Scalable Uncertainty Management

04 – Probabilistic Databases

Rainer Gemulla

Jun 1, 2012

Overview

In this lecture

- Refresher: Finite probability (not presented)
- What is a probabilistic database?
- How can probabilistic information be represented?
- How expressive are these representations?
- How to query probabilistic databases?

Not in this lecture

- Complexity
- Efficiency
- Algorithms

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

Sample space

Definition

The *sample space* Ω of an *experiment* is the set of all possible *outcomes*. We henceforth assume that Ω is finite.

Example

- Toss a coin: $\Omega = \{\,\text{Head}, \text{Tail}\,\}$
- Throw a dice: $\Omega = \{\,1, 2, 3, 4, 5, 6\,\}$

In general, we cannot predict with certainty the outcome of an experiment in advance.

Event

Definition

An *event* $A \subseteq \Omega$ is a subset of the sample space. $\emptyset$ is called the *empty event*, Ω the *trivial event*. Two events A and B are *disjoint* if $A \cap B = \emptyset$.

Example

Coin:

- Outcome is a head: $A = \{\,\text{Head}\,\}$
- Outcome is head or tail: $A = \{\,\text{Head}, \text{Tail}\,\} = \{\,\text{Head}\,\} \cup \{\,\text{Tail}\,\}$
- Outcome is both head and tail: $A = \emptyset = \{\,\text{Head}\,\} \cap \{\,\text{Tail}\,\}$
- Outcome is not head: $A = \{\,\text{Tail}\,\} = \{\,\text{Head}\,\}^c$

Die:

- Outcome is an even number: $A = \{\,2, 4, 6\,\} = \{\,2\,\} \cup \{\,4\,\} \cup \{\,6\,\}$
- Outcome is even and ≤ 3: $A = \{\,2\,\} = \{\,2, 4, 6\,\} \cap \{\,1, 2, 3\,\}$

When $A, B \subseteq \Omega$ are events, so are $A \cup B$, $A \cap B$, and A^c, representing 'A or B', 'A and B', and 'not A', respectively.

Probability space

Definition

A *probability measure* $(2^\Omega, \mathbb{P})$ is a function $\mathbb{P} : 2^\Omega \to [0,1]$ satisfying

a) $\mathbb{P}(\emptyset) = 0$, and $\mathbb{P}(\Omega) = 1$,

b) If $A_1, \ldots, A_n$ are pairwise disjoint, $\mathbb{P}(\bigcup_{i=1}^n A_n) = \sum_{i=1}^n \mathbb{P}(A_n)$.

The triple $(\Omega, 2^\Omega, \mathbb{P})$ is called a *probability space*.

Example

For $\omega \in \Omega$, we write $\mathbb{P}(\omega)$ for $\mathbb{P}(\{\omega\})$; $\{\omega\}$ called *elementary event*.

- Coin: $2^\Omega = \{\emptyset, \{\text{Head}\}, \{\text{Tail}\}, \{\text{Head}, \text{Tail}\}\}$
- Fair coin: $\mathbb{P}(\text{Head}) = \mathbb{P}(\text{Tail}) = \frac{1}{2}$
  Implied: $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{\text{Head}, \text{Tail}\}) = 1$
- Fair dice: $\mathbb{P}(1) = \cdots = \mathbb{P}(6) = \frac{1}{6}$ (rest implied)
- Outcome is even: $\mathbb{P}(\{2,4,6\}) = \mathbb{P}(2) + \mathbb{P}(4) + \mathbb{P}(6) = \frac{1}{2}$
- Outcome is ≤ 3: $\mathbb{P}(\{1,2,3\}) = \mathbb{P}(1) + \mathbb{P}(2) + \mathbb{P}(3) = \frac{1}{2}$

Conditional probability

Definition

If $\mathbb{P}(B) > 0$, then the conditional probability that A occurs given that B occurs is defined to be

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Example

Two dice; prob. that total exceeds 6 given that first shows 3?

- $\Omega = \{1, \ldots, 6\}^2$
- Total exceeds 6: $A = \{(a, b) : a + b > 6\}$
- First shows 3: $B = \{(3, b) : 1 \leq b \leq 6\}$
- $A \cap B = \{(3, 4), (3, 5), (3, 6)\}$
- $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B) / \mathbb{P}(B) = \frac{3}{36} / \frac{6}{36} = \frac{1}{2}$

Independence

Definition

Two events A and B are called *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

If $\mathbb{P}(B) > 0$, implies that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

Example

Two independent events:

- Die shows an even number: $A = \{2, 4, 6\}$
- Die shows at most 4: $B = \{1, 2, 3, 4\}$:
- $\mathbb{P}(A \cap B) = \mathbb{P}(\{2, 4\}) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = \mathbb{P}(A)\,\mathbb{P}(B)$

Not independent:

- Die shows an odd number: $C = \{1, 3, 5\}$
- $\mathbb{P}(A \cap C) = \mathbb{P}(\emptyset) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(C)$

Disjointness $\neq$ independence.

Conditional independence

Definition

Let A, B, C be events with $\mathbb{P}(C) > 0$. A and B are *conditionally independent given* C if $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$.

