



# The Veracity of Big Data

Pierre Senellart



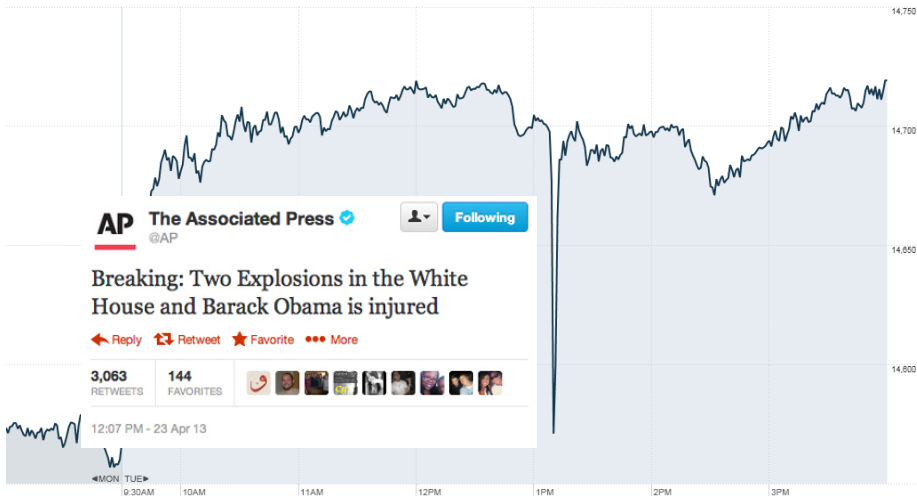
# 23 April 2013, Dow Jones (cnn.com)

Dow



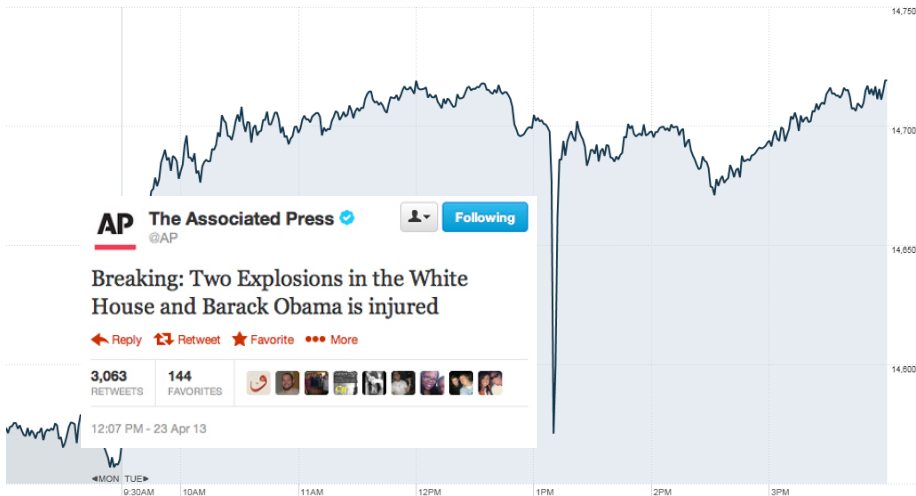
23 April 2013, Dow Jones (cnn.com)

Dow



23 April 2013, Dow Jones (cnn.com)

Dow



- 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
<u>arabic, egypt</u>	406	08-sep-2011	(Seed) 100.0
<u>chinese, republic of china</u>	439	24-oct-2011	100.0
<u>chinese, singapore</u>	421	21-sep-2011	(Seed) 100.0
<u>english, britain</u>	439	24-oct-2011	100.0
<u>english, canada</u>	439	24-oct-2011	(Seed) 100.0
<u>english, england001</u>	439	24-oct-2011	100.0
<u>arabic, morocco</u>	422	23-sep-2011	100.0
<u>cantonese, hong kong</u>	406	08-sep-2011	100.0
<u>english, uk</u>	436	19-oct-2011	100.0
<u>english, south vietnam</u>	427	27-sep-2011	99.9
<u>french, morocco</u>	422	23-sep-2011	99.9
<u>greek, turkey</u>	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 labs

