Dichotomies in the Complexity of Query Answering over Probabilistic Databases

Open Problems in Database Theory, ICDT 2017

Benny Kimelfeld[1]

[1]Thanks to **Dan Suciu** for valuable input!

Tuple-Independent Probabilistic Databases

- A *tuple-independent probabilistic database* [DS04], or **TID** for short, is a pair (D, p) where:
  - D is an ordinary relational database, viewed as a set of *facts*
  - $p : D \to [0, 1]$ associates a probability with each fact

Tuple-Independent Probabilistic Databases

- A *tuple-independent probabilistic database* [DS04], or **TID** for short, is a pair (D, p) where:
  - D is an ordinary relational database, viewed as a set of *facts*
  - $p : D \to [0,1]$ associates a probability with each fact
- Semantics: probability distribution over the subinstances $E \subseteq D$:

$$\Pr(E \mid D, p) \stackrel{\text{def}}{=\!=} \left(\prod_{f \in E} p(f)\right) \times \left(\prod_{f \in D \setminus E} (1 - p(f))\right)$$

Tuple-Independent Probabilistic Databases

- A *tuple-independent probabilistic database* [DS04], or **TID** for short, is a pair (D, p) where:
  - D is an ordinary relational database, viewed as a set of *facts*
  - $p : D \to [0, 1]$ associates a probability with each fact
- Semantics: probability distribution over the subinstances $E \subseteq D$:

$$\Pr(E \mid D, p) \stackrel{\text{def}}{=\!=} \left(\prod_{f \in E} p(f)\right) \times \left(\prod_{f \in D \setminus E} (1 - p(f))\right)$$

- Can simulate and facilitate common models in Statistical Relational Learning (SRL), such as *Markov Logic Networks*, **if** expressive classes of queries can be evaluated efficiently [JS12]

Problem 1: Query Evaluation

The Query Evaluation Problem

Let Q be a boolean query. *Evaluation of Q over TIDs* is the following problem. Given (D, p), compute the probability $\pi_Q(D, p)$ that Q is satisfied by a random database of (D, p).

Problem 1: Query Evaluation

The Query Evaluation Problem

Let Q be a boolean query. *Evaluation of* Q *over TIDs* is the following problem. Given (D, p), compute the probability $\pi_Q(D, p)$ that Q is satisfied by a random database of (D, p). That is:

$$\pi_Q(D, p) \stackrel{\text{def}}{=} \sum_{E \subseteq D \,,\, E \models Q} \Pr(E \mid D, p)$$

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]
- Dichotomy for **RA**$(\sigma, \pi, \bowtie, \rho, -)$, *each relation occurs once* [FO16]
  - We can efficiently recognize the tractable queries

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]
- Dichotomy for **RA**$(\sigma, \pi, \bowtie, \rho, -)$, *each relation occurs once* [FO16]
  - We can efficiently recognize the tractable queries
- **Open:** Dichotomy for full RA (FO)?

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]
- Dichotomy for **RA**$(\sigma, \pi, \bowtie, \rho, -)$, *each relation occurs once* [FO16]
  - We can efficiently recognize the tractable queries

- **Open:** Dichotomy for full RA (FO)?
- **Open:** Dichotomy for natural restricted fragments?
  - e.g., semijoin algebra (guarded FO) **RA**$(\sigma, \pi, \ltimes, \rho, \cup, -)$?

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]
- Dichotomy for **RA**$(\sigma, \pi, \bowtie, \rho, -)$, *each relation occurs once* [FO16]
  - We can efficiently recognize the tractable queries
- **Open:** Dichotomy for full RA (FO)?
- **Open:** Dichotomy for natural restricted fragments?
  - e.g., semijoin algebra (guarded FO) **RA**$(\sigma, \pi, \ltimes, \rho, \cup, -)$?
- **Open:** Dichotomy for (U)CQs on *Block-Independent DBs* (**BID**)?
  - BID: randomly select ≤ 1 tuples from each block of tuples

Dichotomies in Complexity: Known and Unknown

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- Dichotomy for UCQs, or **RA**$(\sigma_=, \pi, \bowtie, \rho, \cup)$: for every Q, evaluation is either in PTime or #P-hard [DS12]
- Dichotomy for **RA**$(\sigma, \pi, \bowtie, \rho, -)$, *each relation occurs once* [FO16]
  - We can efficiently recognize the tractable queries
- **Open:** Dichotomy for full RA (FO)?
- **Open:** Dichotomy for natural restricted fragments?
  - e.g., semijoin algebra (guarded FO) **RA**$(\sigma, \pi, \ltimes, \rho, \cup, -)$?
- **Open:** Dichotomy for (U)CQs on *Block-Independent DBs* (**BID**)?
  - BID: randomly select ≤ 1 tuples from each block of tuples
- **Open:** Dichotomies in the presence of *FDs*?