Example

- Die shows an even number: $A = \{2, 4, 6\}$
- Die shows at most 3: $B = \{1, 2, 3\}$
- $\mathbb{P}(A \cap B) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)$
  $\rightarrow$ A and B are not independent
- Die does not show multiple of 3: $C = \{1, 2, 4, 5\}$
- $\mathbb{P}(A \cap B \mid C) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$
  $\rightarrow$ A and B are conditionally independent given C

Product space

Definition

Let $(\Omega_1, 2^{\Omega_1}, \mathbb{P}_1)$ and $(\Omega_2, 2^{\Omega_2}, \mathbb{P}_2)$ be two probability spaces. Their *product space* is given by $(\Omega_{12}, 2^{\Omega_{12}}, \mathbb{P}_{12})$ with $\Omega_{12} = \Omega_1 \times \Omega_2$ and

$$\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\,\mathbb{P}_2(A_2).$$

Example

Toss two fair dice.

- $\Omega_1 = \Omega_2 = \{1, 2, 3, 4, 5, 6\}$
- $\Omega_{12} = \{(1, 1), \ldots, (6, 6)\}$
- First die: $A_1 = \{1, 2, 3\} \subseteq \Omega_1$
- Second die: $A_2 = \{2, 3, 4\} \subseteq \Omega_2$
- $\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\,\mathbb{P}_2(A_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Product spaces combine the outcomes of several *independent* experiments into one space.

Random variable

Definition

A *random variable* is a function $X : \Omega \to \mathbb{R}$. We will write $\{ X = x \}$ or $\{ X \leq x \}$ for the events $\{ \omega : X(\omega) = x \}$ and $\{ \omega : X(\omega) \leq x \}$, respectively. The *probability mass function* of X is the function $f_X : \mathbb{R} \to [0, 1]$ given by $f_X(x) = \mathbb{P}(X = x)$; its *distribution function* is given by $F_X(x) = \mathbb{P}(X \leq x)$.

Example

Toss two dice:

- Sum of outcomes: $X((a, b)) = a + b$
- $f_X(6) = \mathbb{P}(X = 6) = \mathbb{P}(\{ (1,5), (2,4), (3,3), (4,2), (5,1) \}) = \frac{5}{36}$
- $F_X(3) = \mathbb{P}(X \leq 3) = \mathbb{P}(\{ (1,1), (1,2), (2,1) \}) = \frac{1}{12}$

The notions of conditional probability, independence (consider events $\{ X = x \}$ and $\{ Y = y \}$ for all x and y), and conditional independence also apply to random variables.

Expectation

Definition

The *expected value* of a random variable X is given by

$$\mathbb{E}[X] = \sum_x x\, f_X(x).$$

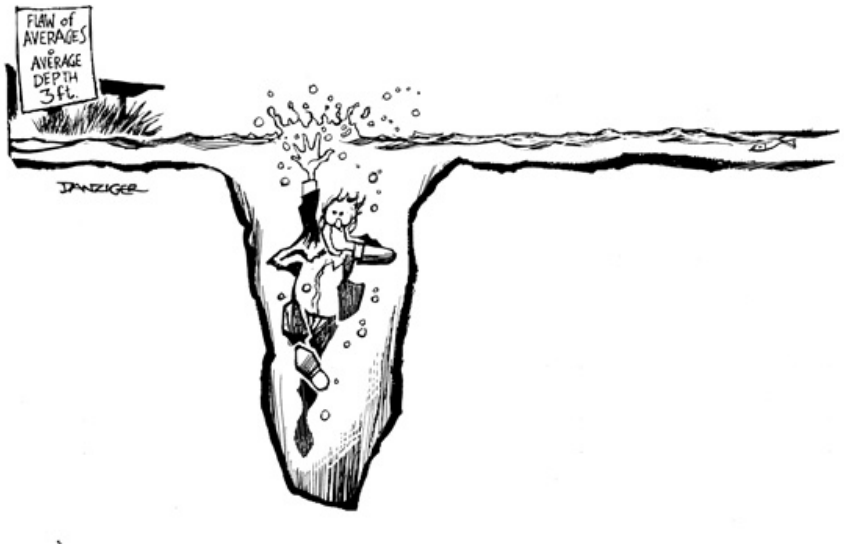
If $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}[g(X)] = \sum_x g(x) f_X(x).$$

Example

- Fair die (with X being identity)
- $\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$
- Consider $g(x) = \lfloor x/2 \rfloor$
- $\mathbb{E}[g(x)] = 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + \cdots + 3 \cdot \frac{1}{6} = 1.5$
- But: $g(\mathbb{E}[X]) = 1$!

Flaw of averages



Mean correct, variance ignored.

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

Be careful with expected values!

Savage, 2009.

Conditional expectation

Definition

Let X, Y be random variables. The *conditional expection* of Y given X is the random variable $\psi(X)$ where

$$\psi(x) = \mathbb{E}[\, Y \mid X = x \,] = \sum_y y \, f_{Y|X}(y \mid x),$$

where $f_{Y|X}(y \mid x) = \mathbb{P}(\, Y = y \mid X = x \,)$.

Example

- *Indicator variable*: $I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$
- Fair die; set $X = I_{\text{even}} = I_{\{\,2,4,6\,\}}$; Y is identity
- $\mathbb{E}[\, Y \mid X = 1 \,] = 1 \cdot 0 + 2 \cdot \frac{1}{3} + 3 \cdot 0 + 4 \cdot \frac{1}{3} + 5 \cdot 0 + 6 \cdot \frac{1}{3} = 4$
- $\mathbb{E}[\, Y \mid X = 0 \,] = 1 \cdot \frac{1}{3} + 2 \cdot 0 + 3 \cdot \frac{1}{3} + 4 \cdot 0 + 5 \cdot \frac{1}{3} + 6 \cdot 0 = 3$
- $\mathbb{E}[\, Y \mid X \,](\omega) = \begin{cases} 4 & \text{if } X(\omega) = 1 \\ 3 & \text{if } X(\omega) = 0 \end{cases}$

Important properties

We use shortcut notation $\mathbb{P}(X)$ for $\mathbb{P}(X = x)$.

