Probabilistic Databases for Query Answering under Missing Values

Leopoldo Bertossi

Joint work with: **Farouk Toumani** (U. Clermont-Ferrand, France)

DBs with Missing Values

- A^o, B^o always observed variables (no MVs)

  $C^\star$ is the observed version of underlying variable C^m that may have missing values

- Occurrence of MVs is governed by a *Missingness Mechanism* (MM)

  Dependencies upon the same or other variables

  MMs first formulated/investigated/applied by Donald Rubin

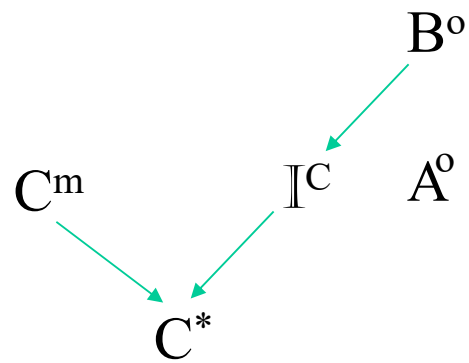
- MMs can represented by Causal or Bayesian Networks (K. Mohan & J. Pearl)

- Values for $C^\star$, including MVs, depend on C^m and B^o via $\mathbb{I}^C$

  $\mathbb{I}^C$: Missingness indicator variable (0 or 1)

	A^o	B^o	$C^\star$
τ_1	a	0	0
τ_2	a	0	0
τ_3	a	1	na
τ_4	a	1	0
τ_5	a	1	na
τ_6	a	1	1
τ_7	a	1	na
τ_8	a	1	2

"observed" DB $D^\star$



$\mathcal{M}$ **MAR** case

- MMs can be used to do classic imputation
- Want to query "underlying" DB D^m, but we only have $D^\star$
- Many possible D^ms ...
  - Determined by what?
  - Which one?
  - With what properties?
  - Computed how?
  - ...
- Instead of usual imputation methods, build a Probabilistic DB

Probabilistic DBs for MVs

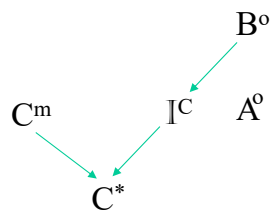
- Each underlying value for C^m has a probability of being the one missing $\quad Dom(C^m) = \{0, 1, 2\}$

  Determined by MG $\mathcal{M}$ and other information in the tuple

	A^o	B^o	$C^\star$
τ_1	a	0	0
τ_2	a	0	0
τ_3	a	1	na
τ_4	a	1	0
τ_5	a	1	na
τ_6	a	1	1
τ_7	a	1	na
τ_8	a	1	2

$$P^{\mathcal{M}}(C^m = 0 \mid A^o = a, B^o = 1, C^\star = \text{na}) \quad ?$$

- Assume:



B^o	$\mathbb{I}^C$	$P_{B^o}(\mathbb{I}^C)$
0	0	1
0	1	0
1	0	$\frac{1}{4}$
1	1	$\frac{3}{4}$

$$P^{\mathcal{M}}(C^m = 0) := \tfrac{1}{2}$$
$$P^{\mathcal{M}}(C^m = 1) = P^{\mathcal{M}}(C^m = 2) := \tfrac{1}{4}$$
(distribution in $\mathcal{M}$)

- Here:

  $p_3^1 \;:=\; P^{\mathcal{M}}(C^m = 0 \mid A^o = a, B^o = 1, C^\star = \text{na}) = P^{\mathcal{M}}(C^m = 0) = \frac{1}{2}$

- Tuples are assumed to be independent

- We obtain:
- A Block-Independent PDB (BID)
- With a collection $\mathcal{W}$ of possible worlds
- They are instances w/o MVs, each with a global probability
- A possible world W with probability $(\frac{1}{2})^3$: a most-probable world
- We can define and do QA as usual on PDBs

$$P^{\mathcal{W}}(Q) := \sum_{\substack{W \in \mathcal{W} \\ W \models Q}} P^{\mathcal{W}}(W)$$

(coming back ...)

- In our case, QA turns out to be #P-hard

	A^o	B^o	C^m	p^{BID}
τ_1	a	0	0	1
τ_2	a	0	0	1
$\mathcal{B}(\tau_3)$	a	1	0	1/2
	a	1	1	1/4
	a	1	2	1/4
τ_4	a	1	0	1
$\mathcal{B}(\tau_5)$	a	1	0	1/2
	a	1	1	1/4
	a	1	2	1/4
τ_6	a	1	1	1
$\mathcal{B}(\tau_7)$	a	1	0	1/2
	a	1	1	1/4
	a	1	2	1/4
τ_8	a	1	2	1

W

	A^o	B^o	C^m
τ_1	a	0	0
τ_2	a	0	0
τ_3^1	a	1	0
τ_4	a	1	0
τ_5^1	a	1	0
τ_6	a	1	1
τ_7^1	a	1	0
τ_8	a	1	2

Classes of Matching Possible Worlds

- Can we do something better?
- Some worlds are "matching", i.e. identical as multisets

