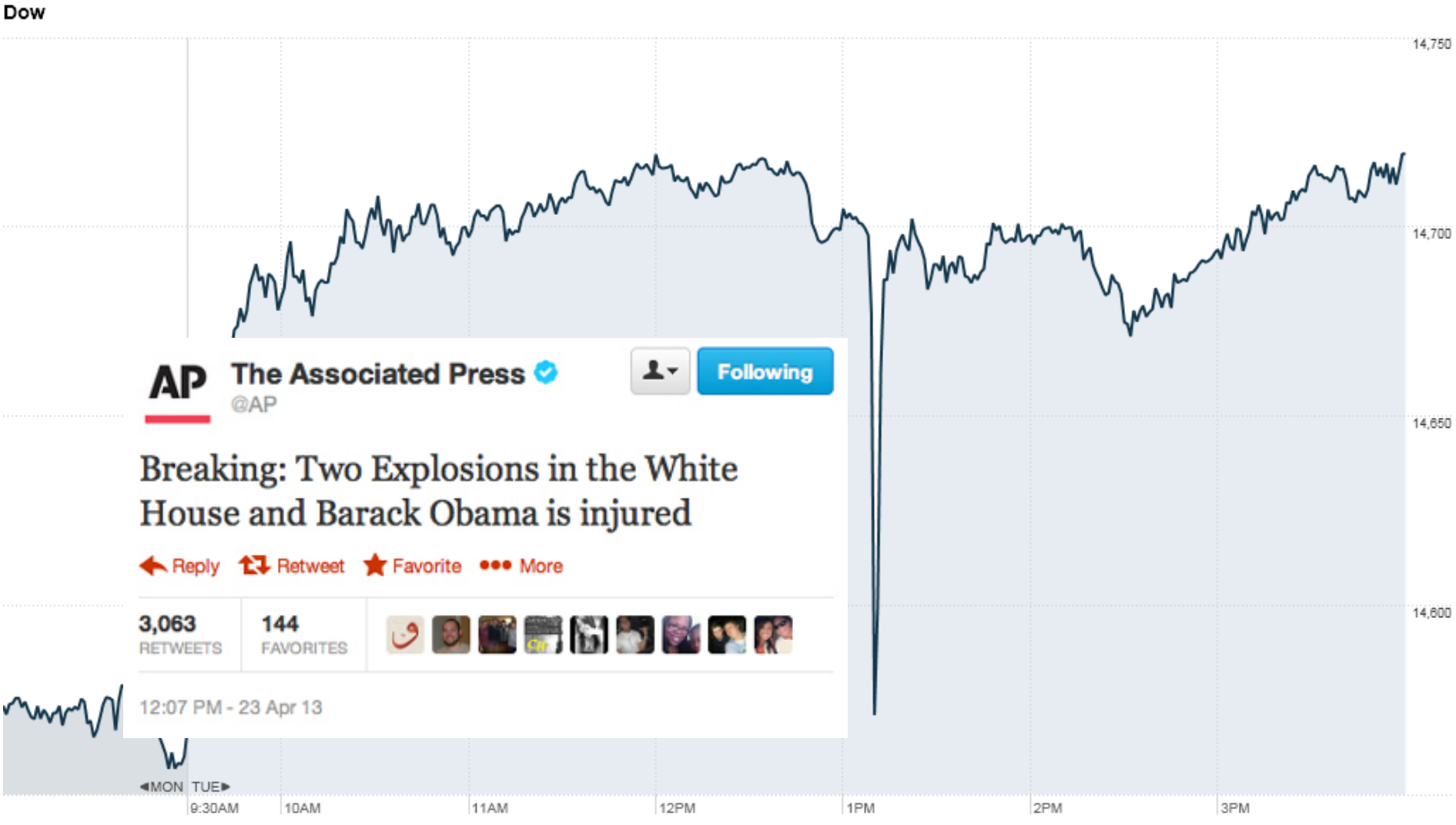
The Veracity of Big Data

Pierre Senellart

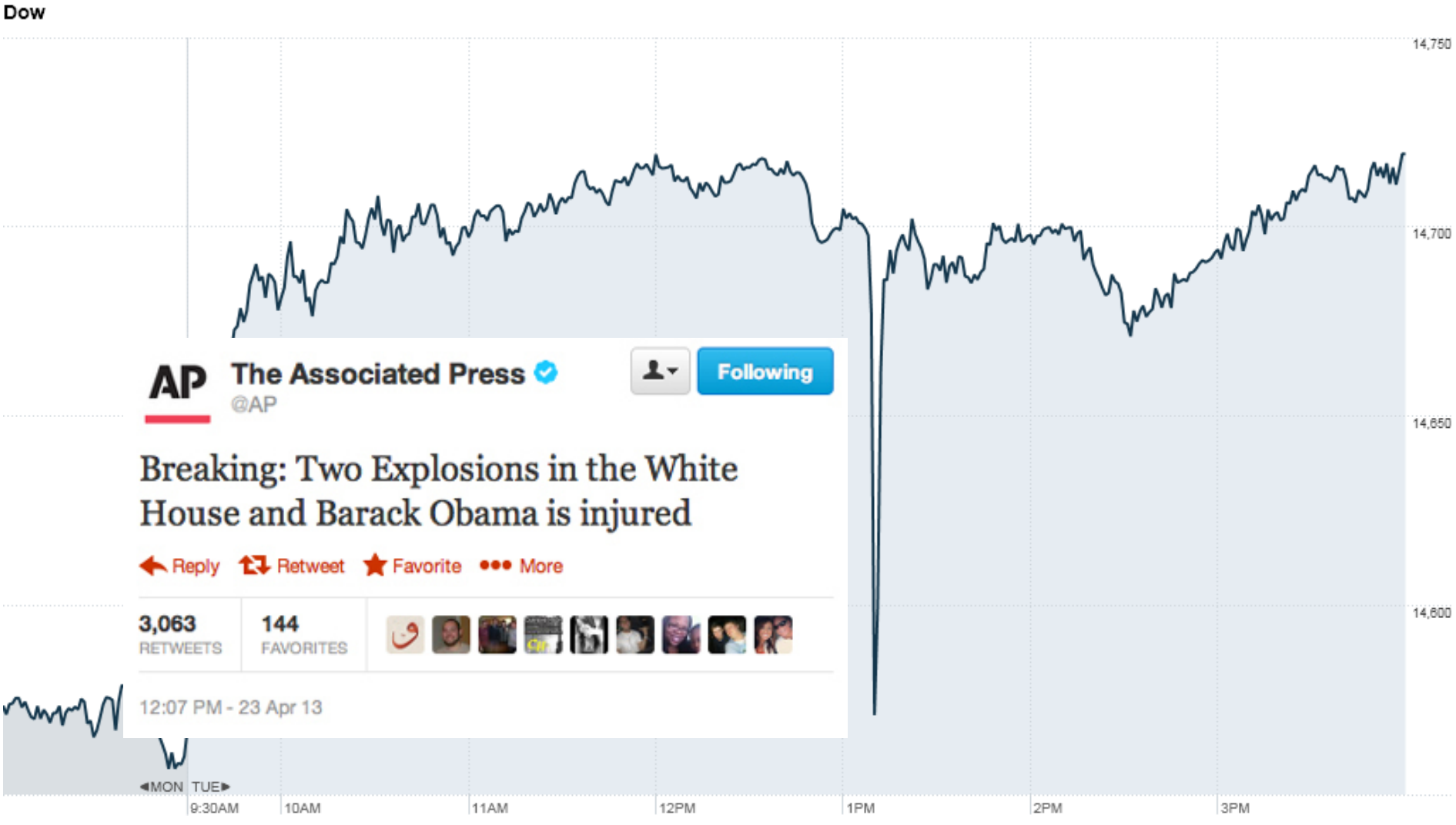
25 May 2016, *AAFD & SFC'16*

23 April 2013, Dow Jones (cnn.com)

23 April 2013, Dow Jones (cnn.com)



23 April 2013, Dow Jones (cnn.com)



- Twitter feed of Associated Press hacked
- Algorithmic trading systems reacting to tweets

