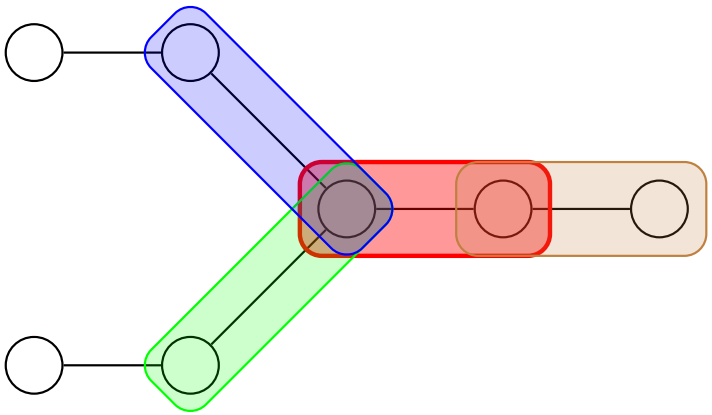
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

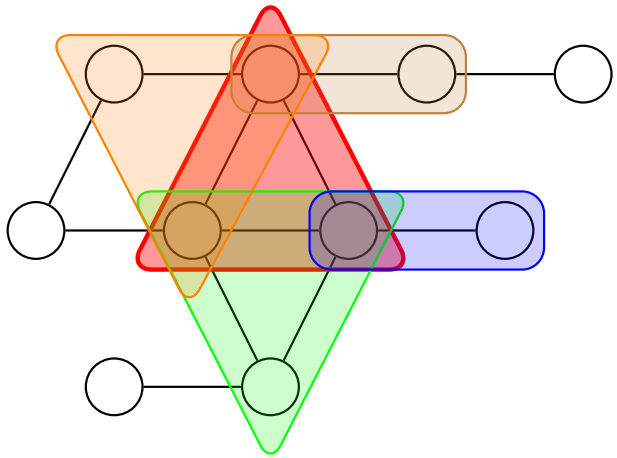




Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

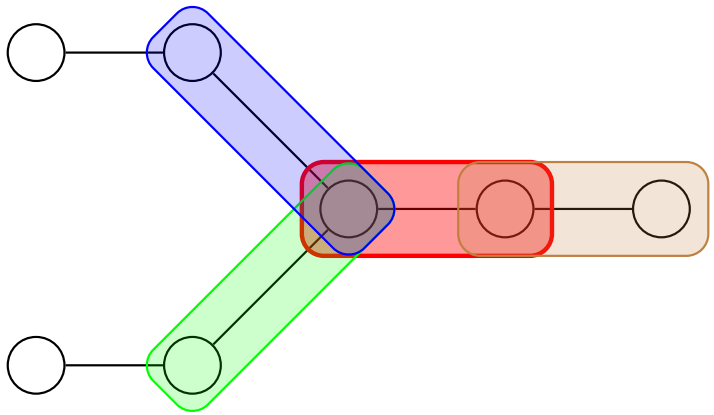


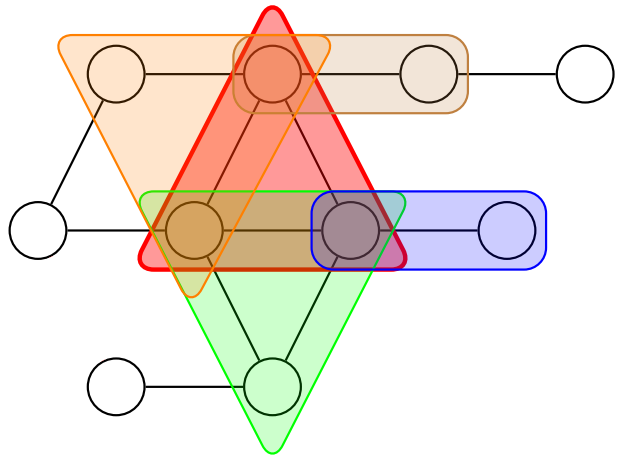


  - Trees have treewidth 1
  - Cycles have treewidth 2
  - k-cliques and $(k-1)$-grids have treewidth $k-1$

Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)





  - Trees have treewidth 1
  - Cycles have treewidth 2
  - k-cliques and $(k-1)$-grids have treewidth $k-1$

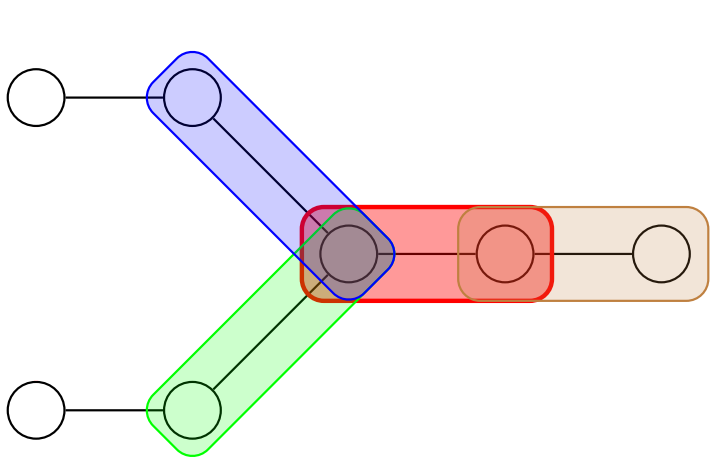
$\rightarrow$ Known results [Courcelle 1990]:

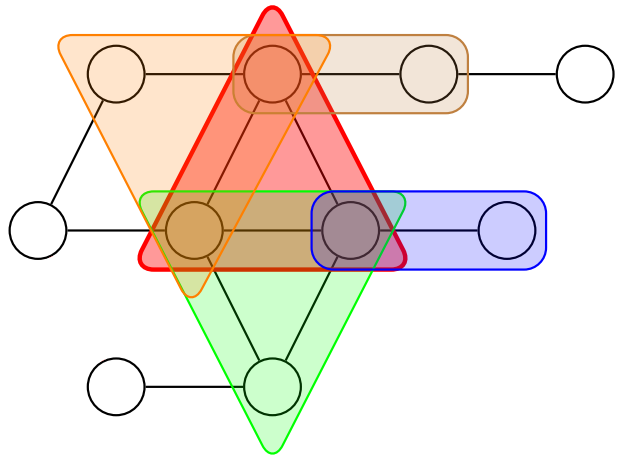
  - $\mathcal{I}$: treelike instances; $\mathcal{Q}$: monadic second-order queries

  $\rightarrow$ non-probabilistic QE is in linear time

Trees and treelike instances

- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)





  - Trees have treewidth 1
  - Cycles have treewidth 2
  - k-cliques and $(k-1)$-grids have treewidth $k-1$

$\rightarrow$ Known results [Courcelle 1990]:

  - $\mathcal{I}$: treelike instances; $\mathcal{Q}$: monadic second-order queries

  $\rightarrow$ non-probabilistic QE is in linear time

$\rightarrow$ Does this extend to probabilistic QE?

Dichotomy for PQE

An instance-based dichotomy result:

Upper bound. [ICALP'15: Amarilli, Bourhis, and Senellart 2015]

For $\mathcal{I}$ the treelike instances and $\mathcal{Q}$ the MSO queries

$\rightarrow$ PQE is in linear time modulo arithmetic costs

Dichotomy for PQE

An instance-based dichotomy result:

Upper bound. [ICALP'15: Amarilli, Bourhis, and Senellart 2015]

For $\mathcal{I}$ the treelike instances and $\mathcal{Q}$ the MSO queries

$\rightarrow$ PQE is in linear time modulo arithmetic costs

- Also for expressive provenance representations
- Also with bounded-treewidth correlations

with 

Dichotomy for PQE

An instance-based dichotomy result:

Upper bound. [ICALP'15: Amarilli, Bourhis, and Senellart 2015]

For $\mathcal{I}$ the treelike instances and $\mathcal{Q}$ the MSO queries

→ PQE is in linear time modulo arithmetic costs

- Also for expressive provenance representations
- Also with bounded-treewidth correlations

Lower bound. [PODS'16: Amarilli, Bourhis, and Senellart 2016]

For any unbounded-tw family $\mathcal{I}$ and $\mathcal{Q}$ the FO queries

→ PQE is #P-hard under RP reductions assuming:

- High-tw instances in $\mathcal{I}$ are easily constructible
- Signature arity is 2 (graphs)

with 

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage:

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage: $f_2 \wedge$

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage: $f_2 \wedge ($

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage: $f_2 \wedge ((g_2 \wedge h_1)$

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage: $f_2 \wedge ((g_2 \wedge h_1) \vee (g_3 \wedge h_2))$

Technical tool: lineages

The lineage of a query q on an instance I:

- Boolean function ϕ whose variables are the facts of I
- A subinstance of I satisfies q iff ϕ is true for that valuation

Example: $q : \exists x\, y\ R(x) \wedge S(x, y) \wedge T(y)$

R	
a	f_1
b	f_2
c	f_3

S		
a	a	g_1
b	v	g_2
b	w	g_3

T	
v	h_1
w	h_2
b	h_3

$\rightarrow$ Lineage: $f_2 \wedge ((g_2 \wedge h_1) \vee (g_3 \wedge h_2))$

$\rightarrow$ For all $\nu : I \rightarrow \{0, 1\}$ we have $\nu(\phi) = 1$ iff $\{F \in I \mid \nu(F) = 1\} \models q$

Using lineages

- Use lineage for PQE:

