Top-k Querying of Unknown Values under Order Constraints

Antoine Amarilli, Yael Amsterdamer, Tova milo, Pierre Senellart

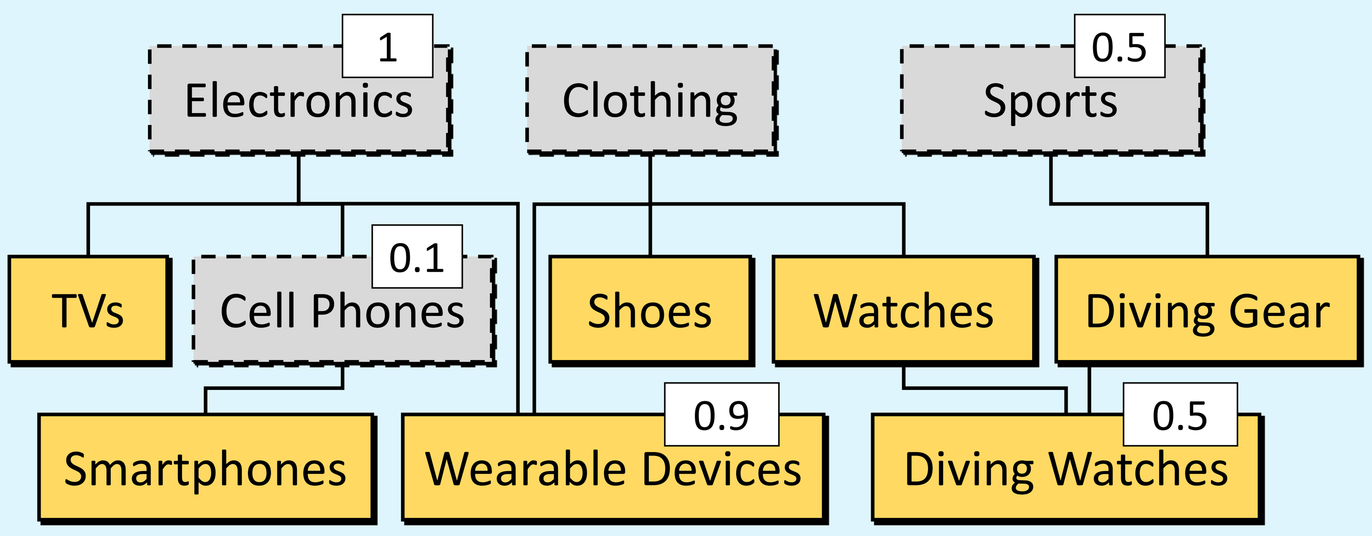
LTCI, Télécom ParisTech, Université Paris-Saclay
Bar Ilan University
Tel Aviv University
DI, École normale supérieure, PSL Research University
Inria Paris

Research Problem:

Given a set of items with unknown values estimate (A) top-k items (B) their values based on known values and partial order constraints

Motivating Example:

Top-k categories compatible with a given product



Model:

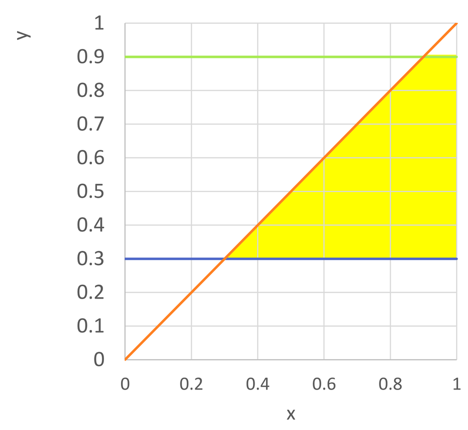
$\mathcal{X} = \{x_1, \dots, x_n\}$ a set of variables, $\mathcal{X}_\sigma$ selected variables

- We assume x_i takes values in [0,1]

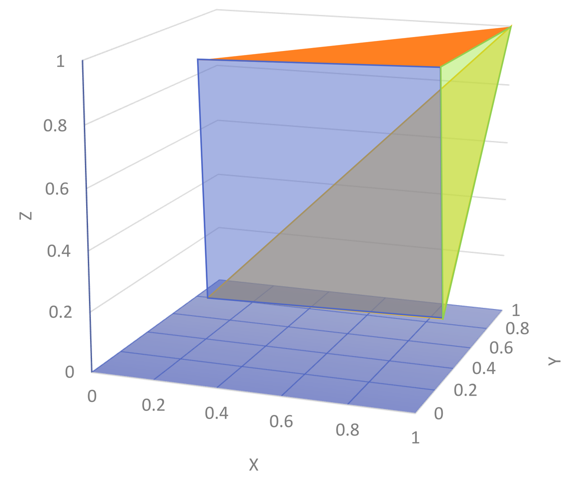
$\mathcal{C}$ a set of order and exact-value constraints over $\mathcal{X}$