Problem 2: Approximate Query Evaluation

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

Problem 2: Approximate Query Evaluation

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- **FPAS** for Q: Numerical algorithm $A(D, p, \epsilon)$ such that:

$$\frac{\pi_Q(D,p)}{(1+\epsilon)} < A(D,p,\epsilon) < (1+\epsilon)\pi_Q(D,p)$$

  - Terminates in polynomial time in the size of (D, p) and in $\frac{1}{\epsilon}$

Problem 2: Approximate Query Evaluation

Evaluating Q: given (D, p), compute $\pi_Q(D, p)$

- **FPAS** for Q: Numerical algorithm $A(D, p, \epsilon)$ such that:

$$\frac{\pi_Q(D, p)}{(1+\epsilon)} < A(D, p, \epsilon) < (1+\epsilon)\pi_Q(D, p)$$

  - Terminates in polynomial time in the size of (D, p) and in $\frac{1}{\epsilon}$

- **FPRAS** for Q: *Randomized* $A(D, p, \epsilon)$ such that:

$$\Pr_A \left[\frac{\pi_Q(D, p)}{(1+\epsilon)} < A(D, p, \epsilon) < (1+\epsilon)\pi_Q(D, p) \right] > 0.99$$

Approximate Evaluation: Known and Unknown

- Every UCQ has an FPRAS
  - Via the *Karp-Luby estimator* [KL83]

Approximate Evaluation: Known and Unknown

- Every UCQ has an FPRAS
  - Via the *Karp-Luby estimator* [KL83]
- Some fragments of UCQ-minus-UCQ have FPRAS, while some are hard to approximate [KRT11]

Approximate Evaluation: Known and Unknown

- Every UCQ has an FPRAS
  - Via the *Karp-Luby estimator* [KL83]
- Some fragments of UCQ-minus-UCQ have FPRAS, while some are hard to approximate [KRT11]
- **Open**: Dichotomies for approximation in RA, or popular fragments with negation
  - Important special cases (arise in translation from SRL, e.g., MLN): universal FO, full dependencies (e.g., full TGDs, EDGs)

Problem 3: Most Probable Database (MPD)

- Let Q be a boolean query (now viewed as a constraint)
- The **MPD** problem for Q:

Given (D, p), compute $\text{argmax}_E \{\Pr(E \mid D, p) \mid E \models Q\}$

Problem 3: Most Probable Database (MPD)

- Let Q be a boolean query (now viewed as a constraint)
- The **MPD** problem for Q:

Given (D, p), compute $\text{argmax}_E \{\Pr(E \mid D, p) \mid E \models Q\}$

  - Also known as: Maximum A-Posteriori (MAP), Most-Probable Explanation (MPE)
  - Again, arises in translation from SRL

Problem 3: Most Probable Database (MPD)

- Let Q be a boolean query (now viewed as a constraint)
- The **MPD** problem for Q:

Given (D, p), compute $\text{argmax}_E \{\Pr(E \mid D, p) \mid E \models Q\}$

  - Also known as: Maximum A-Posteriori (MAP), Most-Probable Explanation (MPE)
  - Again, arises in translation from SRL
- A dichotomy (PTime/NP-hard) is known for the class of unary FDs [GVdBS14]

Problem 3: Most Probable Database (MPD)

- Let Q be a boolean query (now viewed as a constraint)
- The **MPD** problem for Q:

Given (D, p), compute $\text{argmax}_E \{\Pr(E \mid D, p) \mid E \models Q\}$

  - Also known as: Maximum A-Posteriori (MAP), Most-Probable Explanation (MPE)
  - Again, arises in translation from SRL
- A dichotomy (PTime/NP-hard) is known for the class of unary FDs [GVdBS14]
- **Open**: Other / more expressive classes of constraints (e.g., universal FO, full dependencies)?

Questions?

References I

- Nilesh N. Dalvi and Dan Suciu, *Efficient query evaluation on probabilistic databases*, VLDB, Morgan Kaufmann, 2004, pp. 864–875.
- ______, *The dichotomy of probabilistic inference for unions of conjunctive queries*, J. ACM **59** (2012), no. 6, 30.
- Robert Fink and Dan Olteanu, *Dichotomies for queries with negation in probabilistic databases*, ACM Trans. Database Syst. **41** (2016), no. 1, 4:1–4:47.
- Eric Gribkoff, Guy Van den Broeck, and Dan Suciu, *The most probable database problem*, Proceedings of the First International Workshop on Big Uncertain Data (BUDA), 2014.
- Abhay Kumar Jha and Dan Suciu, *Probabilistic databases with markoviews*, PVLDB **5** (2012), no. 11, 1160–1171.

References II

Richard M. Karp and Michael Luby, *Monte-carlo algorithms for enumeration and reliability problems*, 24th Annual Symposium on Foundations of Computer Science, Tucson, Arizona, USA, 7-9 November 1983, IEEE Computer Society, 1983, pp. 56–64.

Sanjeev Khanna, Sudeepa Roy, and Val Tannen, *Queries with difference on probabilistic databases*, PVLDB **4** (2011), no. 11, 1051–1062.