Using lineages

- Use lineage for PQE:
  - Compute a lineage representation efficiently

Using lineages

- Use lineage for PQE:
  - Compute a lineage representation efficiently

  → Probability of the lineage = probability of the query

Using lineages

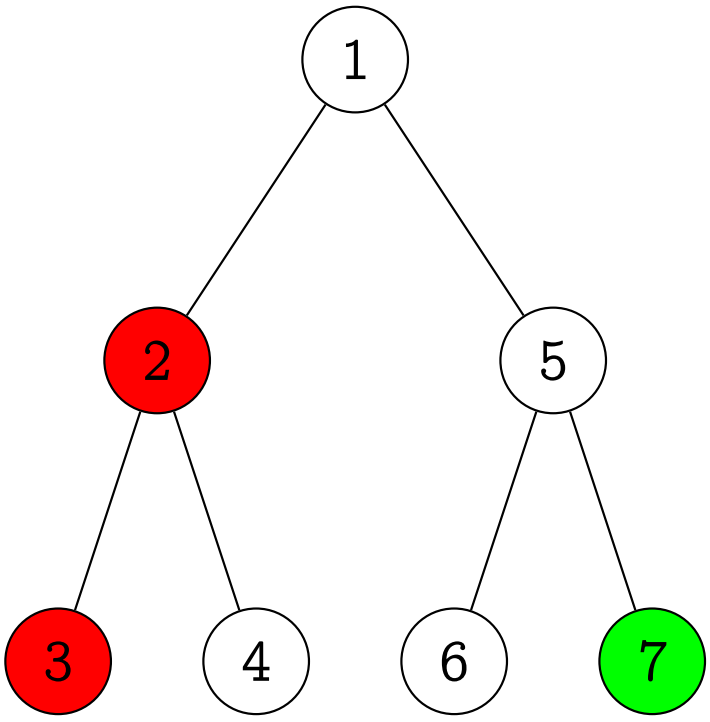
- Use lineage for PQE:
  - Compute a lineage representation efficiently

  → Probability of the lineage = probability of the query
  - Compute the lineage probability efficiently
    (show it is not #P-hard as in the general case)

Uncertain trees

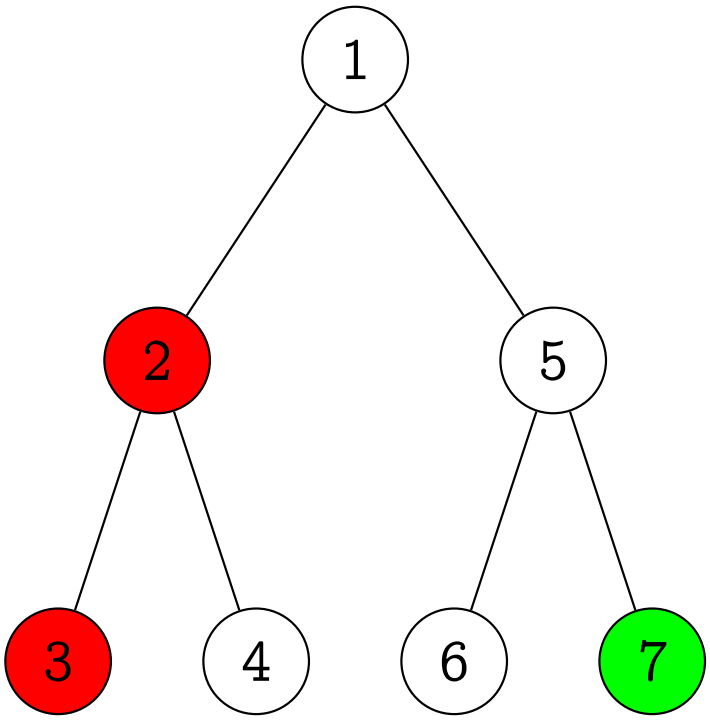
- First compute lineages on uncertain trees then use [Courcelle 1990]

Uncertain trees



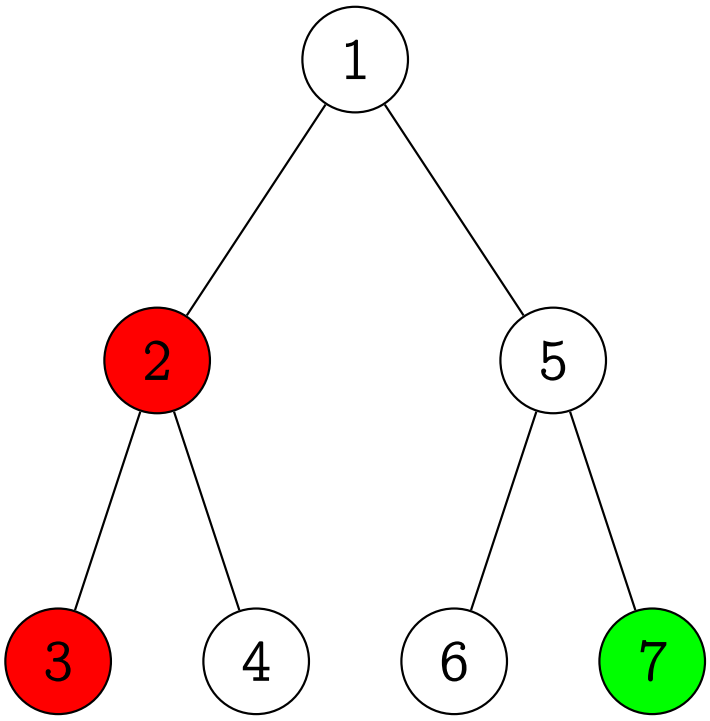
- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded

Uncertain trees



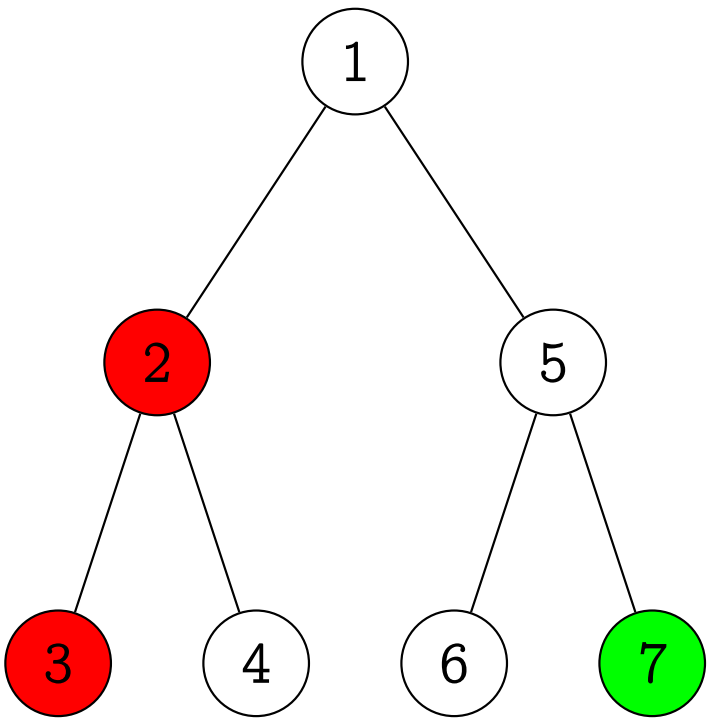
- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded
- A valuation indicates which labels are kept

Uncertain trees



- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded
- A valuation indicates which labels are kept
- Example query:
  "Is there both a red and a green node?"

Uncertain trees

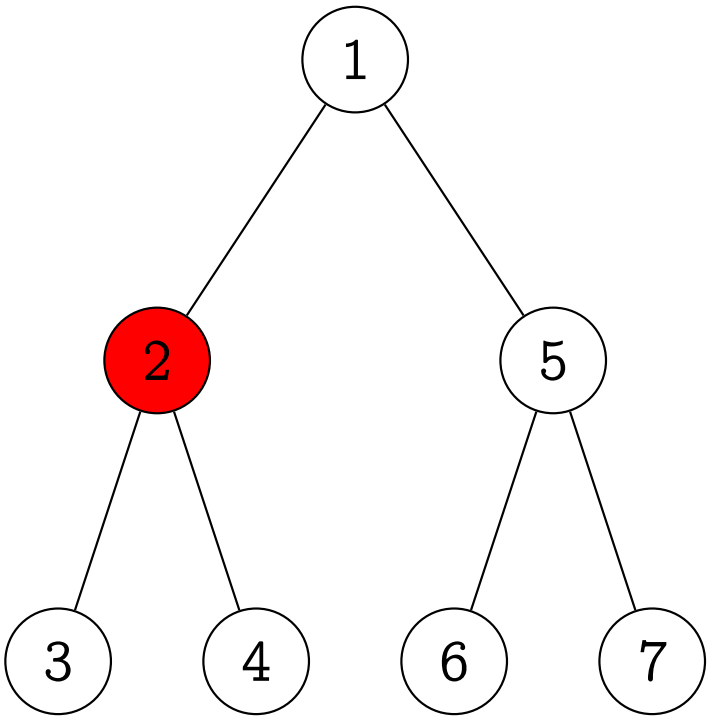


- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded
- A valuation indicates which labels are kept
- Example query:
  "Is there both a red and a green node?"