The Four Vs of Big Data

The Four Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

The Four Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

The Four Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

The Four Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

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, classification, 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, classification, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

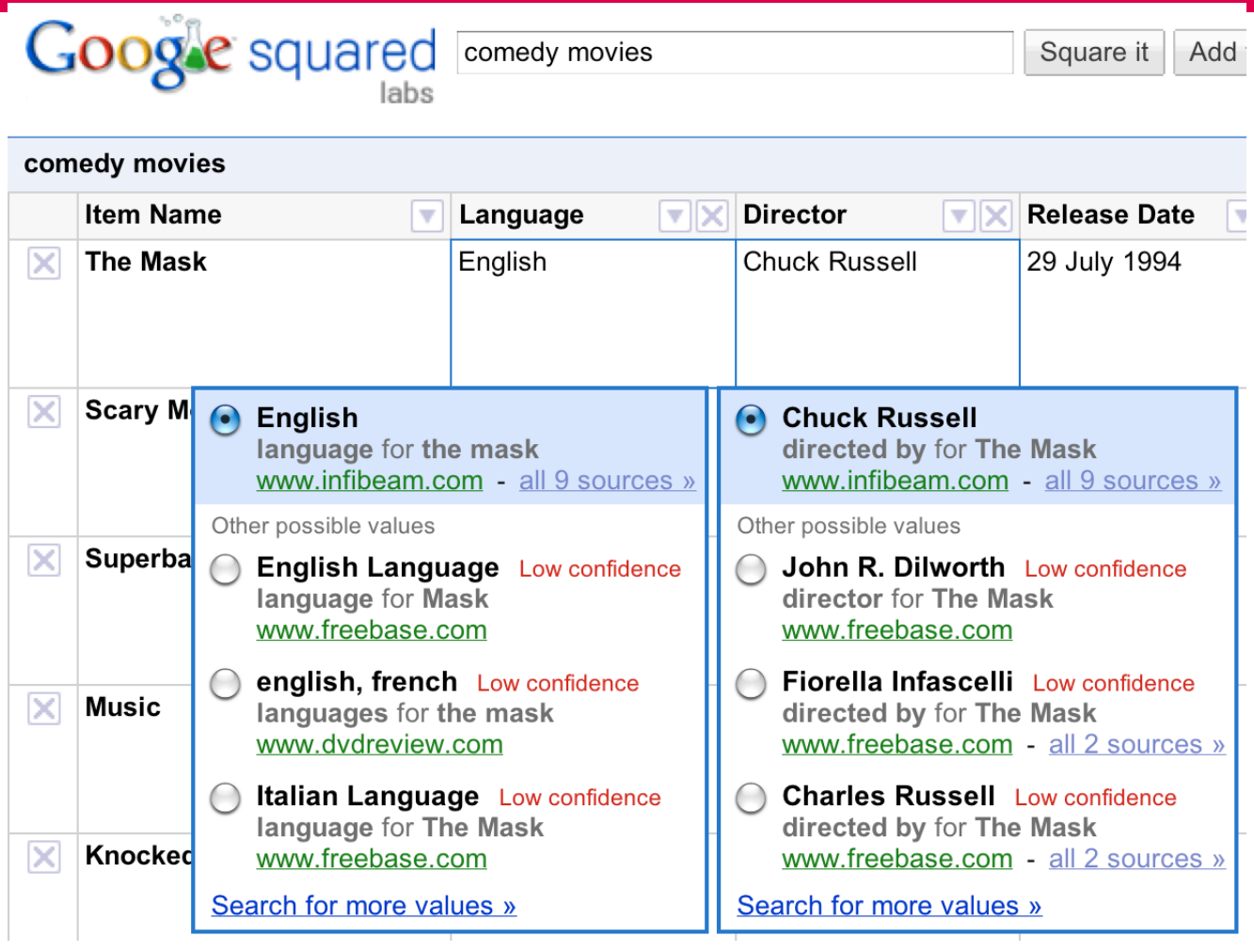
Uncertainty in Web information extraction

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 et al. 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 et al. 2007)

Dealing with Uncertainty

Three main research questions:

- How to estimate the veracity of a source, or of a piece of information? $\Rightarrow$ truth finding

Dealing with Uncertainty

Three main research questions:

- How to estimate the veracity of a source, or of a piece of information? $\Rightarrow$ truth finding
- How to ensure the provenance of a piece of information, to know where it comes from? $\Rightarrow$ provenance management

Dealing with Uncertainty

Three main research questions:

- How to estimate the veracity of a source, or of a piece of information? $\Rightarrow$ truth finding
- How to ensure the provenance of a piece of information, to know where it comes from? $\Rightarrow$ provenance management
- How to efficiently process uncertain data at scale? $\Rightarrow$ probabilistic database management systems

Outline

Outline

Motivating Example

What are the capital cities of European countries?

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia

Voting

Information: redundance

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia
Frequence	**P.** 0.67	**R.** 0.80	**W.** 0.60	Buch. 0.50	Bud. 0.43
	R. 0.33	F. 0.20	K. 0.20	Bud. 0.50	**S.** 0.57
			B. 0.20		

Evaluating Trustworthiness of Sources

Information: redundance, trustworthiness of sources (= average frequence of predicted correctness)

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.60
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.58
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.52
David	Paris	Rome	Bratislava	Budapest	Sofia	0.55
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.51
Fred	Rome	?	?	Budapest	Sofia	0.47
George	Rome	?	?	?	Sofia	0.45
Frequence	**P.** 0.70	**R.** 0.82	**W.** 0.61	**Buch.** 0.53	Bud. 0.46	
weighted	R. 0.30	F. 0.18	K. 0.19	Bud. 0.47	**S.** 0.54	
by trust			B 0.20			

Iterative Fixpoint Computation

Information: redundance, trustworthiness of sources with iterative fixpoint computation

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.65
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.63
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.57
David	Paris	Rome	Bratislava	Budapest	Sofia	0.54
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.49
Fred	Rome	?	?	Budapest	Sofia	0.39
George	Rome	?	?	?	Sofia	0.37
Frequence	**P.** 0.75	**R.** 0.83	**W.** 0.62	**Buch.** 0.57	**Bud.** 0.51	
weighted	R. 0.25	F. 0.17	K. 0.20	Bud. 0.43	S. 0.49	
by trust			B 0.19			

Context and problem

- Context:
  - Set of sources stating facts
  - (Possible) functional dependencies between facts
  - Fully unsupervised setting: we do not assume any information on truth values of facts or inherent trust in sources
- Problem: determine which facts are true and which facts are false
- Real world applications: query answering, source selection, data quality assessment on the web, making good use of the wisdom of crowds

