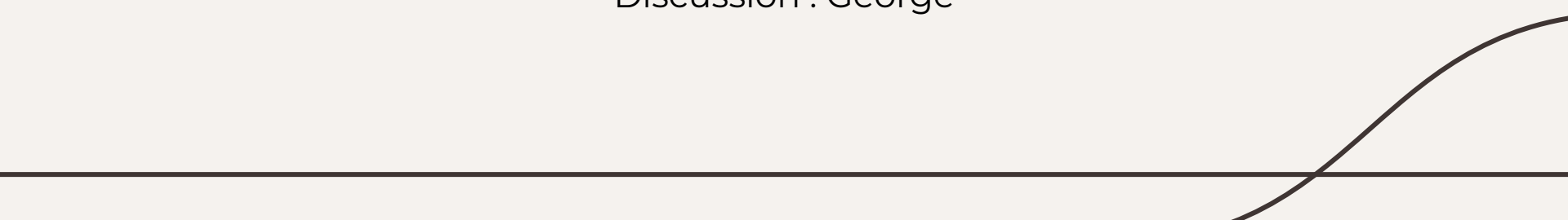




Diamonds in the Dirt

Probabilistic Databases

Presentation : Sandy
Discussion : George



01

Motivation

02

Challenges

03

Semantic

04

Representation

05

**Query
Evaluation**

06

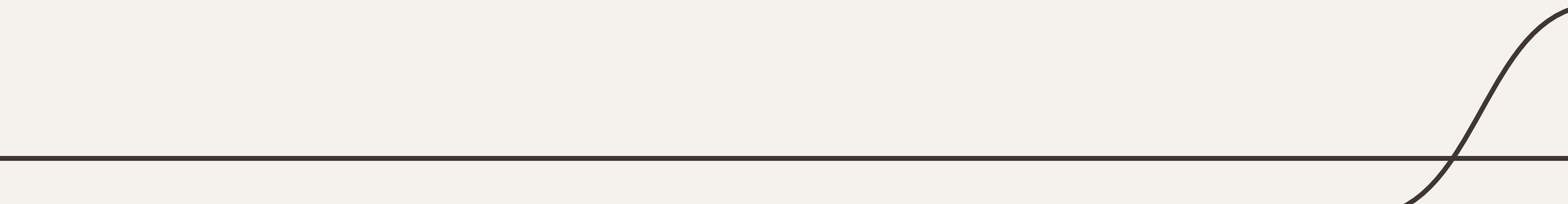
Query Result



01

Motivation

Why do we need probabilistic database?



Motivation of probabilistic database

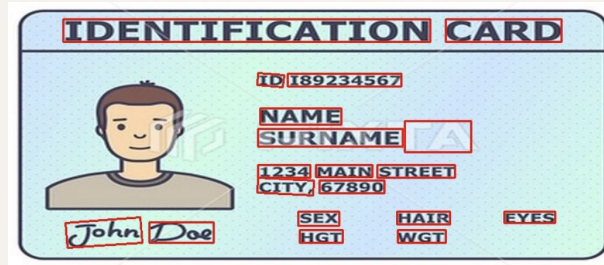
Why do we need probabilistic database when we have relational database?

- In today's databases (ex. Relational database)
 - Data is deterministic
 - Query result is deterministic
 - Data stored in the database is assumed to be accurate and reliable

- However, some data is not precise
 - Sensor data
 - Detector
- Need a database that can store data along with its associated probability

Example of uncertain data

- Sensor data
- Text detector



- High cost of cleaning
- Extractor : Purple Sox

Probabilistic database

- Refer as ProbDMS
- Record in the database is a probabilistic event
- Tuple in the query result is a probabilistic event
 - Database is deterministic, Query answers are probabilistic
 - Database is probabilistic, Query answers are probabilistic
- Similar structure to relational database

Product P

<u>Prod</u>	<u>Price</u>	Color	Shape	P
Gizmo	20	red	oval	p_1
Gizmo	20	blue	square	p_2
Camera	80	green	oval	p_3
Camera	80	red	round	p_4
Camera	80	blue	oval	p_5
iPod	300	white	square	p_6
iPod	300	black	square	p_7

Discussion - Groups of 4

Currently, there are no commercial probabilistic database products available, and only a few prototypes exist. What is the core reason behind this?