Valuation: $\{2, 3, 7\}$

The query is true

Uncertain trees

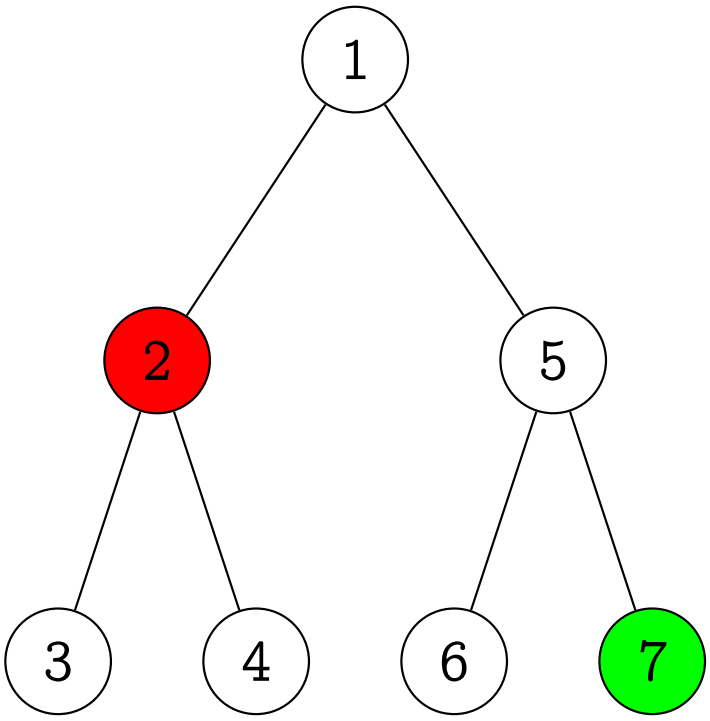


- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded
- A valuation indicates which labels are kept
- Example query:
  "Is there both a red and a green node?"

Valuation: {2}

The query is false

Uncertain trees

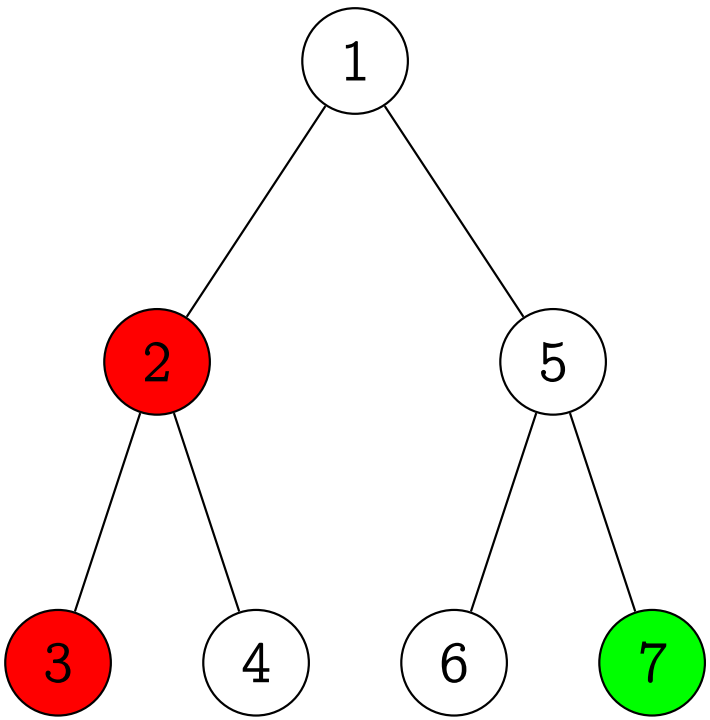


- First compute lineages on uncertain trees then use [Courcelle 1990]
- Uncertain trees: node labels may be discarded
- A valuation indicates which labels are kept
- Example query:
  "Is there both a red and a green node?"

Valuation: $\{2, 7\}$

The query is true

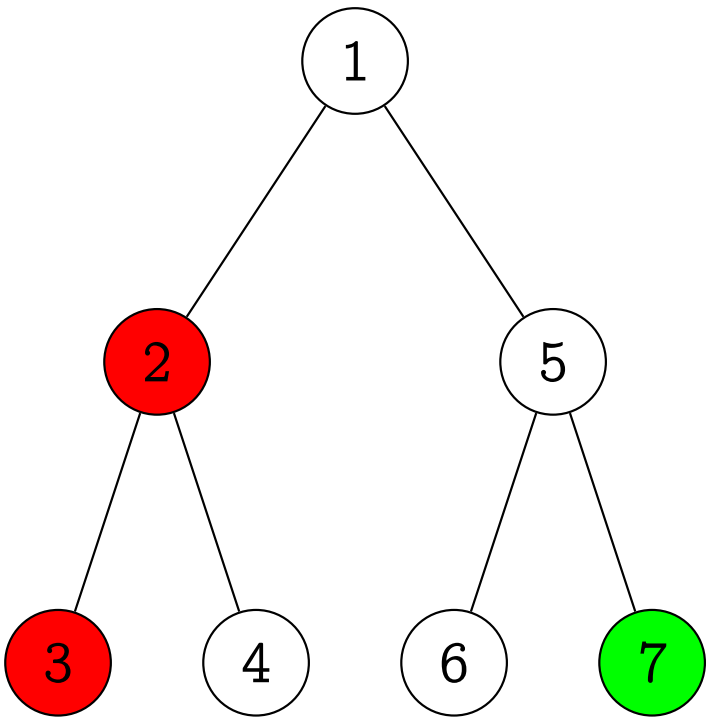
Lineage circuits on trees



q: Is there both a red and a green node?