Outline

Outline

General Model

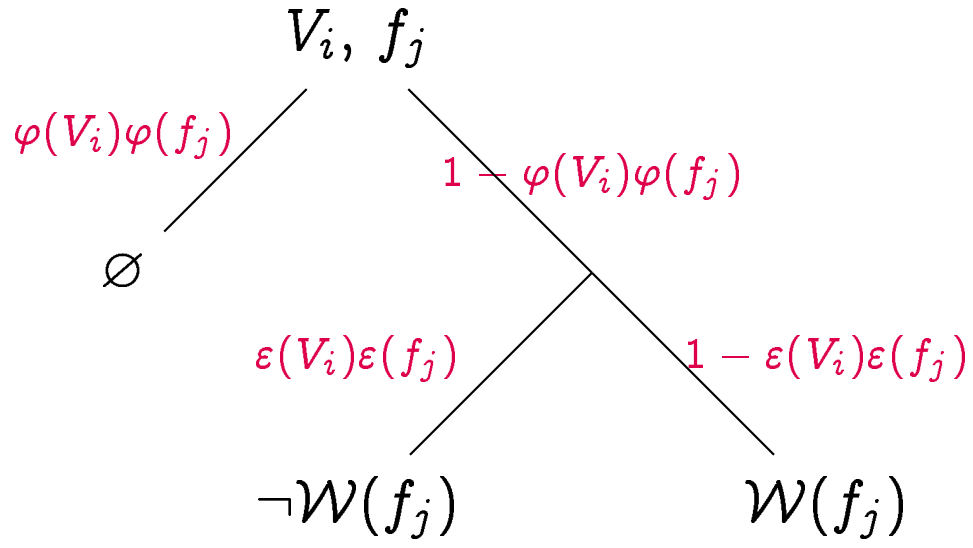
- Set of facts $\mathcal{F} = \{f_1 ... f_n\}$
  - Examples: "Paris is capital of France", "Rome is capital of France", "Rome is capital of Italy"
- Set of views (= sources) $\mathcal{V} = \{V_1 ... V_m\}$, where a view is a partial mapping from $\mathcal{F}$ to $\{\mathrm{T}, \mathrm{F}\}$
  - Example:

    $\neg$ "Paris is capital of France" $\wedge$ "Rome is capital of France"
- Objective: find the most likely real world $\mathcal{W}$ given $\mathcal{V}$ where the real world is a total mapping from $\mathcal{F}$ to $\{\mathrm{T}, \mathrm{F}\}$
  - Example:

    "Paris is capital of France" $\wedge \neg$ "Rome is capital of France" $\wedge$ "Rome is capital of Italy" $\wedge$ ...

Generative Probabilistic Model

(Galland et al. 2010)



- $\varphi(V_i)\varphi(f_j)$: probability that V_i "forgets" f_j
- $\varepsilon(V_i)\varepsilon(f_j)$: probability that V_i "makes an error" on f_j
- Number of parameters: $n + 2(n+m)$
- Size of data: $\tilde{\varphi}nm$ with $\tilde{\varphi}$ the average forget rate

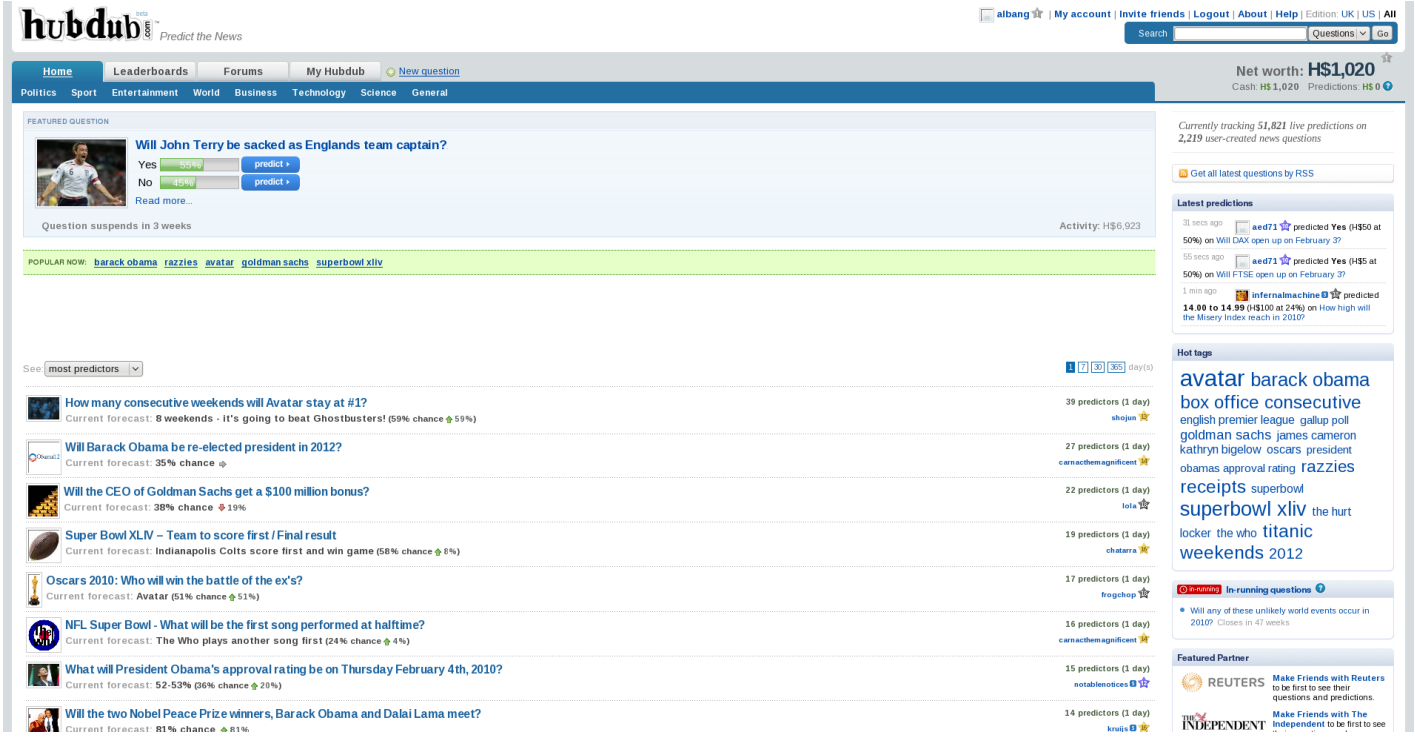
Obvious Approach

- Method: use this generative model to find the most likely parameters given the data
  - Inverse the generative model to compute the probability of a set of parameters given the data
- Not practically applicable:
  - Non-linearity of the model and boolean parameter $\mathcal{W}(f_j)$
    $\Rightarrow$ equations for inversing the generative model very complex
  - Large number of parameters (n and m can both be quite large) $\Rightarrow$ Any exponential technique unpractical

$\Rightarrow$ Heuristic fix-point algorithms (many proposed ones!)

Outline

Hubdub (1/2)



http://www.hubdub.com/

- 357 questions, 1 to 20 answers, 473 participants

Hubdub (2/2)

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	278	292
Counting	340	327
TruthFinder (Yin et al. 2007)	458	274
3-Estimates (Galland et al. 2010)	272	270

General-Knowledge Quiz (1/2)

1. **Where is the city of Ushuaia located?**
   - (•) Don't know
   - ( ) In Italy
   - ( ) In Greece
   - ( ) In Argentina
   - ( ) In the Ivory Coast
   - ( ) In Sweden
   - ( ) In Malaysia

2. **What is the last word of all three parts of Dante's *Divine Comedy* (*Hell* — *Purgatory* — *Paradise*)?**
   - (•) Don't know
   - ( ) "Stars" ("Stelle")
   - ( ) "God" ("Dio")
   - ( ) "Hope" ("Speranza")
   - ( ) "Beatrice"

3. **Who discovered the planet Uranus?**
   - (•) Don't know
   - ( ) Sir William Herschel (in 1781)
   - ( ) Urbain Le Verrier (in 1846)
   - ( ) Clyde Tombaugh (in 1930)
   - ( ) Percival Lowell (in 1894)

`http://www.madore.org/~david/quizz/quizz1.html`

- 17 questions, 4 to 14 answers, 601 participants

General-Knowledge Quiz (2/2)

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	11	6
Counting	12	6
TruthFinder (Yin et al. 2007)	78	77
3-Estimates (Galland et al. 2010)	9	0

Many variations...

