Technical Perspective: Probabilistic Data with Continuous Distributions

Dan Olteanu
Department of Informatics, University of Zurich
dan.olteanu@uzh.ch

The paper entitled "Probabilistic Data with Continuous Distributions" overviews recent work on the foundations of *infinite* probabilistic databases [3, 2]. Prior work on probabilistic databases (PDBs) focused almost exclusively on the *finite* case: A finite PDB represents a discrete probability distribution over a finite set of possible worlds [4]. In contrast, an infinite PDB models a continuous probability distribution over an infinite set of possible worlds. In both cases, each world is a finite relational database instance. Continuous distributions are essential and commonplace tools for reasoning under uncertainty in practice. Accommodating them in the framework of probabilistic databases brings us closer to applications that naturally rely on both continuous distributions and relational databases.

The infinite sample space of possible worlds raises significant technical challenges. I was therefore excited to learn from this work how to address the foundational questions of the representation and querying of probabilistic data in an elegant and technically sound way:

- how to represent succinctly an infinite set of possible worlds in finite space;
- how to get the right semantics for queries over infinite PDBs that natural extends the finite case.

These questions require a deep mix of mathematics and database theory, which can be easily intimidating. This overview paper is to be commended for its approachable style focused on explaining the challenges one at a time and how they were overcome.

Representation. The first question is to effectively represent continuous distributions over an infinite set of possible worlds. The overview paper shows how the formalisms of tuple-independent PDBs [4], block-independent disjoint PDBs [4], and Generative Datalog [1] can be adapted to the infinite setting.

One major difficulty when defining probabilities on uncountable spaces is that we cannot assign a well-defined probability to all subsets of the space, but only to subsets that are *measurable*. Topological Polish spaces and σ-algebras are used to define the event space in a sufficiently generic way, which only depends on the database schema and the domains of the attributes in the schema, while making sure that all sets of interest are measurable, such as those defined by views over PDBs.

Querying. The second question is what would be a well-defined semantics for views that map between infinite PDBs. It turns out that such a semantics is subject to measurability, now lifted to view mappings with respect to the σ-algebras of input and output PDBs. A remarkable result is that the views expressible in relational calculus with aggregation and in Datalog (including variants such as Inflationary Datalog and Least Fixed-Point Logic) are measurable and therefore admit a well-defined semantics over infinite PDBs [3].

The paper also gives a sound semantics to Generative Datalog with continuous distributions that naturally generalises the known case of discrete distributions [2]. Generative Datalog is a declarative probabilistic programming language for relational data, where the Datalog rules may specify distributions at the place of query variables in the head [1]. Deterministic relational data is fed into a generative model that defines a probability distribution over possible database instances. Whenever the rule body is satisfied, the head fires and a new sample is generated for the rule head. In case of continuous distributions, repeatedly sampling under the same satisfaction of the rule body may lead to non-termination. A further challenge is ensuring the order of rule applications is immaterial, which is essential for language declarativeness.

The work overviewed in the paper mirrors development on the foundations of finite PDBs: It investigates representation formalisms and their closure under various query languages. This is a prerequisite for studies on: the tractability of query evaluation on infinite PDBs; exact and approximate query evaluation algorithms; and systems for managing relational data with continuous probability distributions.

1. REFERENCES

[1] V. Bárány, B. ten Cate, B. Kimelfeld, D. Olteanu, and Z. Vagena. Declarative probabilistic programming with datalog. *TODS*, 42(4):22:1–22:35, 2017.
[2] M. Grohe, B. L. Kaminski, J. Katoen, and P. Lindner. Generative datalog with continuous distributions. In *PODS*, pages 347–360, 2020.
[3] M. Grohe and P. Lindner. Infinite probabilistic databases. In *ICDT*, pages 16:1–16:20, 2020.
[4] D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.