comedy movies

Square it Add

Item Name	Language	Director	Release Date
<input type="checkbox"/> The Mask	English	Chuck Russell	29 July 1994
<input type="checkbox"/> Scary M	<input checked="" type="radio"/> English language for the mask <a href="http://www.infibeam.com">www.infibeam.com</a> - all 9 sources » Other possible values	<input checked="" type="radio"/> Chuck Russell directed by for The Mask <a href="http://www.infibeam.com">www.infibeam.com</a> - all 9 sources » Other possible values	
<input type="checkbox"/> Superba	<input type="radio"/> English Language Low confidence language for Mask <a href="http://www.freebase.com">www.freebase.com</a>	<input type="radio"/> John R. Dilworth Low confidence director for The Mask <a href="http://www.freebase.com">www.freebase.com</a>	
<input type="checkbox"/> Music	<input type="radio"/> english, french Low confidence languages for the mask <a href="http://www.dvdreview.com">www.dvdreview.com</a>	<input type="radio"/> Fiorella Infascelli Low confidence directed by for The Mask <a href="http://www.freebase.com">www.freebase.com</a> - all 2 sources »	
<input type="checkbox"/> Knocked	<input type="radio"/> Italian Language Low confidence language for The Mask <a href="http://www.freebase.com">www.freebase.com</a> Search for more values »	<input type="radio"/> Charles Russell Low confidence directed by for The Mask <a href="http://www.freebase.com">www.freebase.com</a> - all 2 sources » Search for more values »	

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? ⇒ **truth finding**
- How to ensure the **provenance** of a piece of information, to know where it comes from? ⇒ **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

Introduction

Truth Finding

Setting

Model

Experiments

Probabilistic Databases

Conclusion



# Outline

Introduction

**Truth Finding**

Setting

Model

Experiments

Probabilistic Databases

Conclusion



## 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
<b>Frequence</b>	<b>P.</b> 0.67 R. 0.33	<b>R.</b> 0.80 F. 0.20	<b>W.</b> 0.60 K. 0.20 B. 0.20	<b>Buch.</b> 0.50 <b>Bud.</b> 0.50	Bud. 0.43 <b>S.</b> 0.57



# Evaluating Trustworthiness of Sources

**Information:** redundance, trustworthiness of sources (= average frequency 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
Frequency weighted by trust	<b>P.</b> 0.70 R. 0.30	<b>R.</b> 0.82 F. 0.18	<b>W.</b> 0.61 K. 0.19 B 0.20	<b>Buch.</b> 0.53 Bud. 0.47	Bud. 0.46 <b>S.</b> 0.54	



# 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
<b>Frequence weighted by trust</b>	<b>P.</b> 0.75 R. 0.25	<b>R.</b> 0.83 F. 0.17	<b>W.</b> 0.62 K. 0.20 B 0.19	<b>Buch.</b> 0.57 Bud. 0.43	<b>Bud.</b> 0.51 S. 0.49	