Modeling of various real-world phenomena:

- Sources copying each other (Dong et al. 2010)
- Complex source dependencies (Pochampally et al. 2014)
- Similarity between attribute values (Yin et al. 2008)
- Correlated group of attributes (Ba et al. 2015)
- Heterogeneous data types (Q. Li et al. 2014)
- ...

See extensive evaluations of different techniques (X. Li et al. 2012; Waguih and Berti-Equille 2014). General problem far from being solved!

Outline

Outline

Different types of uncertainty

Two dimensions:

- Different types:
  - Unknown value: NULL in an RDBMS
  - Alternative between several possibilities: either A or B or C
  - Imprecision on a numeric value: a sensor gives a value that is an approximation of the actual value
  - Confidence in a fact as a whole: cf. information extraction
  - Structural uncertainty: the schema of the data itself is uncertain
- Qualitative (NULL) or Quantitative (95%, low-confidence, etc.) uncertainty

Managing uncertainty

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Managing uncertainty

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Represent all different forms of uncertainty
- Use probabilities to represent quantitative information on the confidence in the data
- Query data and retrieve uncertain results
- Allow adding, deleting, modifying data in an uncertain way
- Bonus (if possible): Keep as well lineage/provenance information, so as to ensure traceability

Why probabilities?

- Not the only option: fuzzy set theory (Galindo et al. 2005), Dempster-Shafer theory (Zadeh 1986)
- Mathematically rich theory, nice semantics with respect to traditional database operations (e.g., joins)
- Some applications already generate probabilities (e.g., statistical information extraction or natural language probabilities)
- In other cases, we "cheat" and pretend that (normalized) confidence scores are probabilities: see this as a first-order approximation

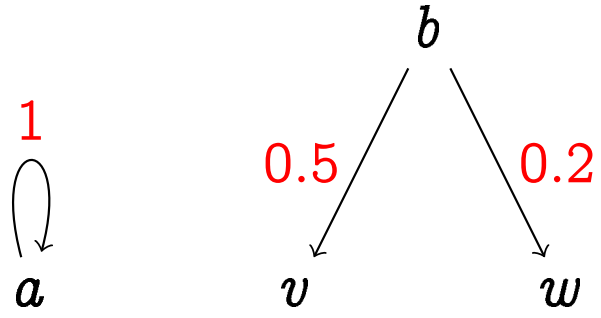
Outline

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2

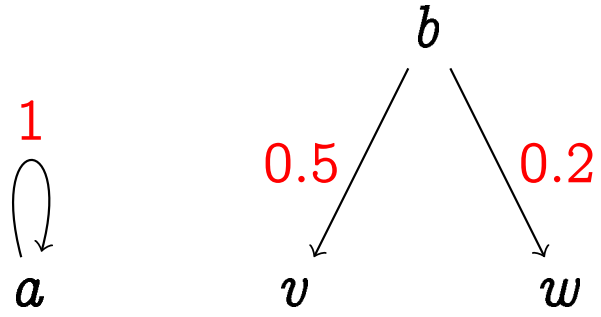
Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



Tuple-independent databases (TID)

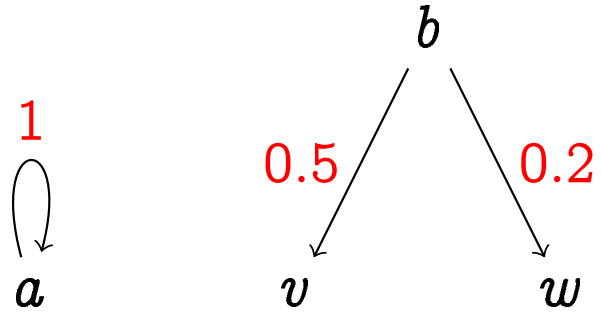
S		
a	a	1
b	v	0.5
b	w	0.2



This TID instance represents the following probability distribution:

Tuple-independent databases (TID)

S		
a	a	1
b	v	0.5
b	w	0.2



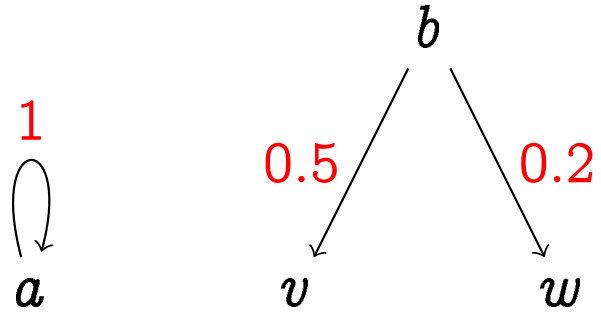
This TID instance represents the following probability distribution:

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



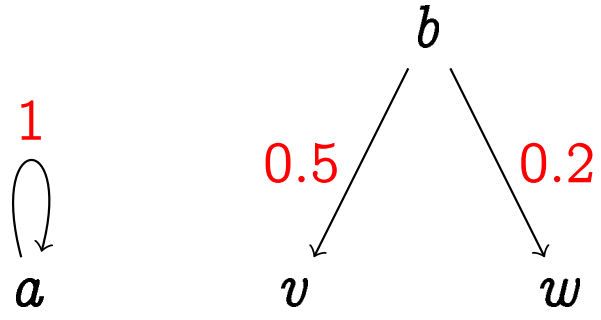
This TID instance represents the following probability distribution:

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:

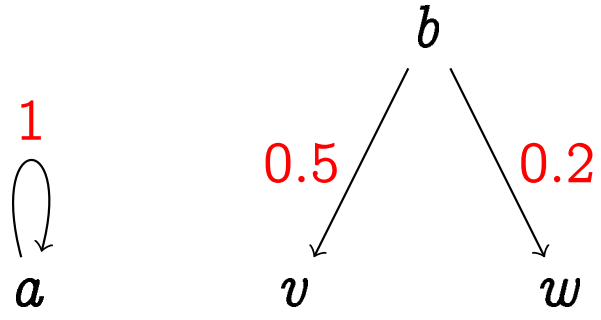
0.5×0.2	
S	
a	a
b	v
b	w

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

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!

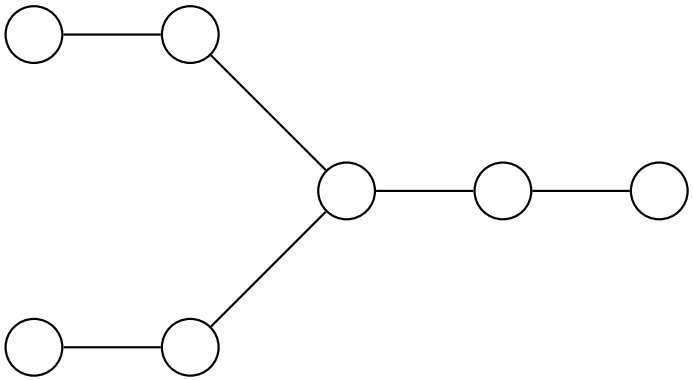
Is there a smaller class $\mathcal{I}$ such that PQE is tractable for a larger $\mathcal{Q}$?