- Which valuations satisfy q? ($\Leftrightarrow$ lineage)

Lineage circuits on trees



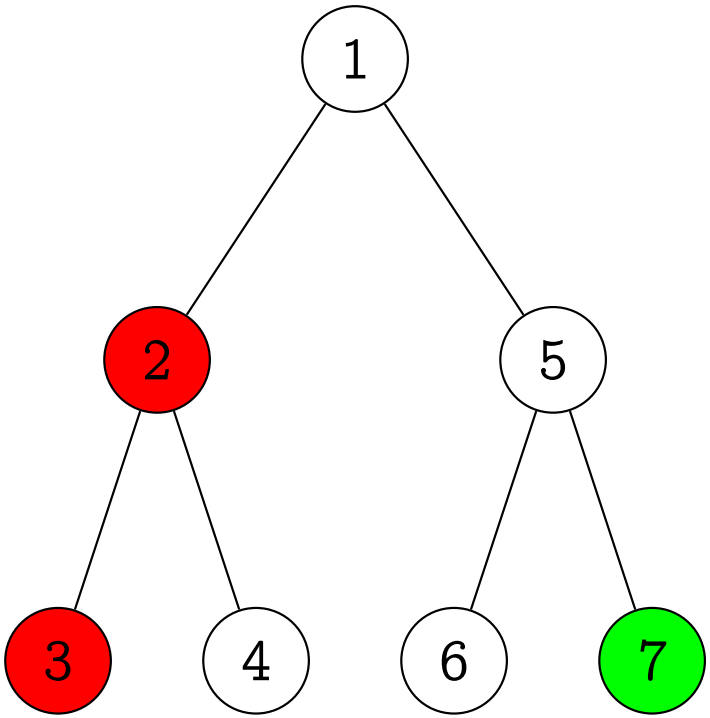
q: Is there both a red and a green node?

- Which valuations satisfy q? ($\Leftrightarrow$ lineage)

- Lineage circuit of a query q on an uncertain tree T
  - Boolean circuit C
  - with input gates g_2, g_3, g_7

  $\rightarrow$ $\nu(T)$ satisfies q iff $\nu(C)$ is true

Lineage circuits on trees

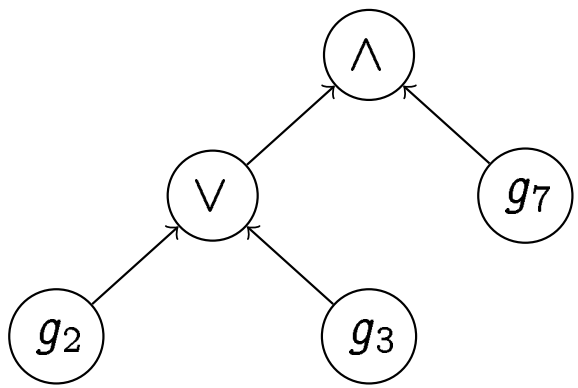


q: Is there both a red and a green node?

- Which valuations satisfy q? ($\Leftrightarrow$ lineage)

- Lineage circuit of a query q on an uncertain tree T
  - Boolean circuit C
  - with input gates g_2, g_3, g_7

  $\rightarrow$ $\nu(T)$ satisfies q iff $\nu(C)$ is true



Upper bound

Theorem

For any query q given as a bottom-up tree automaton A, for any input tree T, we can build a lineage circuit of A on T in linear time in $|A| \cdot |T|$.

Upper bound

Theorem

For any query q given as a bottom-up tree automaton A, for any input tree T, we can build a lineage circuit of A on T in linear time in $|A| \cdot |T|$.

MSO on treelike instances $\Rightarrow$ MSO on trees [Courcelle 1990].

Upper bound

Theorem

For any query q given as a bottom-up tree automaton A, for any input tree T, we can build a lineage circuit of A on T in linear time in $|A| \cdot |T|$.

MSO on treelike instances $\Rightarrow$ MSO on trees [Courcelle 1990].

Theorem

For any fixed MSO query q and $k \in \mathbb{N}$, for any input instance I of treewidth $\leq k$, we can build in linear time in I a lineage circuit of q on I.

Upper bound

Theorem

For any query q given as a bottom-up tree automaton A, for any input tree T, we can build a lineage circuit of A on T in linear time in $|A| \cdot |T|$.

MSO on treelike instances $\Rightarrow$ MSO on trees [Courcelle 1990].

Theorem

For any fixed MSO query q and $k \in \mathbb{N}$, for any input instance I of treewidth $\leq k$, we can build in linear time in I a lineage circuit of q on I.