# 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

## Introduction

## Truth Finding

- Setting

- Model

- Experiments

## Probabilistic Databases

- Uncertainty Management

- Querying Probabilistic Databases

## Conclusion



# Outline

Introduction

**Truth Finding**

Setting

**Model**

Experiments

Probabilistic Databases

Conclusion

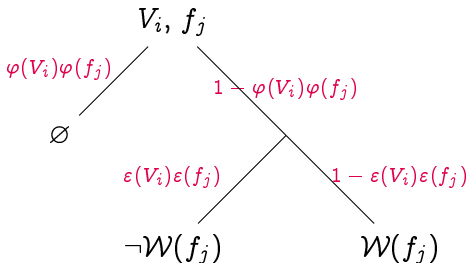


## General Model

- Set of facts  $\mathcal{F} = \{f_1 \dots f_n\}$ 
  - Examples: “Paris is capital of France”, “Rome is capital of France”, “Rome is capital of Italy”
- Set of views (= sources)  $\mathcal{V} = \{V_1 \dots V_m\}$ , where a view is a partial mapping from  $\mathcal{F}$  to  $\{T, F\}$ 
  - Example:
    - “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  $\{T, F\}$ 
  - 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**  $\mathcal{W}(f_j)$   
⇒ equations for inverting the generative model very complex
    - **Large number of parameters** ( $n$  and  $m$  can both be quite large) ⇒  
Any exponential technique unpractical
- ⇒ Heuristic fix-point algorithms (many proposed ones!)



# Outline

Introduction

**Truth Finding**

Setting

Model

**Experiments**

Probabilistic Databases

Conclusion

# Hubdub (1/2)

The screenshot shows the Hubdub website interface. At the top, there's a navigation bar with 'Home', 'Leaderboards', 'Forums', and 'My Hubdub'. Below that, a 'Featured Questions' section displays a question: 'Will John Terry be sacked as England team captain?' with 'Yes' and 'No' options and a 'Read more...' link. A 'POPULARITY' bar lists categories like 'Soccer', 'Politics', 'Business', etc. A 'See most predictions' dropdown is visible. The main content area lists several prediction questions with their current forecasts and chances, such as 'How many consecutive weekends will Avatar stay at #1?' and 'Will Barack Obama be re-elected president in 2012?'. A sidebar on the right shows 'Net worth: \$1,020', 'Currently tracking 51,822 live predictions on 3,129 user-created trivia questions', and 'Latest predictions' with a 'Hot topic' section for 'Avatar Barack Obama box office consecutive english premier bagga gallop and godman sachs james cameron kadhwan dagloww cackes presicion obamas approval rating 1822395 receipts superbowl superbowl xliiv the hart locker the site titanic weekends 2012'. At the bottom of the sidebar, there are sections for 'Featured Partner' including Reuters and Entertainment Weekly.

<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	458	274
(Yin et al. 2007)		
3-Estimates	272	270
(Galland et al. 2010)		





# 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	78	77
(Yin et al. 2007)		
3-Estimates	9	0
(Galland et al. 2010)		



## 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

Introduction

Truth Finding

**Probabilistic Databases**

Uncertainty Management

Querying Probabilistic Databases

Conclusion



Introduction

Truth Finding

**Probabilistic Databases**

Uncertainty Management

Querying Probabilistic Databases

Conclusion



# 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

Introduction

Truth Finding

**Probabilistic Databases**

Uncertainty Management

Querying Probabilistic Databases

Conclusion



# Tuple-independent databases (TID)

**S**

---

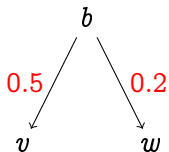
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2

---



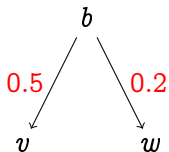
# Tuple-independent databases (TID)

<b>S</b>		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



# Tuple-independent databases (TID)

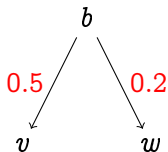
<b>S</b>		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



This TID instance represents the following **probability distribution**:

# Tuple-independent databases (TID)

S		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



This TID instance represents the following **probability distribution**:

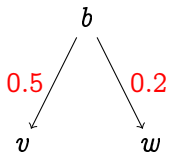
$$0.5 \times 0.2$$

S	
<i>a</i>	<i>a</i>
<i>b</i>	<i>v</i>
<i>b</i>	<i>w</i>



# Tuple-independent databases (TID)

<b>S</b>		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



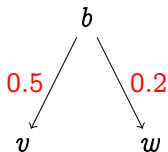
This TID instance represents the following **probability distribution**:

$0.5 \times 0.2$	$0.5 \times (1 - 0.2)$
<b>S</b>	<b>S</b>
<i>a</i> <i>a</i>	<i>a</i> <i>a</i>
<i>b</i> <i>v</i>	<i>b</i> <i>v</i>
<i>b</i> <i>w</i>	



# Tuple-independent databases (TID)

S		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



This TID instance represents the following **probability distribution**:

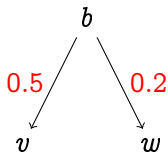
$0.5 \times 0.2$	$0.5 \times (1 - 0.2)$	$(1 - 0.5) \times 0.2$
S	S	S
<i>a</i> <i>a</i>	<i>a</i> <i>a</i>	<i>a</i> <i>a</i>
<i>b</i> <i>v</i>	<i>b</i> <i>v</i>	
<i>b</i> <i>w</i>		<i>b</i> <i>w</i>





# Tuple-independent databases (TID)

S		
<i>a</i>	<i>a</i>	1
<i>b</i>	<i>v</i>	0.5
<i>b</i>	<i>w</i>	0.2



This TID instance represents the following **probability distribution**:

$0.5 \times 0.2$	$0.5 \times (1 - 0.2)$	$(1 - 0.5) \times 0.2$	$(1 - 0.5) \times (1 - 0.2)$
S	S	S	S
<i>a</i> <i>a</i>	<i>a</i> <i>a</i>	<i>a</i> <i>a</i>	<i>a</i> <i>a</i>
<i>b</i> <i>v</i>	<i>b</i> <i>v</i>		
<i>b</i> <i>w</i>		<i>b</i> <i>w</i>	



## 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	
<i>a</i>	1
<i>b</i>	0.4
<i>c</i>	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		S		
<i>a</i>	1	<i>a</i>	<i>a</i>	1
<i>b</i>	0.4	<i>b</i>	<i>v</i>	0.5
<i>c</i>	0.6	<i>b</i>	<i>w</i>	0.2



## Query evaluation on probabilistic instances

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

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

<b>R</b>	<b>S</b>	<b>T</b>
<i>a</i> 1	<i>a a</i> 1	<i>v</i> 0.3
<i>b</i> 0.4	<i>b v</i> 0.5	<i>w</i> 0.7
<i>c</i> 0.6	<i>b w</i> 0.2	<i>b</i> 1



## Query evaluation on probabilistic instances

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

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

<b>R</b>	<b>S</b>	<b>T</b>
<i>a</i> 1	<i>a a</i> 1	<i>v</i> 0.3
<i>b</i> 0.4	<i>b v</i> 0.5	<i>w</i> 0.7
<i>c</i> 0.6	<i>b w</i> 0.2	<i>b</i> 1



## Query evaluation on probabilistic instances

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

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

<b>R</b>		<b>S</b>		<b>T</b>	
<i>a</i>	1	<i>a</i>	<i>a</i>	0.3	
<i>b</i>	0.4	<i>b</i>	<i>v</i>	0.5	
<i>c</i>	0.6	<i>b</i>	<i>w</i>	0.2	
				<i>b</i>	1





## Query evaluation on probabilistic instances

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

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

<b>R</b>	<b>S</b>	<b>T</b>
<i>a</i> 1	<i>a a</i> 1	<i>v</i> 0.3
<i>b</i> 0.4	<i>b v</i> 0.5	<i>w</i> 0.7
<i>c</i> 0.6	<i>b w</i> 0.2	<i>b</i> 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

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

<b>R</b>		<b>S</b>		<b>T</b>	
<i>a</i>	1	<i>a</i>	<i>a</i>	<i>v</i>	0.3
<i>b</i>	0.4	<i>b</i>	<i>v</i>	<i>w</i>	0.7
<i>c</i>	0.6	<i>b</i>	<i>w</i>	<i>b</i>	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

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

<b>R</b>		<b>S</b>		<b>T</b>	
<i>a</i>	1	<i>a</i>	<i>a</i>	<i>v</i>	0.3
<i>b</i>	0.4	<i>b</i>	<i>v</i>	<i>w</i>	0.7
<i>c</i>	0.6	<i>b</i>	<i>w</i>	<i>b</i>	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