Trees and treelike instances

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

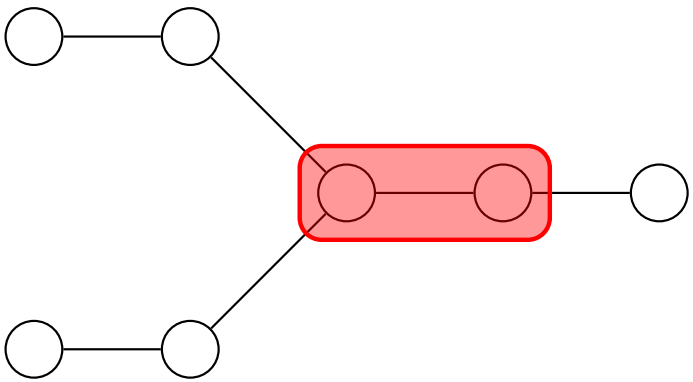
Trees and treelike instances

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



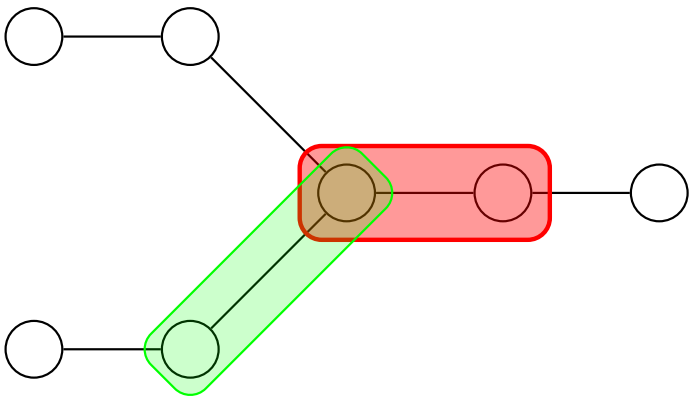
Trees and treelike instances

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



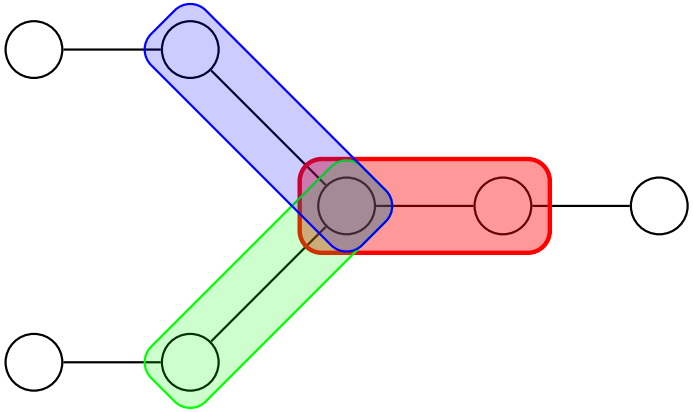
Trees and treelike instances

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



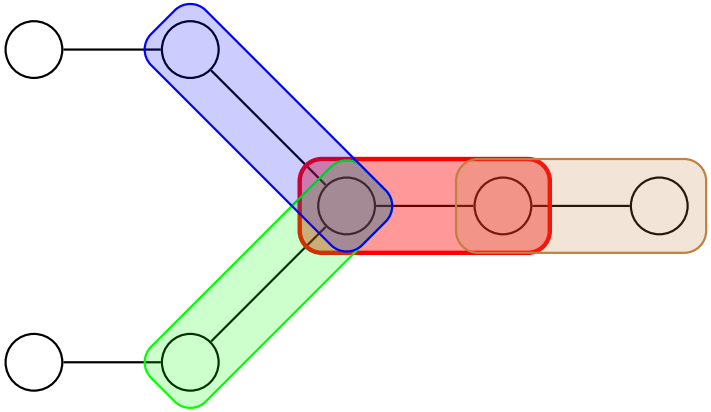
Trees and treelike instances

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



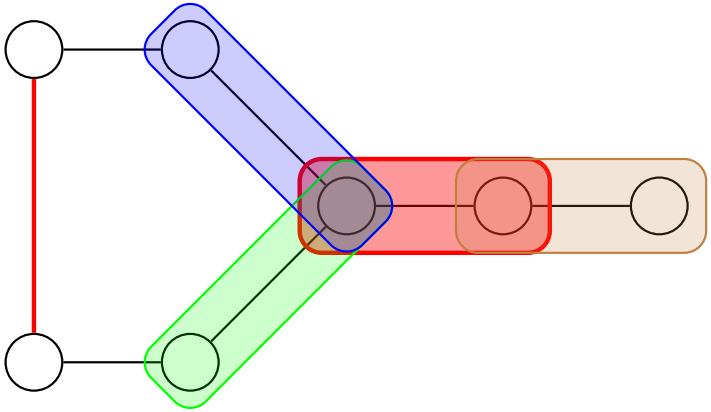
Trees and treelike instances

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



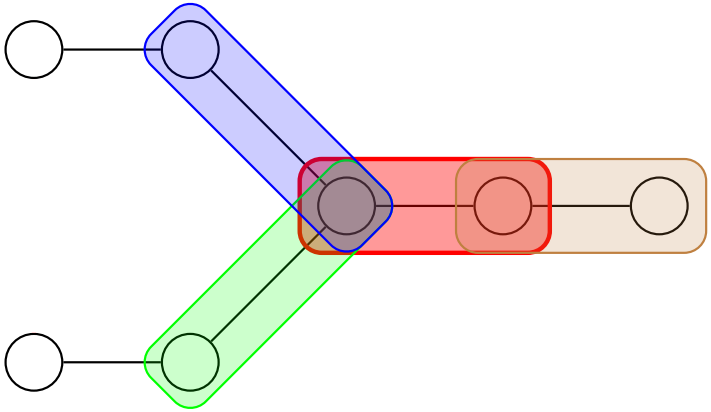
Trees and treelike instances

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



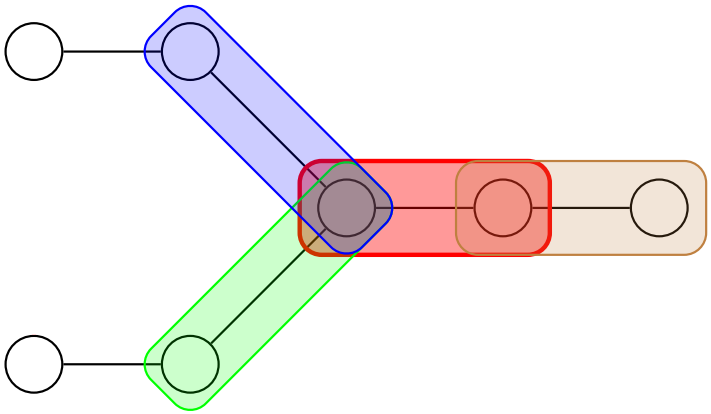
Trees and treelike instances

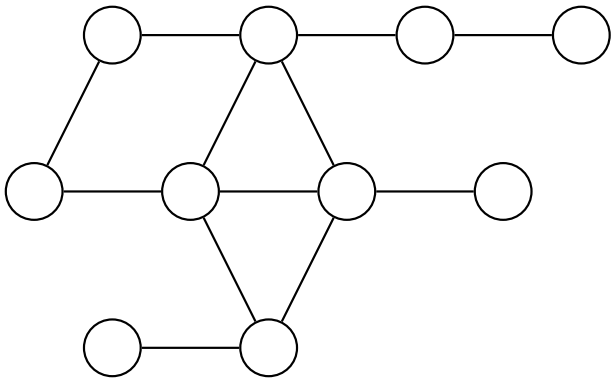
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)