Theorem

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

$$\textit{If } B \supseteq A, \quad \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$$

$$\mathbb{P}(X) = \sum_y \mathbb{P}(X, Y = y) \qquad \textit{(sum rule)}$$

$$\mathbb{P}(X, Y) = \mathbb{P}(Y \mid X)\mathbb{P}(X) \qquad \textit{(product rule)}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \textit{(Bayes theorem)}$$

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b \qquad \textit{(linearity of expectation)}$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X] \qquad \textit{(law of total expectation)}$$

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

Amateur bird watching

- Bird watcher's observations

Sightings

Name	Bird	Species	
Mary	Bird-1	Finch: 0.8 ‖ Toucan: 0.2	t_1
Susan	Bird-2	Nightingale: 0.65 ‖ Toucan: 0.35	t_2
Paul	Bird-3	Humming bird: 0.55 ‖ Toucan: 0.45	t_3

- Which species may have been sighted? $\rightarrow$ CWA, possible tuples

ObservedSpecies

Species		
Finch	0.80	$(t_1, 1)$
Toucan	0.71	$(t_1, 2) \vee (t_2, 2) \vee (t_3, 2)$
Nightingale	0.65	$(t_2, 1)$
Humming bird	0.55	$(t_3, 1)$

Probabilistic databases quantify uncertainty.

What do probabilities mean?

- Multiple interpretations of probability
- Frequentist interpretation
  - Probability of an event = relative frequency when repeated often
  - Coin, n trials, n_{H} observed heads

$$\lim_{n\to\infty} \frac{n_{\mathrm{H}}}{n} = \frac{1}{2} \implies \mathbb{P}(\mathrm{H}) = \frac{1}{2}$$

- Bayesian interpretation
  - Probability of an event = degree of belief that event holds
  - Reasoning with "background knowledge" and "data"
  - Prior belief + model + data $\rightarrow$ posterior belief
    - ★ Model parameter: θ = true "probability" of heads
    - ★ Prior belief: $\mathbb{P}(\theta)$
    - ★ Likelihood (model): $\mathbb{P}(n_{\mathrm{H}}, n \mid \theta)$
    - ★ Bayes theorem: $\mathbb{P}(\theta \mid n_{\mathrm{H}}, n) \propto \mathbb{P}(n_{\mathrm{H}}, n \mid \theta)\,\mathbb{P}(\theta)$
    - ★ Posterior belief: $\mathbb{P}(\theta \mid n_{\mathrm{H}}, n)$

But... what do probabilities really mean? And where do they come from?

- Answers differ from application to application, e.g.,
  - Information extraction $\rightarrow$ from probabilistic models
  - Data integration $\rightarrow$ from background knowledge & expert feedback
  - Moving objects $\rightarrow$ from particle filters
  - Predictive analytics $\rightarrow$ from statistical models
  - Scientific data $\rightarrow$ from measurement uncertainty
  - Fill in missing data $\rightarrow$ from data mining
  - Online applications $\rightarrow$ from user feedback
- Semantics sometimes precise, sometimes less so
- Often: Convert model scores to $[0, 1]$
  - Larger value $\rightarrow$ higher confidence
  - Carries over to queries: higher probability of an answer $\rightarrow$ more credible
  - Ranking often more informative than precise probabilities

Many applications can benefit from a platform that manages probabilistic data.

Probabilistic database

Example

Sightings

Name	Bird	Species
Mary	Bird-1	Finch: 0.8 ‖ Toucan: 0.2
Susan	Bird-2	Nightingale: 0.65 ‖ Toucan: 0.35
Paul	Bird-3	Humming bird: 0.55 ‖ Toucan: 0.45

Possible worlds:

N	B	S
M	1	F
S	2	N
P	3	H

0.286

N	B	S
M	1	F
S	2	N
P	3	T

0.234

N	B	S
M	1	F
S	2	T
P	3	H

0.154

N	B	S
M	1	F
S	2	T
P	3	T

0.126

N	B	S
M	1	T
S	2	N
P	3	H

0.0715

N	B	S
M	1	T
S	2	N
P	3	T

0.0585

N	B	S
M	1	T
S	2	T
P	3	H

0.0385

N	B	S
M	1	T
S	2	T
P	3	T

0.0315

Definition

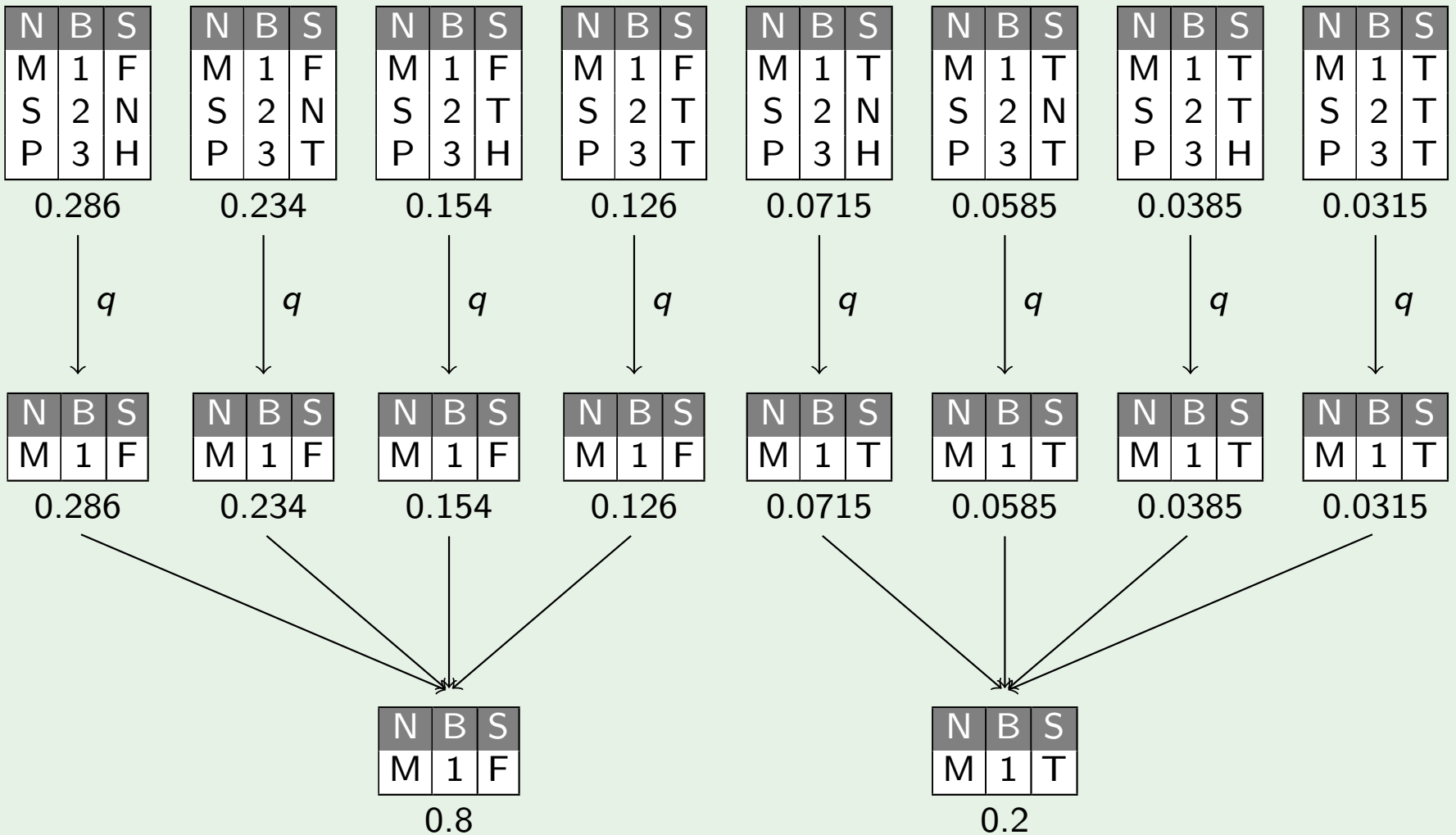
A (finite) *probabilistic database* (*p-database*, *PDB*) is a probability space $\mathcal{D} = (\mathcal{I}, \mathbb{P})$ over a (finite) incomplete database $\mathcal{I}$ in which w.l.o.g. $\mathbb{P}(I) > 0$ for all $I \in \mathcal{I}$.

A PDB associates a nonzero probability to each *possible world* $I \in \mathcal{I}$.

Possible answer set semantics (example)

Example

What did Mary see? $\rightarrow q(R) = \sigma_{\text{Name='Mary'}}(R)$



Possible answer set semantics

Definition

The *possible answer set* to a query q on a probabilistic database $\mathcal{D} = (\mathcal{I}, \mathbb{P})$ is the probability space $\mathcal{D}_q = (q(\mathcal{I}), \mathbb{P}_q)$, where $q(\mathcal{I})$ is the possible answer set to q on $\mathcal{I}$, and

$$\mathbb{P}_q(J) = \mathbb{P}(q(I) = J) = \mathbb{P}(\{I \in \mathcal{I} : q(I) = J\}) = \sum_{I \in \mathcal{I} : q(I) = J} \mathbb{P}(I).$$

