Leveraging the Structure of Uncertain Data
Antoine Amarilli
joint work with Pierre Bourhis and Pierre Senellart

Probabilistic Databases
A dating website uses machine learning to classify users as active/inactive, and chats as flirtatious or not.

We assume independence across all these facts
→ Cannot represent correlations, like "Monoamorous users are flirtatious with ≤1 person"

Results

Problem Statement: Which structural hypotheses on probabilistic data make query evaluation tractable?

Query Evaluation

Query: are there active users engaged in flirtatious chat?
On deterministic data, this is easy:
→ YES
→ NO

On probabilistic data, evaluation becomes much harder:
→ The task is intractable (#P-hard) for many queries even when the query is fixed (i.e., in data complexity)
→ What can we do? (especially on simple, realistic data?)

Lineages

Lineage \( \varphi \) of an arbitrary Boolean query \( q \) on database \( D \):
\( \varphi \) is a Boolean formula (or circuit) on the facts of \( D \) such that \( D' \subseteq D \) satisfies \( q \) iff \( \varphi \) holds for the valuation of \( D' \).

Our main technical results:
- Lineage circuits for tree automata on uncertain trees can be computed in \( O(n) \)
- Extends to bounded treewidth data
- The circuit has low treewidth itself so probability computation is 1

Ongoing work with Mikaël Monet, Silviu Maniu, and Pierre Senellart

- Computing tree decompositions on real datasets

The Paris road network (4.3M nodes and 5.4M edges) has treewidth ≤ 521 (computed using heuristics)

Most of the network is covered by a partial decomposition of width 5

- Tractability in combined complexity for restricted queries

Our method on low-treewidth data is linear in the data but hides high complexity in the query
→ Lower bound: Tractable complexity in TID data and query seems unlikely, even for tree-shaped queries and data
→ Can we compute lineages more efficiently in some cases?