Trees and treelike instances

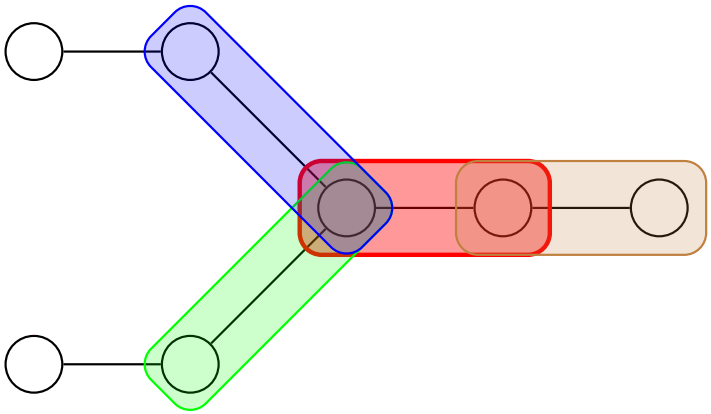
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

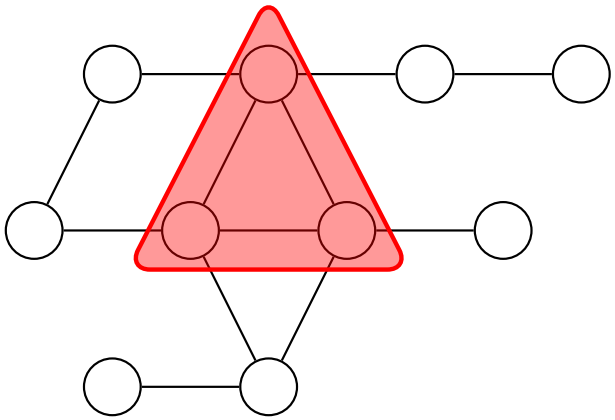




Trees and treelike instances

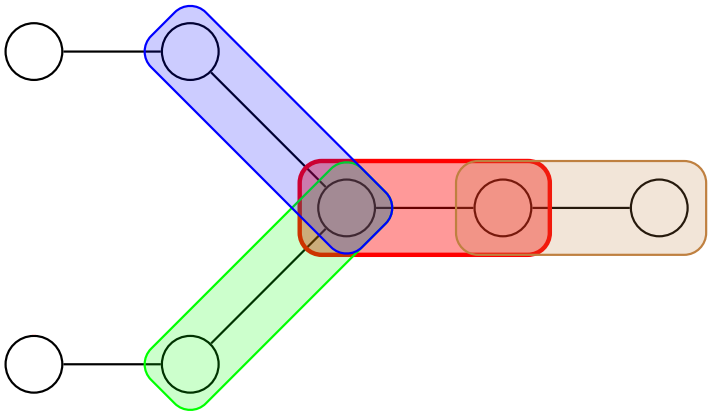
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

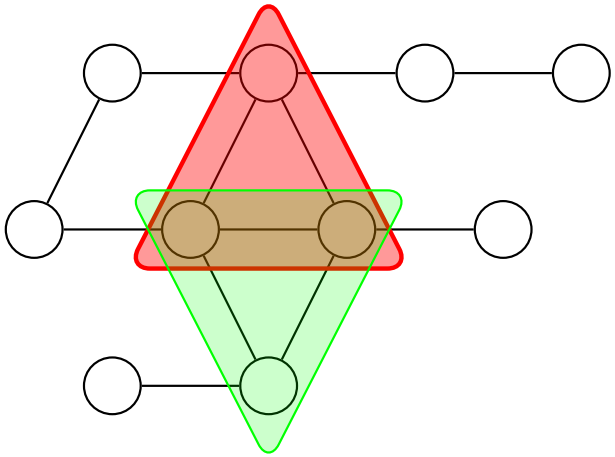




Trees and treelike instances

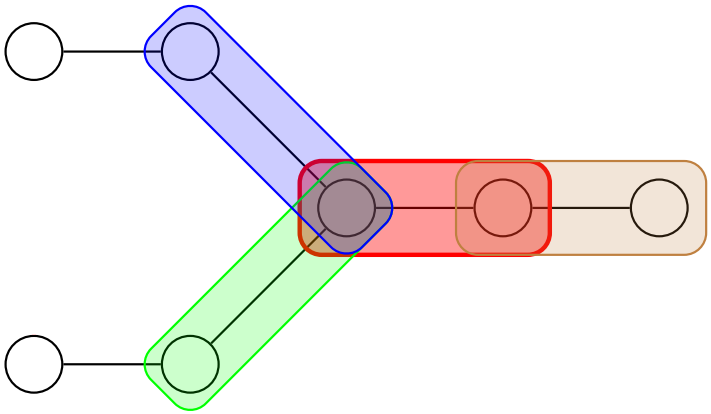
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

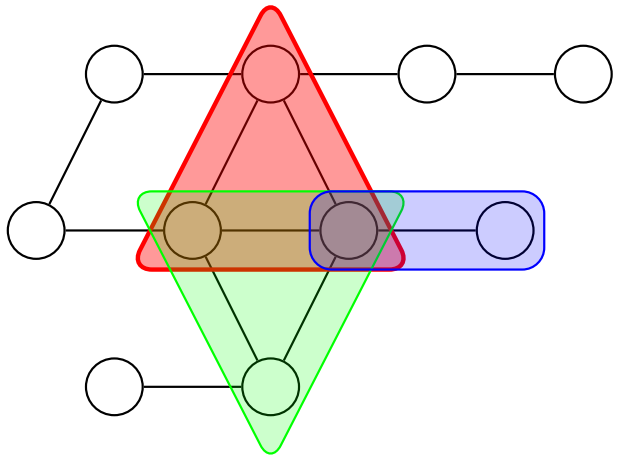




Trees and treelike instances

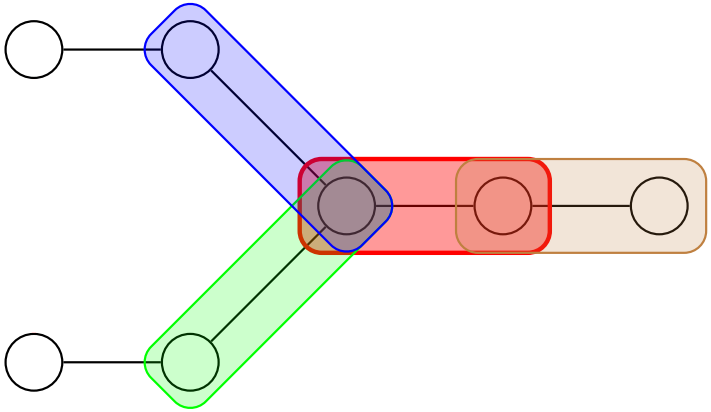
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