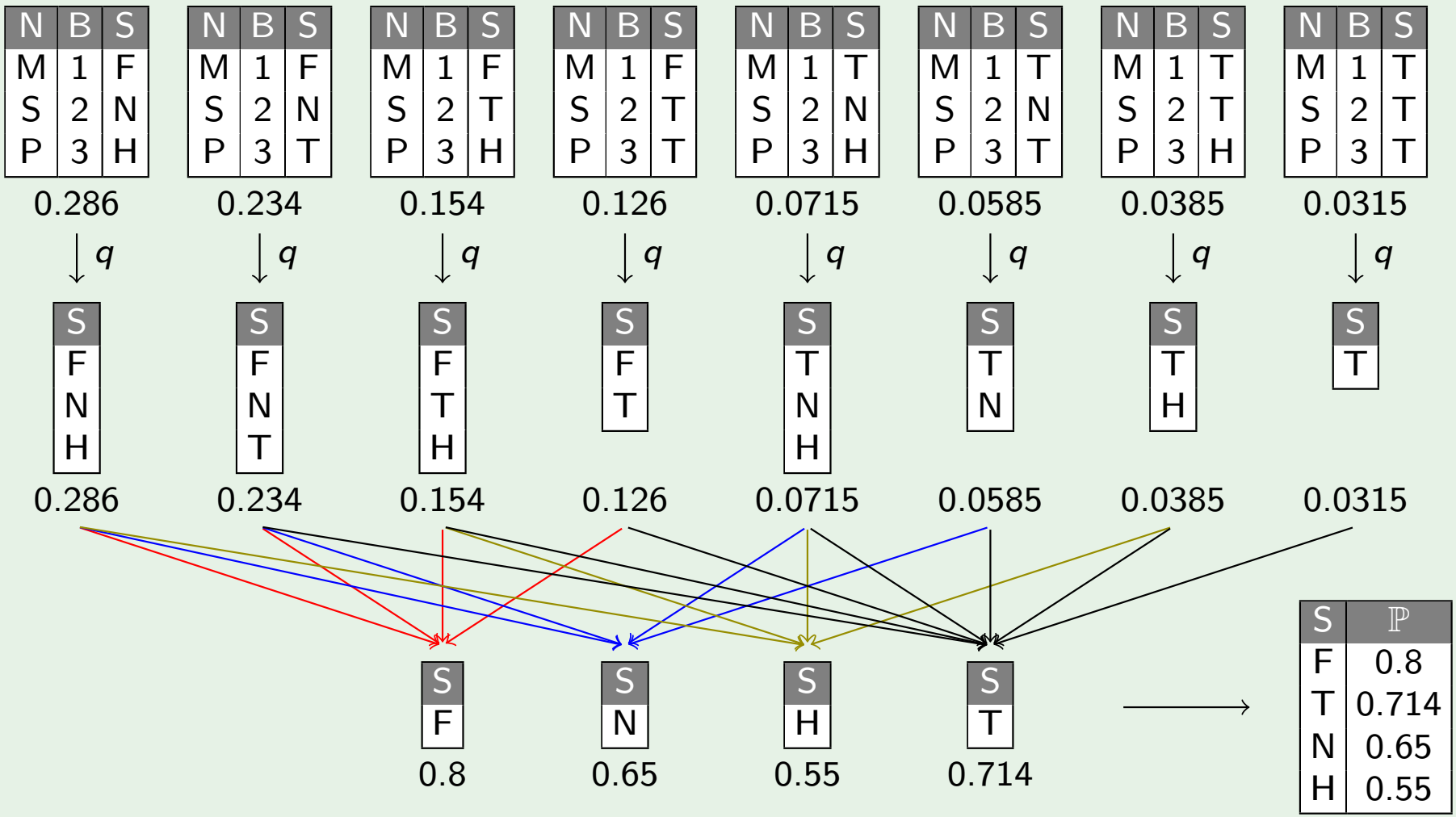
We refer to $\mathcal{D}_q$ as the *image* of $\mathcal{D}$ under q.

- Cf. definition for incomplete databases
- $|q(\mathcal{I})| \leq |\mathcal{I}|$ since each instance of $\mathcal{I}$ gives precisely one result $q(I)$

Possible tuple semantics (example)

Example

Which species have been sighted? $\rightarrow q(R) = \pi_{\text{Species}}(R)$



Possible tuple semantics

Definition

Let $\mathcal{D} = (\mathcal{I}, \mathbb{P})$ be a probabilistic database. A tuple t is a *possible answer* to a query q if there exists a possible world $I \in \mathcal{I}$ such that $t \in q(I)$. The *marginal probability* of t is given by

$$\mathbb{P}\left(t \in q(I) \right) = \sum_{I \in \mathcal{I} : t \in q(I)} \mathbb{P}\left(I \right).$$

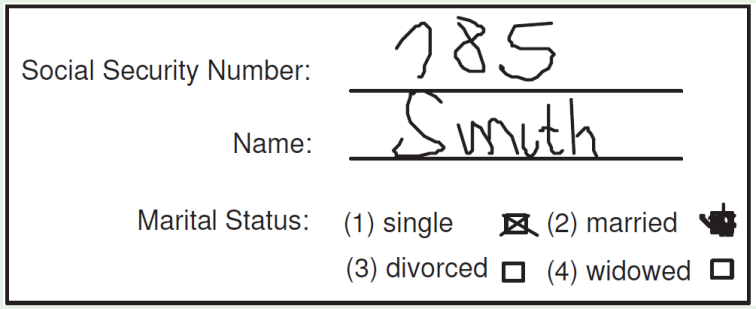
- A tuple t is a *certain answer* if $\mathbb{P}\left(t \in q(I) \right) = 1$;
  equivalently, $(\forall I \in \mathcal{I})\, t \in q(I)$
  $\rightarrow$ Certain answer tuple semantics as before (q-information).
  $\rightarrow$ Weak representation results carry over.
- Possible tuple semantics is the main focus of probabilistic databases

Outline

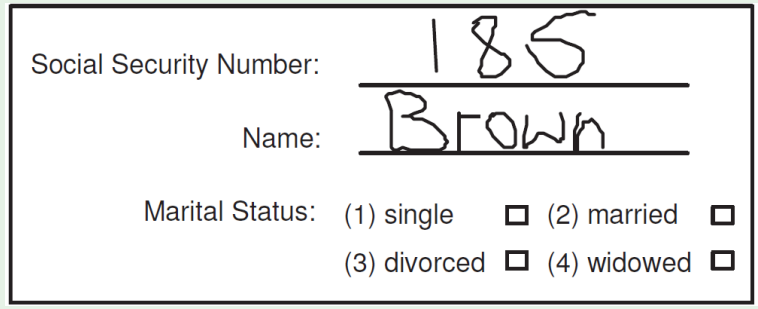
1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

Motivating example

Example



Form 1



Form 2

Ambiguity:

- Is Smith single or married?
- What is the martial status of Brown?
- What is Smith's social security number: 185 or 785?
- What is Brown's social security number: 185 or 186?