- Lack of real-world use cases?

Could you think of any use cases where probabilistic database systems is needed in your field?

- Imperfections?

Do you find any potential imperfections in the probabilistic database that prevents it from wider application?

02

Challenges

What are the challenges of creating the new system?

Question

- How do we store (represent) a probabilistic database?
 - How do we answer queries using our chosen representation?
 - How do we present the result of queries to the user?
-

Challenges


- Scalability
 - As system get better at representing (complex) relationship, it would be harder to manage large amount of data
 - Support for complex SQL query
 - In order to support decision, complex queries with aggregate must be supported in order to benefit from queries
 - Efficient query execution
 - Query has two parts : query + probabilistic inference
 - Good user interface
 - Decide which tuples should be return to users
-



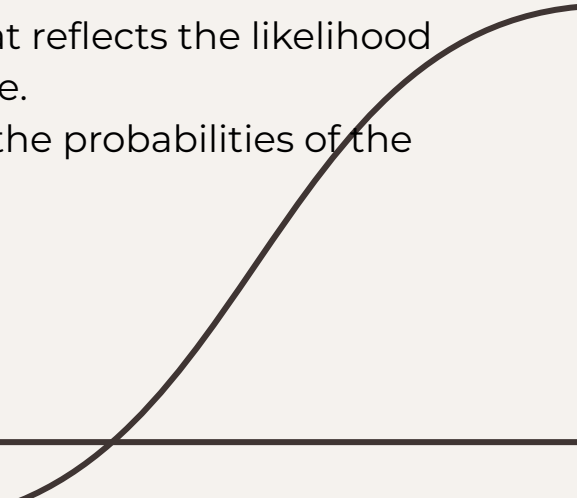
03

Semantic

“possible worlds model”



“possible worlds model”

- Semantic of ProbDMB
 - Possible world : a regular relational database
 - Probabilistic database: representation of a probability distribution over possible worlds
 - $PDB = (W, P)$
 - Each possible world is associated with a probability that reflects the likelihood or belief in that particular configuration of the database.
 - When querying, probabilities are calculated based on the probabilities of the possible worlds that satisfy the query conditions.
- 
-

Example of possible worlds model

- Assume that there are 6 possible readings for detection of three fields in the form
- Create an incomplete Database is a finite set of database instances $W = (W_1, \dots, W_6)$, each W is a possible world

Social Security Number: <u>185</u> Name: <u>Smith</u> Marital Status: (1) single <input checked="" type="checkbox"/> (2) married <input checked="" type="checkbox"/> (3) divorced <input type="checkbox"/> (4) widowed <input type="checkbox"/>	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>185</td><td>Brown</td><td>1</td></tr></tbody></table>	SSN	N	M	185	Smith	1	185	Brown	1	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>185</td><td>Brown</td><td>2</td></tr></tbody></table>	SSN	N	M	185	Smith	1	185	Brown	2
SSN	N	M																		
185	Smith	1																		
185	Brown	1																		
SSN	N	M																		
185	Smith	1																		
185	Brown	2																		
Social Security Number: <u>185</u> Name: <u>Brown</u> Marital Status: (1) single <input type="checkbox"/> (2) married <input type="checkbox"/> (3) divorced <input type="checkbox"/> (4) widowed <input type="checkbox"/>	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>185</td><td>Brown</td><td>3</td></tr></tbody></table>	SSN	N	M	185	Smith	1	185	Brown	3	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>185</td><td>Brown</td><td>4</td></tr></tbody></table>	SSN	N	M	185	Smith	1	185	Brown	4
SSN	N	M																		
185	Smith	1																		
185	Brown	3																		
SSN	N	M																		
185	Smith	1																		
185	Brown	4																		
	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>186</td><td>Brown</td><td>1</td></tr></tbody></table>	SSN	N	M	185	Smith	1	186	Brown	1	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>186</td><td>Brown</td><td>2</td></tr></tbody></table>	SSN	N	M	185	Smith	1	186	Brown	2
SSN	N	M																		
185	Smith	1																		
186	Brown	1																		
SSN	N	M																		
185	Smith	1																		
186	Brown	2																		
	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>186</td><td>Brown</td><td>3</td></tr></tbody></table>	SSN	N	M	185	Smith	1	186	Brown	3	<table border="1"><thead><tr><th>SSN</th><th>N</th><th>M</th></tr></thead><tbody><tr><td>185</td><td>Smith</td><td>1</td></tr><tr><td>186</td><td>Brown</td><td>4</td></tr></tbody></table>	SSN	N	M	185	Smith	1	186	Brown	4
SSN	N	M																		
185	Smith	1																		
186	Brown	3																		
SSN	N	M																		
185	Smith	1																		
186	Brown	4																		