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

<b>R</b>		<b>S</b>		<b>T</b>	
<i>a</i>	1	<i>a a</i>	1	<i>v</i>	0.3
<i>b</i>	0.4	<i>b v</i>	0.5	<i>w</i>	0.7
<i>c</i>	0.6	<i>b w</i>	0.2	<i>b</i>	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

→ 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		S			T	
<i>a</i>	1	<i>a</i>	<i>a</i>	1	<i>v</i>	0.3
<i>b</i>	0.4	<i>b</i>	<i>v</i>	0.5	<i>w</i>	0.7
<i>c</i>	0.6	<i>b</i>	<i>w</i>	0.2	<i>b</i>	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

→ Probability:  $0.4 \times$



## Query evaluation on probabilistic instances

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

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

<b>R</b>		<b>S</b>		<b>T</b>	
<i>a</i>	1	<i>a a</i>	1	<i>v</i>	0.3
<i>b</i>	0.4	<i>b v</i>	0.5	<i>w</i>	0.7
<i>c</i>	0.6	<i>b w</i>	0.2	<i>b</i>	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

→ 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		S			T	
<i>a</i>	1	<i>a</i>	<i>a</i>	1	<i>v</i>	0.3
<i>b</i>	0.4	<i>b</i>	<i>v</i>	0.5	<i>w</i>	0.7
<i>c</i>	0.6	<i>b</i>	<i>w</i>	0.2	<i>b</i>	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

→ 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

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

<b>R</b>		<b>S</b>		<b>T</b>
<i>a</i>	1	<i>a a</i>	1	<i>v</i> 0.3
<i>b</i>	0.4	<i>b v</i>	0.5	<i>w</i> 0.7
<i>c</i>	0.6	<i>b w</i>	0.2	<i>b</i> 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

→ 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 **P**TIME for any  $q \in \mathcal{S}$  on all instances
  - PQE is **#P-hard** for any  $q \in \mathcal{Q} \setminus \mathcal{S}$  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 **P**TIME for any  $q \in \mathcal{S}$  on all instances
  - PQE is **#P-hard** for any  $q \in \mathcal{Q} \setminus \mathcal{S}$  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 **P**TIME for any  $q \in \mathcal{S}$  on all instances
  - PQE is **#P-hard** for any  $q \in \mathcal{Q} \setminus \mathcal{S}$  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**  $\mathcal{Q}$ ?



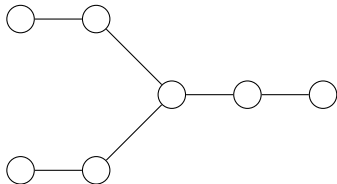
## Trees and treelike instances

- 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**)

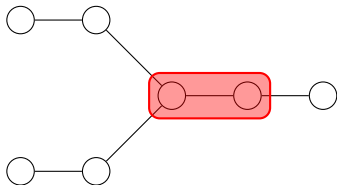






## Trees and treelike instances

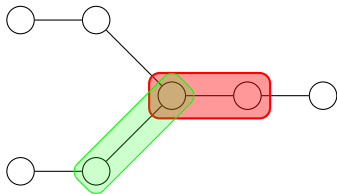
- Idea: let  $\mathcal{I}$  be treelike instances (constant bound on treewidth)





## Trees and treelike instances

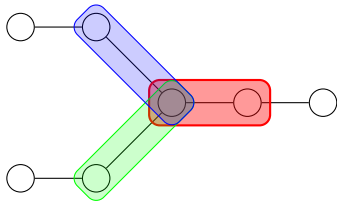
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





## Trees and treelike instances

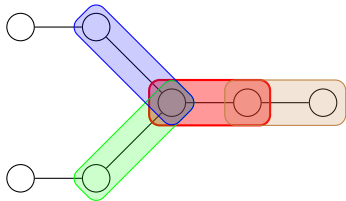
- Idea: let  $\mathcal{I}$  be treelike instances (constant bound on treewidth)