The lineage circuits are themselves treelike, hence:

Upper bound

Theorem

For any query q given as a bottom-up tree automaton A, for any input tree T, we can build a lineage circuit of A on T in linear time in $|A| \cdot |T|$.

MSO on treelike instances $\Rightarrow$ MSO on trees [Courcelle 1990].

Theorem

For any fixed MSO query q and $k \in \mathbb{N}$, for any input instance I of treewidth $\leq k$, we can build in linear time in I a lineage circuit of q on I.

The lineage circuits are themselves treelike, hence:

Corollary

Probabilistic query evaluation of MSO queries on treelike instances is in linear time up to arithmetic costs.

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

$\rightarrow$ Restrict to arity-2 (= labeled graphs) for technical reasons

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

$\rightarrow$ Restrict to arity-2 (= labeled graphs) for technical reasons

$\rightarrow$ Impose that $\mathcal{I}$ is tw-constructible:

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

$\rightarrow$ Restrict to arity-2 (= labeled graphs) for technical reasons

$\rightarrow$ Impose that $\mathcal{I}$ is tw-constructible:

- Given $k \in \mathbb{N}$, we can construct in time $\mathrm{Poly}(k)$ an instance of $\mathcal{I}$ of treewidth $\geq k$

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

$\rightarrow$ Restrict to arity-2 (= labeled graphs) for technical reasons

$\rightarrow$ Impose that $\mathcal{I}$ is tw-constructible:

- Given $k \in \mathbb{N}$, we can construct in time Poly(k) an instance of $\mathcal{I}$ of treewidth $\geq k$

Theorem

There is a first-order query q such that for any unbounded-tw, tw-constructible, arity-2 instance family $\mathcal{I}$, probabilistic query eval for q on $\mathcal{I}$ is #P-hard under RP reductions.

Lower bound

- Class $\mathcal{I}$ of unbounded-treewidth instances, query q in class $\mathcal{Q}$
- Show that probabilistic query evaluation of q on $\mathcal{I}$ is hard

$\rightarrow$ Restrict to arity-2 (= labeled graphs) for technical reasons

$\rightarrow$ Impose that $\mathcal{I}$ is tw-constructible:

- Given $k \in \mathbb{N}$, we can construct in time Poly(k) an instance of $\mathcal{I}$ of treewidth $\geq k$

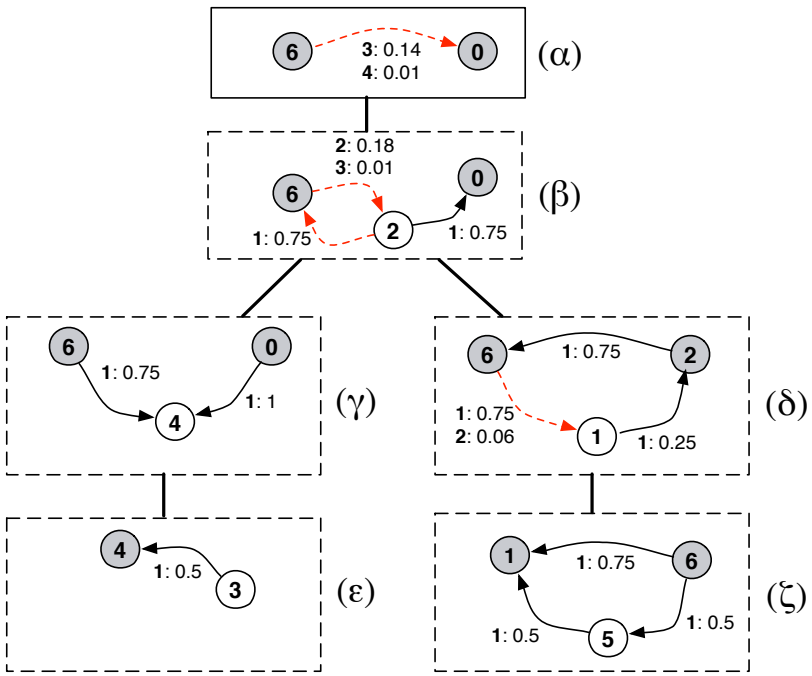
Theorem

There is a first-order query q such that for any unbounded-tw, tw-constructible, arity-2 instance family $\mathcal{I}$, probabilistic query eval for q on $\mathcal{I}$ is #P-hard under RP reductions.