Probabilistic database:

- Here: $2 \cdot 4 \cdot 2 \cdot 2 = 32$ possible readings $\rightarrow$ can easily store all of them
- 200M people, 50 questions, 1 in 10000 ambiguous (2 options) $\rightarrow 2^{10^6}$ possible readings
- Each reading is a table with 50 columns and 200M rows!

Probabilistic representation system

Finiteness assumption: Throughout our entire treatment of PDBs.

Definition

A *probabilistic representation system* consists of a set $\mathscr{T}$ of tables and a function Mod that associates to each table $T \in \mathscr{T}$ a probabilistic database $\mathrm{Mod}(T)$.

Definition

A probabilistic representation system is *complete* if it can represent any probabilistic database.

Definition

Let $(\mathscr{T}, \mathrm{Mod})$ be a probabilistic representation system and $\mathscr{L}$ be a query language. The probabilistic representation system obtained by *closing* $\mathscr{T}$ *under* $\mathscr{L}$ is the set of tables $\{ (T, q) \mid T \in \mathscr{T}, q \in \mathscr{L} \}$ and function $\mathrm{Mod}(T, q) = q(\mathrm{Mod}(T))$.

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

pc-table (example)

Example

FID	SSN	Name	
1	185	Smith	$X = 1$
1	785	Smith	$X \neq 1$
2	185	Brown	$Y = 1 \wedge X \neq 1$
2	186	Brown	$Y \neq 1 \vee X = 1$

$\underline{V}$	D	$\mathbb{P}$
X	1	0.2
X	2	0.8
Y	1	0.3
Y	2	0.7

FID	SSN	Name
1	185	Smith
2	186	Brown

$\{X \mapsto 1, Y \mapsto 1\}$
$\{X \mapsto 1, Y \mapsto 2\}$
$0.2 \cdot 0.3 + 0.2 \cdot 0.7$
$= 0.2$

FID	SSN	Name
1	785	Smith
2	185	Brown

$\{X \mapsto 2, Y \mapsto 1\}$
$0.8 \cdot 0.3$
$= 0.24$

FID	SSN	Name
1	785	Smith
2	186	Brown

$\{X \mapsto 2, Y \mapsto 2\}$
$0.8 \cdot 0.7$
$= 0.56$

pc-tables

Definition

A *probabilistic c-table* (pc-table) is pair $(T, \mathbb{P})$, where T ia a c-table and $\mathbb{P}$ a probability distribution over the set of assignments Θ of $\mathrm{Var}(T)$ such that all variables are independent.

$$\mathrm{Mod}(T) = \{\, \theta(T) : \theta \in \Theta \,\}$$

$$\mathbb{P}(I) = \sum_{\theta \in \Theta : \theta(T) = I} \mathbb{P}(\theta)$$

- Variables are independent
  $\rightarrow$ need only specify probabilities of form $\mathbb{P}(X = a)$
- $\mathbb{P}$ can be stored in a standard relation storing (variable, value, probability)-triples

Completeness of pc-tables

Theorem

pc-tables are a complete representation system.

Proof.

Let $\mathcal{D} = (\mathcal{I}, \mathbb{P})$ be a probabilistic database with $\mathcal{I} = \{ I^1, \ldots, I^n \}$ and $I^k = \{ t_{k1}, \ldots, t_{kn_k} \}$. Let X be a random variable with domain $\{ 1, \ldots, n \}$. Set $\mathbb{P}(X = k) = \mathbb{P}(\mathcal{I}^k)$ and use the c-table:

$\alpha(\mathcal{I})$	
t_{11}	$X = 1$
$\vdots$	
t_{1n_1}	$X = 1$
t_{21}	$X = 2$
$\vdots$	
t_{2n_2}	$X = 2$
t_{31}	$X = 3$
$\vdots$	

Completeness of pc-tables (example)

Example

I^1

FID	SSN	Name
1	185	Smith
2	186	Brown

0.2

I^2

FID	SSN	Name
1	785	Smith
2	185	Brown

0.24

I^3

FID	SSN	Name
1	785	Smith
2	186	Brown

0.56

FID	SSN	Name	
1	185	Smith	$X = 1$
2	186	Brown	$X = 1$
1	785	Smith	$X = 2$
2	185	Brown	$X = 2$
1	785	Smith	$X = 3$
2	186	Brown	$X = 3$

$\underline{V}$	D	$\mathbb{P}$
X	1	0.2
X	2	0.24
X	3	0.56

pc-tables are strong

Theorem

pc-tables are strong under $\mathcal{RA}$.

Proof.