## Trees and treelike instances

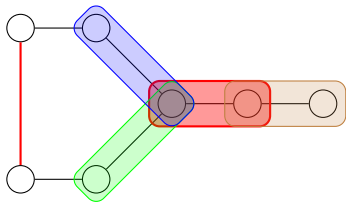
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

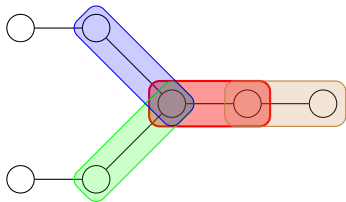
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

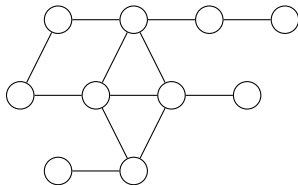
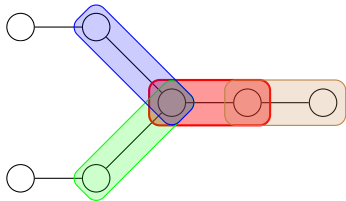
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





## Trees and treelike instances

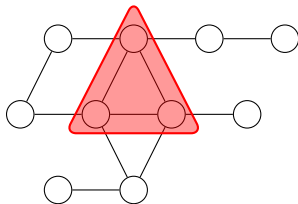
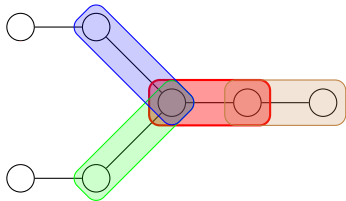
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





## Trees and treelike instances

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

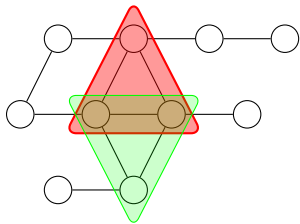
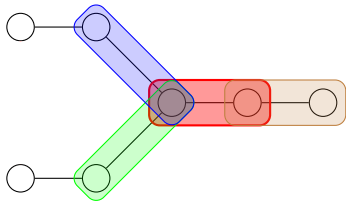






# Trees and treelike instances

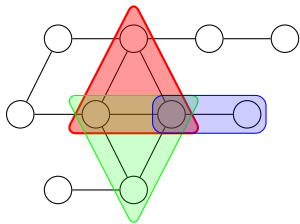
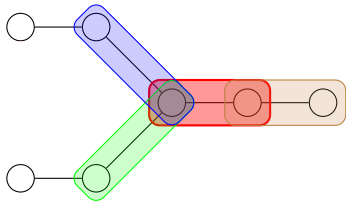
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

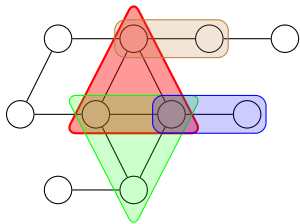
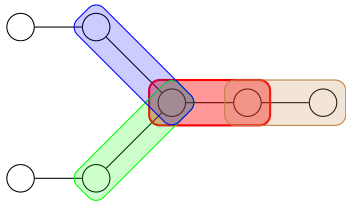
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

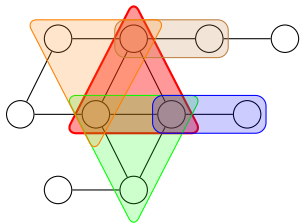
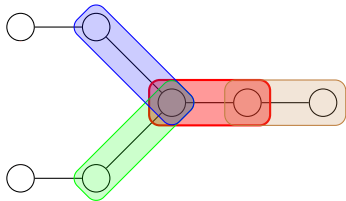
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