Example of possible worlds model

- Each world is associated with a possibility P , showing the probability that this world is present
- Summation of the probability over every worlds equal to 1
- Marginal probability
 - ex: (SSN=185, N=Smith, M=1)
 - W_1 & W_3
 - $P(\text{Tuple}) = 0.1 + 0.1 = 0.2$

$W_1 : P(W_1) = 0.1$		
SSN	N	M
185	Smith	1
185	Brown	1

$W_2 : P(W_2) = 0.1$		
SSN	N	M
185	Smith	1
185	Brown	2

$W_3 : P(W_3) = 0.1$		
SSN	N	M
185	Smith	1
185	Brown	3

$W_4 : P(W_4) = 0.1$		
SSN	N	M
185	Smith	1
185	Brown	4

$W_5 : P(W_5) = 0.3$		
SSN	N	M
185	Smith	1
186	Brown	1

$W_6 : P(W_6) = 0.3$		
SSN	N	M
185	Smith	1
186	Brown	2

Discussion - Groups of 4

Since no mature systems have been implemented, there are many foreseeable fields of challenges and potential improvements for future probabilistic databases.

- What are the challenges in designing user interfaces for probabilistic databases? How can we effectively communicate uncertainty to users in a intuitive way?
- The trade-off between the need for accurately modelling complex correlations vs. maintaining system scalability and performance: Which one do you think is more crucial when designing the system? what extra criteria should be considered that guide the designer in making the trade-off?



04

Representation

How does probabilistic database look like?



BID: representation of ProbDMS

- p refer to probabilistic of the tuple
- Possible tuples are broken into block
 - Tuples in the same block is disjoint
 - Tuples in different block is independent

Researchers:

	<u>Name</u>	Affiliation	P	
t_1^1	Fred	U. Washington	$p_1^1 = 0.3$	$X_1 = 1$
t_1^2		U. Wisconsin	$p_1^2 = 0.2$	$X_1 = 2$
t_1^3		Y! Research	$p_1^3 = 0.5$	$X_1 = 3$
t_2^1	Sue	U. Washington	$p_2^1 = 1.0$	$X_2 = 1$
t_3^1	John	U. Wisconsin	$p_3^1 = 0.7$	$X_3 = 1$
t_3^2		U. Washington	$p_3^2 = 0.3$	$X_3 = 2$
t_4^1	Frank	Y! Research	$p_4^1 = 0.9$	$X_4 = 1$
t_4^2		M. Research	$p_4^2 = 0.1$	$X_4 = 2$

(a)

Services:

	<u>Name</u>	Conference	Role	P	
S_1	Fred	VLDB	Session Chair	$q_1 = 0.2$	$Y_1 = 1$
S_2	Fred	VLDB	PC Member	$q_2 = 0.8$	$Y_2 = 1$
S_3	John	SIGMOD	PC Member	$q_3 = 0.7$	$Y_3 = 1$
S_4	John	VLDB	PC Member	$q_4 = 0.7$	$Y_4 = 1$
S_5	Sue	SIGMOD	Chair	$q_5 = 0.5$	$Y_5 = 1$

(b)

Lineage : C-table

- History / provenance of tuple
- helps in understanding the reliability and trustworthiness
- C-table
 - Each tuple is annotated with a boolean expression over hidden variables

Location	
U. Washington	$(X_1 = 1) \wedge (Y_1 = 1) \vee (X_1 = 1) \wedge (Y_2 = 1) \vee (X_3 = 2) \wedge (Y_4 = 1)$
U. Wisconsin	$(X_1 = 2) \wedge (Y_1 = 1) \vee (X_1 = 2) \wedge (Y_2 = 1) \vee (X_3 = 1) \wedge (X_4 = 1)$
Y! Research	$(X_1 = 3) \wedge (Y_1 = 1) \vee (X_1 = 3) \wedge (Y_2 = 1)$

Discussion - Groups of 2

The future trend of probabilistic databases.