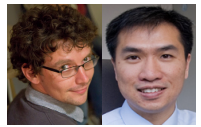
Proven by extracting arbitrary graphs as minors of high-treewidth families using [Chekuri and Chuzhoy 2014]

Application: Efficient querying of uncertain graphs

[BUDA'14: Maniu, Cheng, and Senellart 2014]



- Problem: Optimize query evaluation on probabilistic graphs
- Challenge: Real graph data is not treelike
- Methodology: Build partial tree decompositions and use different query evaluation techniques on treelike parts and on the rest of the data

with 

Outline

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management
- Possible interactions with existing teams at DI:

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management
- Possible interactions with existing teams at DI:

Antique. Static analysis of JS programs for Web data extraction (e.g., [WWW'12: Benedikt, Furche, Savvides, and Senellart 2012])

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management
- Possible interactions with existing teams at DI:

Antique. Static analysis of JS programs for Web data extraction (e.g., [WWW'12: Benedikt, Furche, Savvides, and Senellart 2012])

Data, Sierra. Data science at large, machine learning applications to large, complex, structured data

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management
- Possible interactions with existing teams at DI:

Antique. Static analysis of JS programs for Web data extraction (e.g., [WWW'12: Benedikt, Furche, Savvides, and Senellart 2012])

Data, Sierra. Data science at large, machine learning applications to large, complex, structured data

Talgo. Graph algorithms, approximation algorithms

Integration within the DI: Research

- Introducing a new research area on data management at ENS, forming a new team
- Possible creation of a new Inria project-team on Web data management
- Possible interactions with existing teams at DI:

Antique. Static analysis of JS programs for Web data extraction (e.g., [WWW'12: Benedikt, Furche, Savvides, and Senellart 2012])

Data, Sierra. Data science at large, machine learning applications to large, complex, structured data

Talgo. Graph algorithms, approximation algorithms

- Participate in the Big Data initiatives at the level of PSL

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped
- Also available for:

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped
- Also available for:
  - Student tutoring

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped
- Also available for:
  - Student tutoring
  - Coaching for programming competitions (esp., ACM-ICPC)

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped
- Also available for:
  - Student tutoring
  - Coaching for programming competitions (esp., ACM-ICPC)
  - Helping with the entrance competition

Integration within the DI: Teaching

- Available to teach existing L3 modules as needed: algorithms & programming, formal languages & complexity, operating systems, information theory… and of course databases
- May propose a new M1 course if relevant: Web information retrieval, Web search, big data management, natural language processing, modern Web development…
- Can continue my existing MPRI M2 class on Web data management with Serge Abiteboul, possibly revamped
- Also available for:
  - Student tutoring
  - Coaching for programming competitions (esp., ACM-ICPC)
  - Helping with the entrance competition
  - …

Merci.


Amarilli, A., P. Bourhis, and P. Senellart (July 2015). "Provenance Circuits for Trees and Treelike Instances". In: *Proc. ICALP*. Kyoto, Japan, pp. 56–68.


— (June 2016). "Tractable Lineages on Treelike Instances: Limits and Extensions". In: *Proc. PODS*. San Francisco, USA.


Amarilli, A., S. Maniu, and P. Senellart (Aug. 2015). "Intensional Data on the Web". In: *SIGWEB Newsletter*.


Basu, D., Q. Lin, W. Chen, H. T. Vo, Z. Yuan, P. Senellart, and S. Bressan (2016). "Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning". In: *Transactions on Large-Scale Data and Knowledge-Centered Systems*.


Benedikt, M., T. Furche, A. Savvides, and P. Senellart (Apr. 2012). "ProFoUnd: Program-analysis–based Form Understanding". In: *Proc. WWW*. Demonstration. Lyon, France, pp. 313–316.

- Chekuri, C. and J. Chuzhoy (2014). "Polynomial Bounds for the Grid-Minor Theorem". In: *Proc. STOC*.
- Courcelle, B. (1990). "The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs". In: *Inf. Comput.* 85.1.
- Dalvi, N. and D. Suciu (2012). "The Dichotomy of Probabilistic Inference for Unions of Conjunctive Queries". In: *J. ACM* 59.6.
- Faheem, M. and P. Senellart (Dec. 2015). "Adaptive Web Crawling through Structure-Based Link Classification". In: *Proc. ICADL*. Seoul, South Korea, pp. 39–51.
- Fink, R., A. Hogue, D. Olteanu, and S. Rath (2011). "SPROUT^2: a squared query engine for uncertain web data". In: *SIGMOD*.


Gouriten, G., S. Maniu, and P. Senellart (Sept. 2014). "Scalable, Generic, and Adaptive Systems for Focused Crawling". In: *Proc. Hypertext*. Douglas Engelbart Best Paper Award. Santiago, Chile, pp. 35–45.


Lei, S., S. Maniu, L. Mo, R. Cheng, and P. Senellart (Aug. 2015). "Online Influence Maximization". In: *Proc. KDD*. Sydney, Australia, pp. 645–654.


Maniu, S., R. Cheng, and P. Senellart (June 2014). "ProbTree: A Query-Efficient Representation of Probabilistic Graphs". In: *Proc. BUDA*. Workshop without formal proceedings. Snowbird, USA.


Suchanek, F. M., G. Kasneci, and G. Weikum (2007). "YAGO: A Core of Semantic Knowledge. Unifying WordNet and Wikipedia". In: *WWW*, pp. 697–706. ISBN: 978-1-59593-654-7.