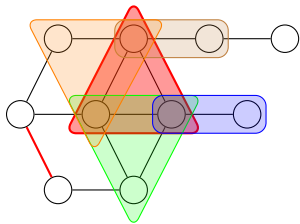
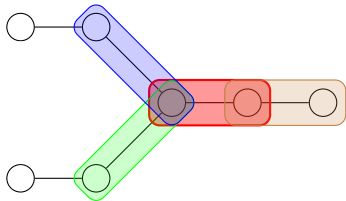
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

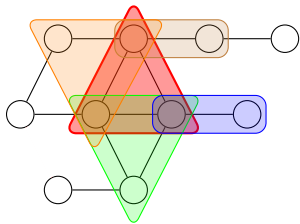
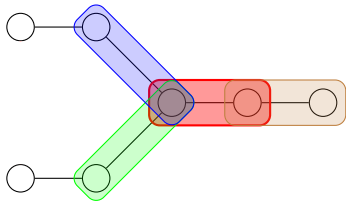
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





## Trees and treelike instances

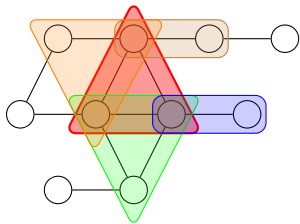
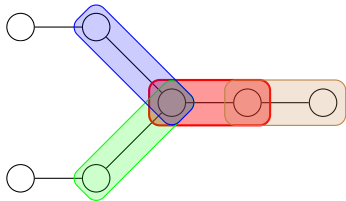
- Idea: let  $\mathcal{I}$  be **treelike instances** (constant bound on **treewidth**)





# Trees and treelike instances

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

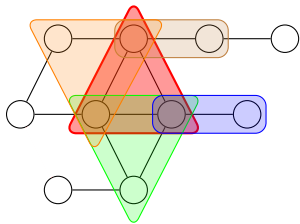
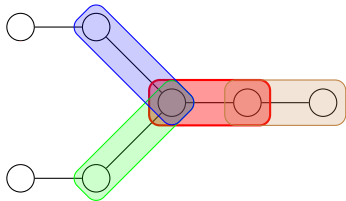


- **Trees** have treewidth 1
- **Cycles** have treewidth 2
- **$k$ -cliques** and  **$(k - 1)$ -grids** have treewidth  $k - 1$



# 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  $k - 1$

→ Known results (Courcelle 1990):

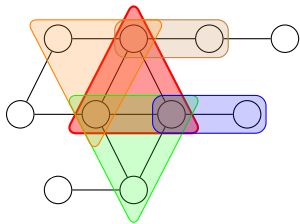
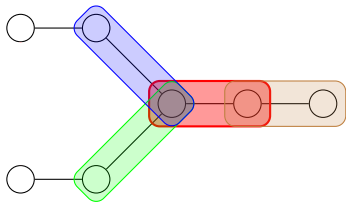
- $\mathcal{I}$ : **treelike** instances;  $\mathcal{Q}$ : **monadic second-order** queries
- **non-probabilistic** QE is in **linear time**





# 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  $k - 1$

→ Known results (Courcelle 1990):

- $\mathcal{I}$ : **treelike instances**;  $\mathcal{Q}$ : **monadic second-order queries**
- **non-probabilistic QE** is in **linear time**

→ 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**

→ 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**

- 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**

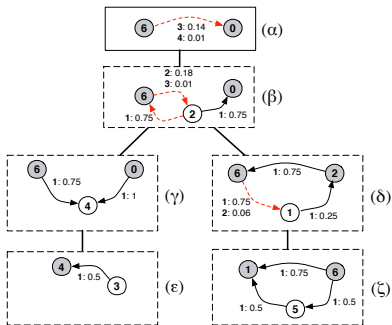
- PQE is in **linear time** modulo arithmetic costs
  - Also for expressive **provenance representations**
  - Also with bounded-treewidth **correlations**

**Lower bound.** (Amarilli et al. 2016)

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

- 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

Introduction

Truth Finding

Probabilistic Databases

Conclusion



# 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.