- Exact values are rationals

$\mathrm{pw}(\mathcal{C})$ - possible worlds, all valuations of $\mathcal{X}$ that satisfy $\mathcal{C}$



$x \geq y, 0.3 = w \leq y \leq z = 0.9$



$x \geq y, 0.3 = w \leq y \leq z$

We assume a uniform pdf over $\mathrm{pw}(\mathcal{C})$

Top-k semantics: k items with highest expected values

- Desirable properties (in literature)
- A value estimate compatible with top-k
- Interpolation over posets – independent contribution

See the paper for other variants

Results

Algorithm for Top-k and Interpolation

Proposition: if $\mathcal{C}$ implies a total order, the expected value of $x \in \mathcal{X}$ can be computed in PTIME

Fragment. Distribution independent from other fragments.
Marginals follow (rescaled) Beta distribution by connection to order statistics

$$\underset{\alpha}{\underset{\parallel}{x_0}} \leq x_1 \leq \cdots \leq x_{i-1} \leq \overbrace{\underset{v_i}{\underset{\parallel}{x_i}} \leq x_{i+1} \leq \cdots \leq x_{j-1} \leq \underset{v_j}{\underset{\parallel}{x_j}}} \leq x_{j+1} \leq \cdots \leq x_n \leq \underset{\beta}{\underset{\parallel}{x_{n+1}}}$$

Theorem: for general $\mathcal{C}$, interpolation and top-k are in $\mathrm{FP}^{\#P}$

- Weighted sum of expected values over linear extensions of $\mathcal{C}$, weights by the probability of each ordering
- Nondeterministically sum over linear extensions

See the paper for full algorithm

Hardness (tight bounds)

Theorem [Rademacher 2007]: interpolation is $\mathrm{FP}^{\#P}$-hard

- Even without exact-value constraints

Theorem: top-k is $\mathrm{FP}^{\#P}$-hard even without expected values

- Reduction from interpolation!
- Use top-k to compare the value of x to fresh exact-value
- Bound the denominator
- Rational number identification using polynomial # comparisons

Approximations

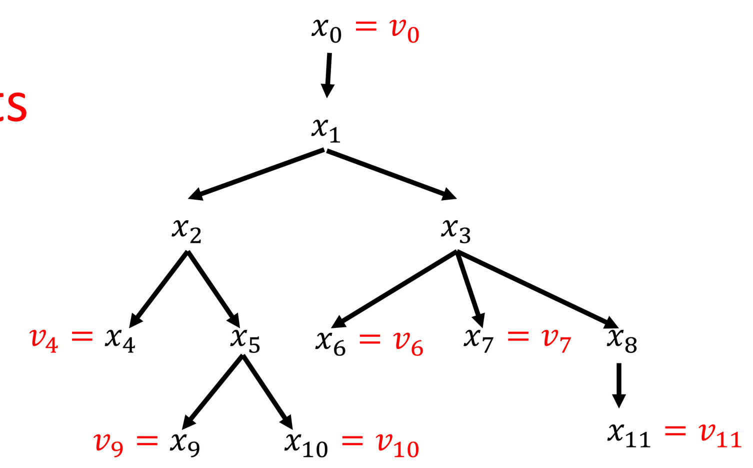
- Solving with high probability is hard unless $NP \subseteq BPP$
- An FPRAS with bounded error ε on expected value
  - By connection to volume computation
  - High degree PTIME complexity, may admit efficient implementations

Splitting Lemma

- The *influence relation* $x \leftrightarrow x'$ is the symmetric, reflexive, and transitive closure of $\prec$ on $\mathcal{X} \backslash \mathcal{X}_{exact}$
- Its equivalence classes are used in a definition of *uninfluenced decomposition* $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_m$ of $\mathcal{C}$
- Proof by a bijection over possible worlds

Tractable Cases

tree-shaped constraint sets



Theorem: if $\mathcal{C}$ is tree-shaped,
$\mathrm{V}(\mathcal{C})$ can be computed in time $\mathrm{O}(|\mathcal{X}^2|)$

- Bottom-up processing, propagating a piecewise polynomial function for the volume of subtree based on root's parent value
- Complexity proof by induction

Theorem: if $\mathcal{C}$ is tree-shaped,
x's marginal can be computed in $\mathrm{O}(|\mathcal{X}_{exact}| \cdot |\mathcal{X}^2|)$

- A similar bottom-up scheme
- Computing the pdf as a function from v to $\mathrm{V}(\mathcal{C}_{x=v})$
- $|\mathcal{X}_{exact}|$ factor is due to the pieces of the polynomial

Corollary: if $\mathcal{C}$ is decomposable to (reverse-)tree-shaped, interpolation and top-k can be solved in PTIME