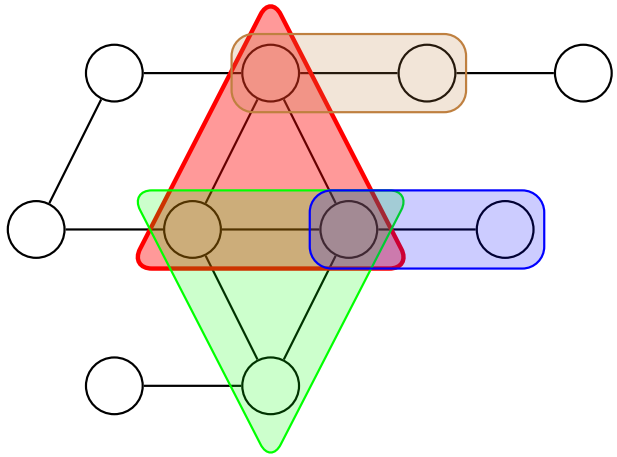




Trees and treelike instances

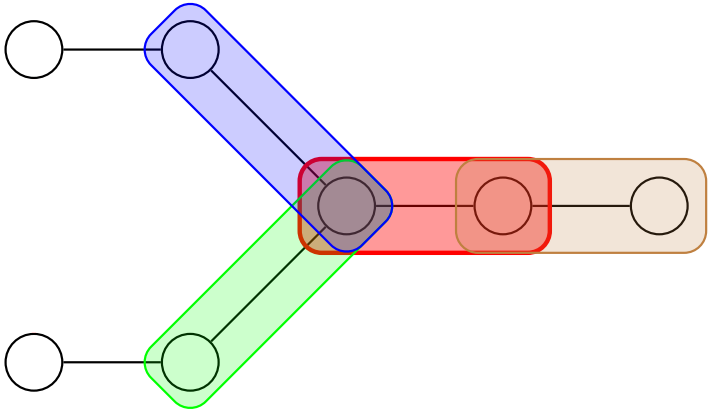
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

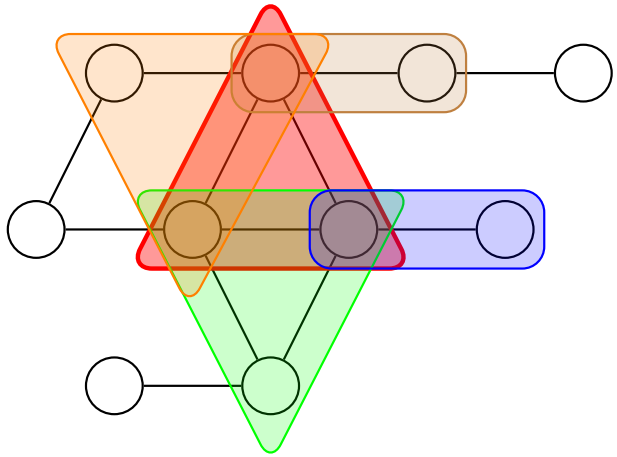




Trees and treelike instances

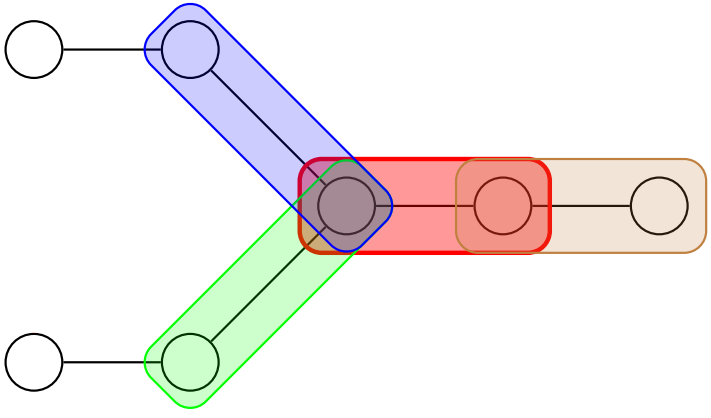
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

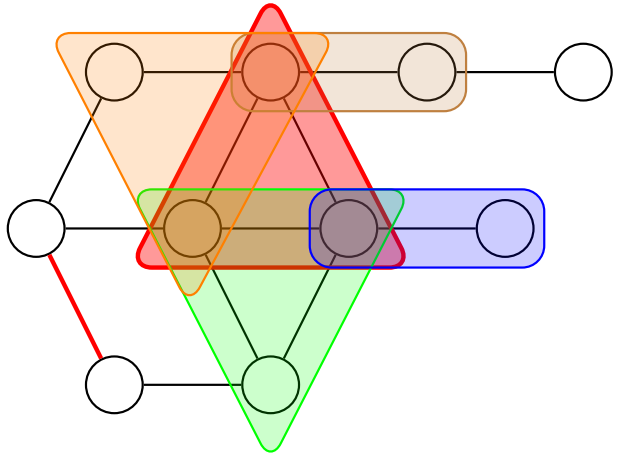




Trees and treelike instances

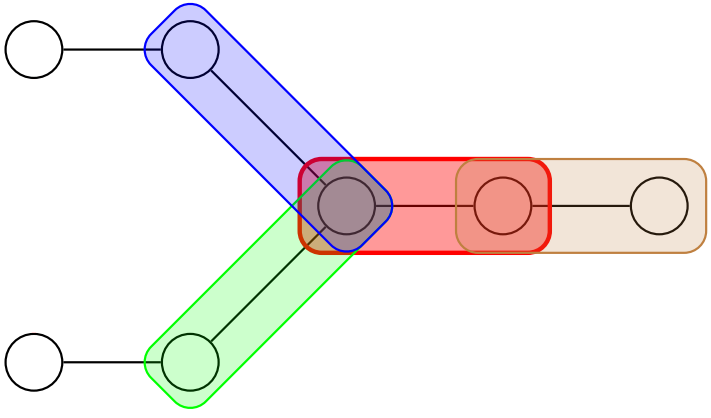
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

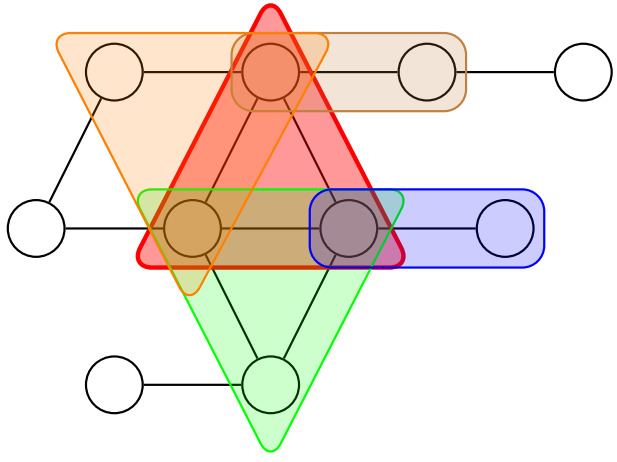




Trees and treelike instances

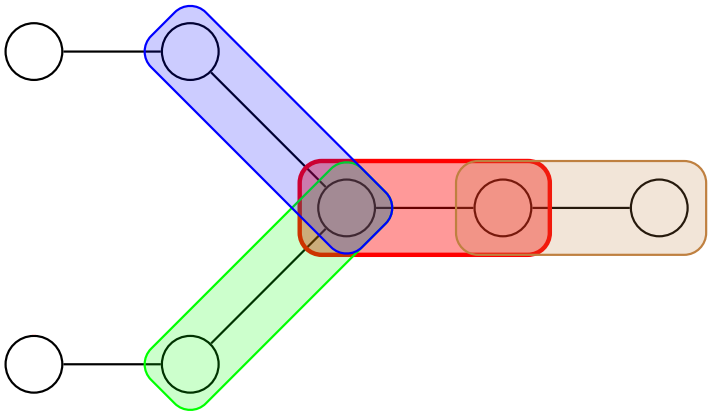
- Idea: let $\mathcal{I}$ be treelike instances (constant bound on treewidth)