Given a pc-table $(T, \mathbb{P})$ and a query q, the resulting pc-table is given by $(\bar{q}(T), \mathbb{P})$, where $\bar{q}$ is the c-table algebra query corresponding to q. □

Example

R

FID	SSN	Name	
1	185	Smith	$X = 1$
1	785	Smith	$X \neq 1$
2	185	Brown	$Y = 1 \wedge X \neq 1$
2	186	Brown	$Y \neq 1 \vee X = 1$

V	D	$\mathbb{P}$
X	1	0.2
X	2	0.8
Y	1	0.3
Y	2	0.7

$\pi_{\text{SSN}}(R)$

SSN	
185	$X = 1 \vee (Y = 1 \wedge X \neq 1)$
785	$X \neq 1$
186	$Y \neq 1 \vee X = 1$

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

Tuple-independent databases (p?-tables)

Definition

In a *tuple-independent probabilistic database* T, each tuple $t \in T$ is marked with a probability $p_t > 0$. We have $\mathrm{Mod}(T) = (\mathcal{I}, \mathbb{P})$ where $\mathcal{I} = \{ I \subseteq T : \mathbb{P}(I) > 0 \}$ and

$$\mathbb{P}(I) = \Big(\prod_{t \in I} p_t\Big)\Big(\prod_{t \notin I}(1 - p_t)\Big).$$

Example (Nell)

Recently-Learned Facts twitter Refresh

instance	iteration	date learned	confidence
dried_squash_seeds is a nut	225	28-mar-2011	99.5
sinnett_thorn_mountain_cave is a cave	225	28-mar-2011	99.7
vail_road is a street	224	26-mar-2011	98.4
harold_macmillan is a scientist	225	28-mar-2011	96.6
n32207 is a ZIP code	224	26-mar-2011	99.4
wday_tv collaborates with bbc_news	224	26-mar-2011	96.9
times controls friedman	227	03-apr-2011	96.9
support_personnel is a profession that is a kind of professionals	224	26-mar-2011	96.9
nbc_news is a newspaper in the city washington__dc	224	26-mar-2011	99.2
twitter operates the website twitter_com	225	28-mar-2011	100.0

Completeness

Theorem

Tuple-independent databases are not complete.

Proof.

They can only represent databases in which all tuples are independent events. E.g., they cannot represent

$$\left\{ \underset{0.5}{\boxed{a}}, \underset{0.5}{\boxed{b}} \right\} \quad \text{or} \quad \left\{ \underset{0.1}{\boxed{}}, \underset{0.1}{\boxed{a}}, \underset{0.1}{\boxed{b}}, \underset{0.7}{\boxed{\begin{matrix} a \\ b \end{matrix}}} \right\}.$$

□

Theorem

The closure of tuple-independent databases under positive $\mathcal{RA}$ is not complete.

Closure under $\mathcal{RA}$

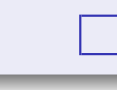
Theorem

The closure of tuple-independent databases under $\mathcal{RA}$ is complete.

Proof.

Let $\mathcal{D} = (\mathcal{I}, \mathbb{P})$ be a probabilistic database with $\mathcal{I} = \{ I^1, \ldots, I^n \}$. To obtain a tuple-independent database, use n certain EDB predicates $R^1, \ldots, R^n$ with $I(R^k) = I^k$ and one tuple-independent table W that contains tuples $\{1, \ldots, n\}$ with $p_k = \mathbb{P}\left(I^k \mid \{ I_1, \ldots, I_{k-1} \}^c \right)$. Write a query that selects relation R^k iff $\operatorname{argmin}_{t:W(t)} = k$:

$$
\begin{array}{ll}
R(\mathbf{x}) \leftarrow W(1), R^1(\mathbf{x}) & p_1 = \mathbb{P}\left(I^1 \right) \\
R(\mathbf{x}) \leftarrow \neg W(1), W(2), R^2(\mathbf{x}) & p_2 = \mathbb{P}\left(I^2 \mid \{ I^1 \}^c \right) \\
R(\mathbf{x}) \leftarrow \neg W(1), \neg W(2), W(3), R^3(\mathbf{x}) & p_3 = \mathbb{P}\left(I^3 \mid \{ I^1, I^2 \}^c \right) \\
\vdots & \vdots \\
R(\mathbf{x}) \leftarrow \neg W(1), \ldots, \neg W(n-1), W(n), R^n(\mathbf{x}) & p_n = 1
\end{array}
$$



Closure under $\mathcal{RA}$ (example)

Example

$I^1 = I(R^1)$

FID	SSN	Name
1	185	Smith
2	186	Brown

0.2

$I^2 = I(R^2)$

FID	SSN	Name
1	785	Smith
2	185	Brown

0.24

$I^3 = I(R^3)$

FID	SSN	Name
1	785	Smith
2	186	Brown

0.56

$$
\begin{aligned}
R(f,s,n) &\leftarrow W(1), R^1(f,s,n) && p_1 = 0.2 \\
R(f,s,n) &\leftarrow \neg W(1), W(2), R^2(f,s,n) && p_2 = 0.24/(1-0.2) \\
R(f,s,n) &\leftarrow \neg W(1), \neg W(2), W(3), R^3(f,s,n) && p_3 = 0.56/(1-0.2-0.24)
\end{aligned}
$$