World	Notation	$\mathbf{P}^{\mathcal{W}}(\mathbf{W})$
$\mathbf{W_1}$	$\mathbf{W^{[000]}}$	$(\frac{1}{2})^3 = \mathbf{0.125}$
W_2	$W^{[001]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
W_3	$W^{[002]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
W_4	$W^{[010]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
W_5	$W^{[011]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
$\mathbf{W_6}$	$\mathbf{W^{[012]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_7	$W^{[020]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
$\mathbf{W_8}$	$\mathbf{W^{[021]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_9	$W^{[022]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{10}	$W^{[100]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
W_{11}	$W^{[101]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
$\mathbf{W_{12}}$	$\mathbf{W^{[102]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{13}	$W^{[110]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{14}	$W^{[111]}$	$(\frac{1}{4})^3 = 0.015$
W_{15}	$W^{[112]}$	$(\frac{1}{4})^3 = 0.015$
$\mathbf{W_{16}}$	$\mathbf{W^{[120]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{17}	$W^{[121]}$	$(\frac{1}{4})^3 = 0.015$
W_{18}	$W^{[122]}$	$(\frac{1}{4})^3 = 0.015$
W_{19}	$W^{[200]}$	$(\frac{1}{2})^2 \times \frac{1}{4} = 0.06$
$\mathbf{W_{20}}$	$\mathbf{W^{[201]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{21}	$W^{[202]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
$\mathbf{W_{22}}$	$\mathbf{W^{[210]}}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{23}	$W^{[211]}$	$(\frac{1}{4})^3 = 0.015$
W_{24}	$W^{[212]}$	$(\frac{1}{4})^3 = 0.015$
W_{25}	$W^{[220]}$	$(\frac{1}{4})^2 \times \frac{1}{2} = 0.03$
W_{26}	$W^{[221]}$	$(\frac{1}{4})^3 = 0.015$
W_{27}	$W^{[222]}$	$(\frac{1}{4})^3 = 0.015$

	A^o	B^o	C^m
τ_1	a	0	0
τ_2	a	0	0
τ_3^1	**a**	**1**	**0**
τ_4	a	1	0
τ_5^3	**a**	**1**	**2**
τ_6	a	1	1
τ_7^1	a	1	0
τ_8	a	1	2

	A^o	B^o	C^m
τ_1	a	0	0
τ_2	a	0	0
τ_3^3	**a**	**1**	**2**
τ_4	a	1	0
τ_5^1	**a**	**1**	**0**
τ_6	a	1	1
τ_7^1	a	1	0
τ_8	a	1	2

$P^{\mathcal{W}}(W^{[020]}) = 1 \times 1 \times \frac{1}{2} \times 1 \times \frac{1}{4} \times 1 \times \frac{1}{2} \times 1.$

$P^{\mathcal{W}}(W^{[200]}) = 1 \times 1 \times \frac{1}{4} \times 1 \times \frac{1}{2} \times 1 \times \frac{1}{2} \times 1.$

Two matching worlds

Group them in classes of matching worlds

Probability of a class is the sum of the worlds' probabilities (that may differ)

Most Compliant Classes

- All worlds in a class return the same query answer
- Some worlds, so as their classes, may be more "compliant" with the MG than others ...
- Compute the (same) empirical distribution of members in a class:

$$P^E_W(t) := mult^W(t)/||W||$$

(count as multisets)

Classes C	Worlds	$P^{\mathcal{C}}(C)$	$KLD(C)$
C_1	$W^{[002]}, W^{[020]}, W^{[200]}$	0.188	0.431
C_2	$W^{[011]}, W^{[101]}, W^{[110]}$	0.094	0.518
C_3	$W^{[012]}, W^{[021]}, W^{[102]}, W^{[120]}, W^{[201]}, W^{[210]}$	0.188	0.452
C_4	$W^{[111]}$	0.016	0.711
C_5	$W^{[001]}, W^{[010]}, W^{[100]}$	0.188	0.431
C_6	$W^{[022]}, W^{[202]}, W^{[220]}$	0.094	0.711
C_7	$W^{[112]}, W^{[121]}, W^{[211]}$	0.047	0.929
C_8	$W^{[000]}$	0.125	0.431
C_9	$W^{[122]}, W^{[212]}, W^{[221]}$	0.047	0.972
C_{10}	$W^{[222]}$	0.016	1.111

- Kullback-Leibler Divergence, for comparison with $\mathcal{M}$-induced distribution:

(T: set of different tuples)

$$Div_{KL}(P^E_W \parallel P^{\mathcal{M}}) := \sum_{t \in T} P^E_W(t) \ln \frac{P^E_W(t)}{P^{\mathcal{M}}(t)}$$

- Here, one most compliant class (there may be several)

Algorithmic and Complexity Results

- We have several results on hardness of: arXiv:2604.06520
  - Computation of Most-Probable Classes
  - Computation of Most-Compliant Classes
- We can efficiently computing one most-compliant class
  - By reduction to computing a "minimum-cost flow" in bipartite graphs
  - Once we have such a class, we can compute an MC-world, its probability, and query it

    Actually, a most probable in that class
  - The results applies to a broad family of compliance measures, in particular to KL
  - Also efficiently a Pareto-optimal world
- Our implementation is more efficient than all common baseline imputation methods

Back to the BID

- What about QA directly on our BID?
- Our BID is not a tuple-independent PDB (TID)

  For TIDs there are available methods and implementations
- Use "Marko-Views" to transform the BID into a TID

  (A. Kumar Jha & D. Suciu, 2012)
- Impose soft-constraints à la Markov-Logic Networks on BID

  Requesting: "at least one tuple per block"
  "at most one tuple per block"
- Defined Marko-Views are their violation views
- Getting tuples into them has a high cost, i.e. heavy weight, leading to:
  - Changing the initial probabilities (via weights)
  - DB tuples become independent

- Observed DB $D^\star$ is queried via the resulting TID
- QA on the resulting TID is done with ProvSQL (P. Senellart et al.)
- It uses query-lineage and knowledge compilation for QA
- We have run experiments with relatively large DBs
- We are in the process of analysing, comparing and reporting on QA ...
- Comparison with:
  - Previously described approach
  - The initial, baseline DB D on which MVs were introduced using the MG, to create $D^\star$