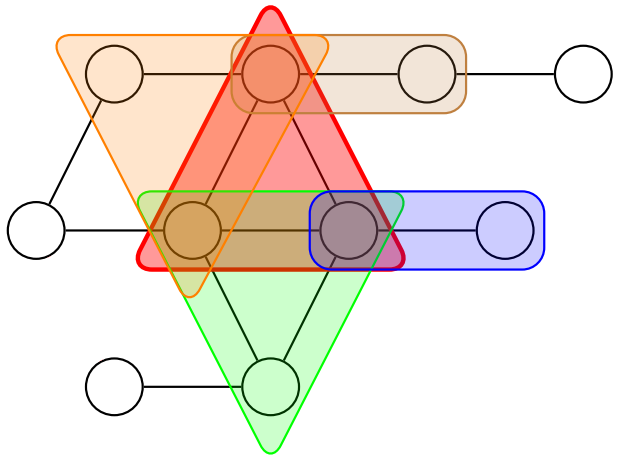




Trees and treelike instances

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

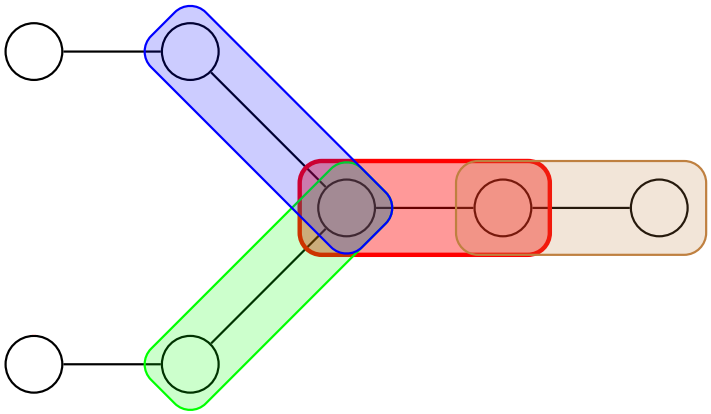


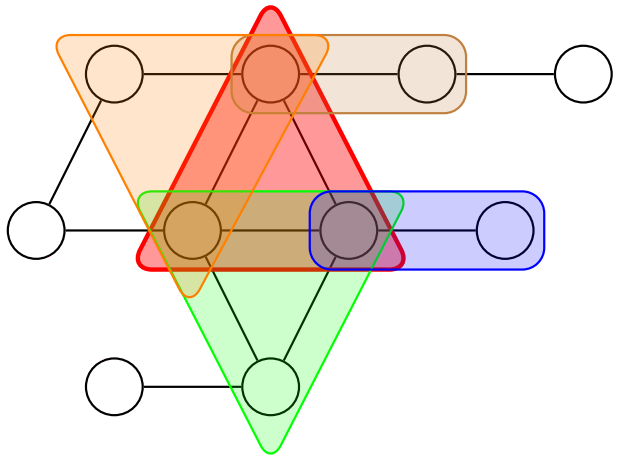


  - Trees have treewidth 1
  - Cycles have treewidth 2
  - k-cliques and $(k-1)$-grids have treewidth $k-1$

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 $k-1$

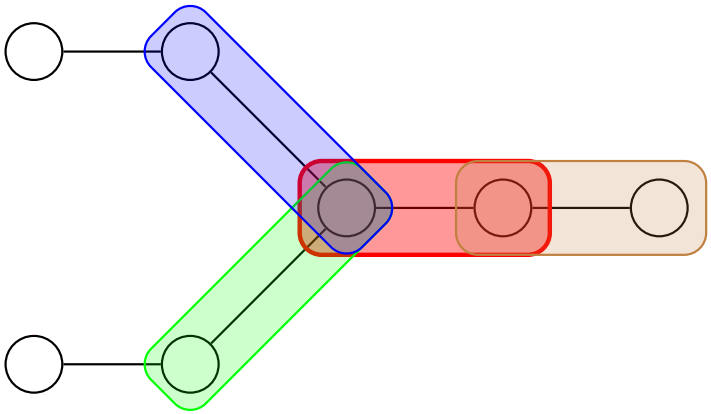
$\rightarrow$ Known results (Courcelle 1990):

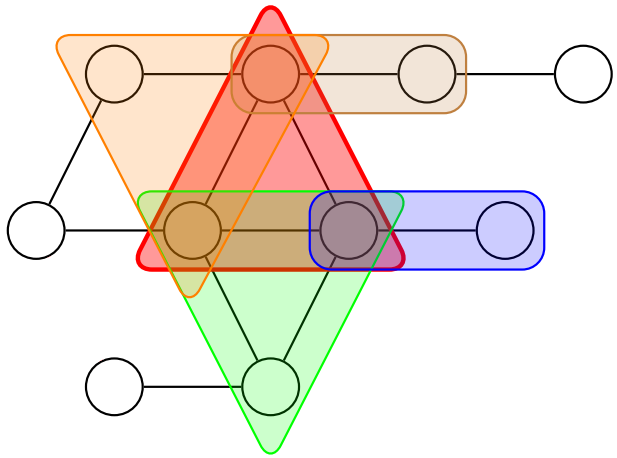
  - $\mathcal{I}$: treelike instances; $\mathcal{Q}$: monadic second-order queries

  $\rightarrow$ 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. (Amarilli et al. 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. (Amarilli et al. 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

Dichotomy for PQE

An instance-based dichotomy result:

Upper bound. (Amarilli et al. 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

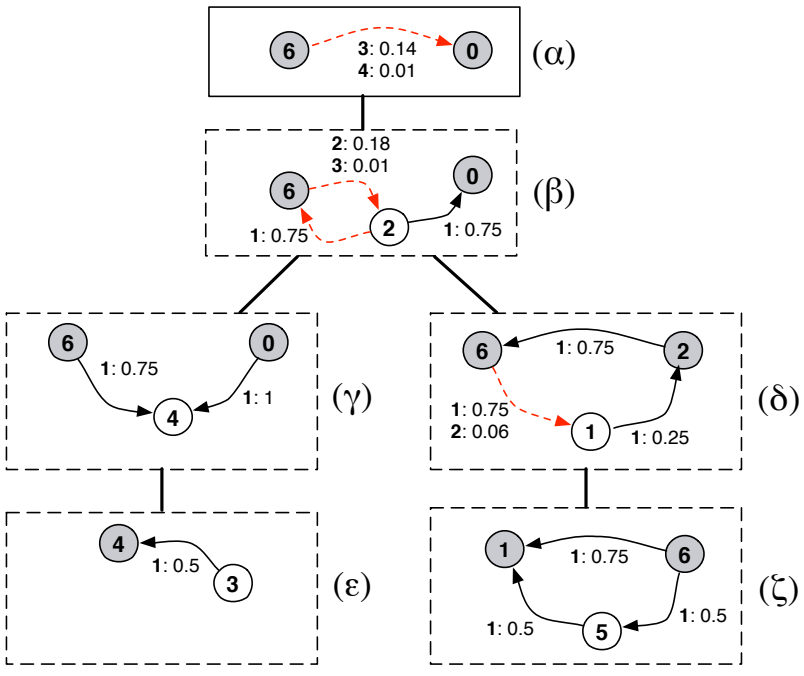
Lower bound. (Amarilli et al. 2016)

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

$\rightarrow$ PQE is #P-hard under RP reductions assuming:

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

Application: Efficient querying of uncertain graphs (Maniu et al. 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

Outline

Conclusion

- The real world is uncertain
- Tools we use to process the real world introduce uncertainty
- Need for principled methods to:
  - Estimate uncertainty (veracity, truthfulness...) of information
  - Properly manage the confidence (probability, level of certainty...) in the information
  - Keep information on the provenance of data

Conclusion

- The real world is uncertain
- Tools we use to process the real world introduce uncertainty
- Need for principled methods to:
  - Estimate uncertainty (veracity, truthfulness...) of information
  - Properly manage the confidence (probability, level of certainty...) in the information
  - Keep information on the provenance of data

Merci.