W

World	$\mathbb{P}$	
1	0.2	$\mathbb{P}(\operatorname{argmin}_{t:W(t)} = 1) = 0.2$
2	0.3	$\mathbb{P}(\operatorname{argmin}_{t:W(t)} = 2) = 0.3 \cdot (1-0.2) = 0.24$
3	1	$\mathbb{P}(\operatorname{argmin}_{t:W(t)} = 3) = 1 \cdot (1-0.2) \cdot (1-0.3) = 0.56$

Probabilistic database design

- Database normalization $\rightarrow$ Minimize redundancy/correlations
- Tuple-independent databases are good building blocks
  - No correlations between tuples
  - No constraints
  - Database normalization can be applied
- Decompose complex databases into tuple-independent databases

Example (Nell)

- nellExtraction: extracted relations
  (tuple probability = belief that extracted tuple is correct)
- nellSource: source of extraction
  (tuple probability = belief that source is correct)
- Correlation via views

ProducesProduct(x, y) $\leftarrow$ nellExtraction(x, 'ProducesProduct', y, s), nellSource(s)

Tuple-independent databases can be stored in standard relations.

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

BID tables

- Relations are partitioned into blocks
- Events within a block a disjoint; events across blocks are independent
  $\rightarrow$ Block-independent-disjoint database
- Blocks are identified by key attributes

Example

FID	SSN	Name	
1	185	Smith	$X = 1$
1	785	Smith	$X = 2$
2	175	Brown	$Y = 1$
2	186	Brown	$Y = 2$

$\underline{V}$	D	$\mathbb{P}$
X	1	0.8
X	2	0.2
Y	1	0.5
Y	2	0.5

$\rightarrow$

$\underline{\text{FID}}$	SSN	Name	$\mathbb{P}$
1	185	Smith	0.8
1	785	Smith	0.2
2	175	Brown	0.5
2	186	Brown	0.5

Theorem

BID-tables extended with PJR queries are a complete representation system.

U-tables (MayBMS)

- Goal: completeness + natural representation in RDBMS
- Restrict pc-table conditions to forms $X_1 = a_1 \wedge \ldots \wedge X_k = a_k$
- Conditions $\rightarrow$ U-tables (usually: one per set of correlated attributes)
- Distribution over assignments $\rightarrow$ BID-table (*world table*)

Example

R

FID	SSN	Name	
1	185	Smith	$X = 1$
1	785	Smith	$X = 2$
2	185	Brown	$Y = 1 \wedge X = 2$
2	186	Brown	$Y = 2$
2	186	Brown	$X = 1$

W

$\underline{V}$	D	$\mathbb{P}$
X	1	0.2
X	2	0.8
Y	1	0.3
Y	2	0.7

T

V_1	D_1	V_2	D_2	FID	SSN	Name
X	1	X	1	1	185	Smith
X	2	X	2	1	785	Smith
Y	1	X	2	2	185	Brown
Y	2	Y	2	2	186	Brown
X	1	X	1	2	186	Brown

Reconstruction via joins: $R(f, s, n) \leftarrow T(v_1, d_1, v_2, d_2, f, s, n), W(v_1, d_1), W(v_2, d_2)$

Theorem

U-databases are complete. They can compute/represent results of nr-datalog queries conveniently (i.e., in polynomial time and space).

Or-set tables

Example

Probabilistic or-set tables (= probabilistic finite-domain Codd tables):

Sightings

Name	Bird	Species
Mary	Bird-1	Finch: 0.8 ‖ Toucan: 0.2
Susan	Bird-2	Nightingale: 0.65 ‖ Toucan: 0.35
Paul	Bird-3	Humming bird: 0.55 ‖ Toucan: 0.45

Probabilistic ?-or-set tables (Trio):

Sightings

Name	Bird	Species	
Mary	Bird-1	Finch: 0.8 ‖ Toucan: 0.2	
Susan	Bird-2	Nightingale: 0.65 ‖ Toucan: 0.10	?
Paul	Bird-3	Humming bird	0.55

Outline

1. Refresher: Finite Probability
2. Probabilistic Databases
3. Probabilistic Representation Systems
   - pc-tables
   - Tuple-independent databases
   - Other common representation systems
4. Summary

Lessons learned

- Probabilistic databases quantify uncertainty
- Probabilistic database = incomplete database + probability distribution
- Many notions and results from incomplete databases carry over
- Queries can be analyzed in terms of
  1. Possible answer sets
  2. Certain answer tuples (same as incomplete databases)
  3. Possible answer tuples (main focus of PDBs)
- pc-tables $\rightarrow$ complete, strong under $\mathcal{RA}$
- Tuple-independent tables $\rightarrow$ complete when closed under $\mathcal{RA}$ (Good probabilistic database design)
- BID-tables $\rightarrow$ complete when closed under PJR queries
- U-databases $\rightarrow$ complete, handle positive $\mathcal{RA}$ well, easy to represent in an RDBMS

Suggested reading

- Charu C. Aggarwal (Ed.)
  Managing and Mining Uncertain Data (Chapter 2)
  Springer, 2009

- Dan Sucio, Dan Olteanu, Christopher Ré, Christoph Koch
  Probabilistic Databases (Chapter 2)
  Not yet published (But you'll get copies!)

- Charu C. Aggarwal (Ed.)
  Managing and Mining Uncertain Data (Chapter 5 $\rightarrow$ Trio)
  Springer, 2009

- Charu C. Aggarwal (Ed.)
  Managing and Mining Uncertain Data (Chapter 6 $\rightarrow$ MayBMS)
  Springer, 2009