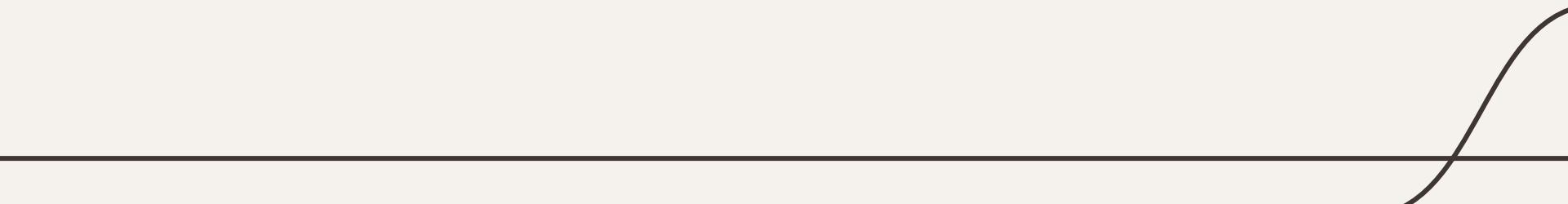
- Are there any emerging trends or technologies not covered in the paper that are shaping the future of probabilistic databases?
 - Probabilistic inference is closely related to machine learning and AI, which of the three aspects described in the paper is most likely to be the first that can be done by AI?
-



05

Query Evaluation

How to evaluate query?



Method 1

- Separate query and lineage evaluation from probabilistic inference
 - Various algorithms
- Probabilistic inference take up long time

Method 2

- Integrate probabilistic inference with the query computation
 - Benefit
 - Use query optimization technology
 - What and How
 - What user what → query
 - How to execute query → plan
 - Safety
 - Safe query
 - Can push probabilistic inference inside query plan
 - Safe plan
 - Allow probabilities to be computed in relational algebra
-

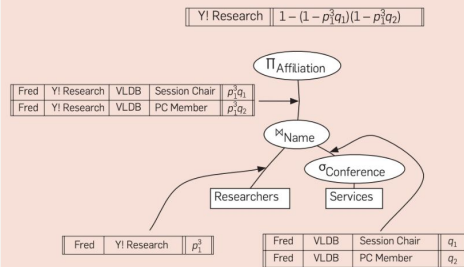
Choosing the query plan

- Efficiency & Safety
- Need to ensure correctness of the optimized plan
- Search for low cost & safe

```

SELECT x.Affiliation, confidence( )
FROM Researchers x, Services y
WHERE x.Name = y.Name
      and y.Conference = 'VLDB'
GROUP BY x.Affiliation
    
```

(a)

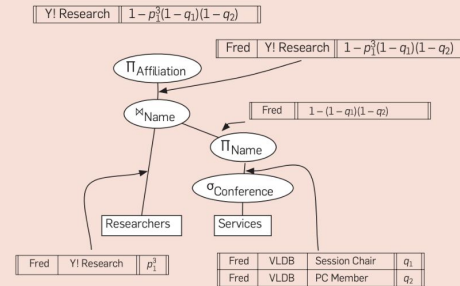


(b)

```

SELECT x.Affiliation, 1-prod(1-x.P*y.P)
FROM Researchers x, (SELECT Name, 1-(1-prod(P))
                     FROM Services
                     WHERE Conference = 'VLDB'
                     GROUP BY Name) y
WHERE x.Name = y.Name
GROUP BY x.Affiliation
    
```

(d)



(c)

06

Query Result

How result is calculated from query?

Top-k Query answering

- Impossible and meaningless to show all tuples in results
- Rank tuples and restrict to show only top k tuples
- Rank based on decreasing order of the output probabilities

Discussion - Groups of 2

Monthly journals such as Communication of the ACM are read by a broad range of computer scientists and researchers, not just database practitioners.

- What factors should the authors keep in mind while writing articles for such a diverse audience?
 - What should be the readers' mindsets?
 - Apart from surveys or literature reviews, what other types of publications are best suited for this type of monthly journal?
-

Conclusion

- Goal of the probabilistic database is to reduce cost of using uncertain data
- Finding “diamond in dirt”