Tractable Lineages on Treelike Instances: Limits and Extensions

Antoine Amarilli[1], Pierre Bourhis[2], Pierre Senellart[1,3]
[1]Télécom ParisTech, Université Paris-Saclay; [2]CRIStAL, UMR 9189, CNRS, Université Lille 1; [3]IPAL, CNRS, NUS

Query Evaluation on Probabilistic Instances

- **Tuple Independent Database (TID)** instances:

Each tuple is **present** or **absent** with given probability assuming **independence** across tuples

Example TID $\mathcal{I}$

R		
a	b	100%
b	c	40%
c	a	75%

Probability distribution

30%	10%	45%	15%
a b	a b	a b	a b
b c	b c		
c a		c a	

- **Boolean query:** e.g., conjunctive query (CQ)

Example: Q: $\exists xyz\ \mathbf{R}(x, y)\ \mathbf{R}(y, z)\ \mathbf{R}(z, x)$

- **Probabilistic query evaluation (PQE):**

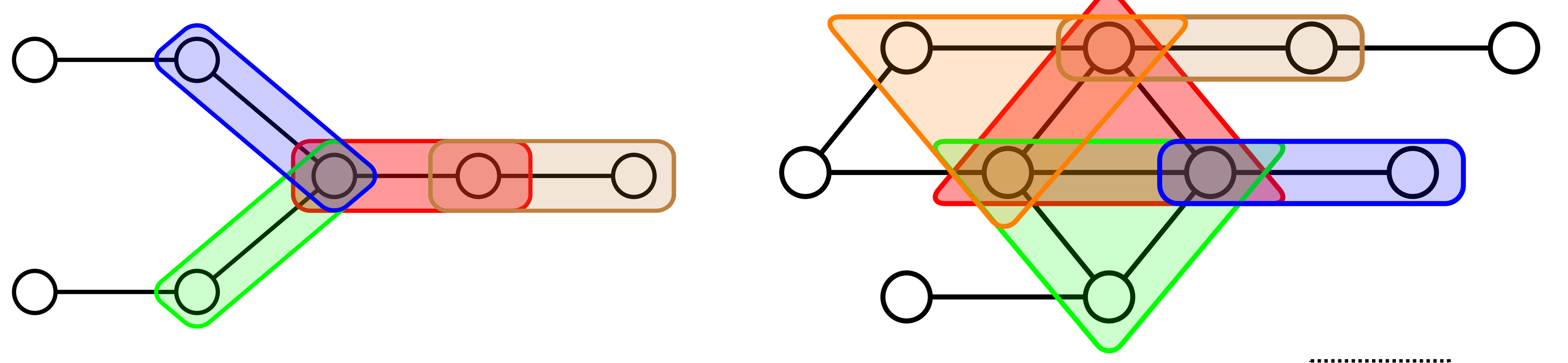
compute the probability that a TID $\mathcal{I}$ satisfies a query Q
Data complexity: Q is fixed, input is the TID $\mathcal{I}$
Example: there is 30% probability that $\mathcal{I}$ satisfies Q

→ **Existing work** (Dalvi & Suciu, [DS12]):
PQE is **intractable** (#P-hard) for CQs, except **safe** CQs

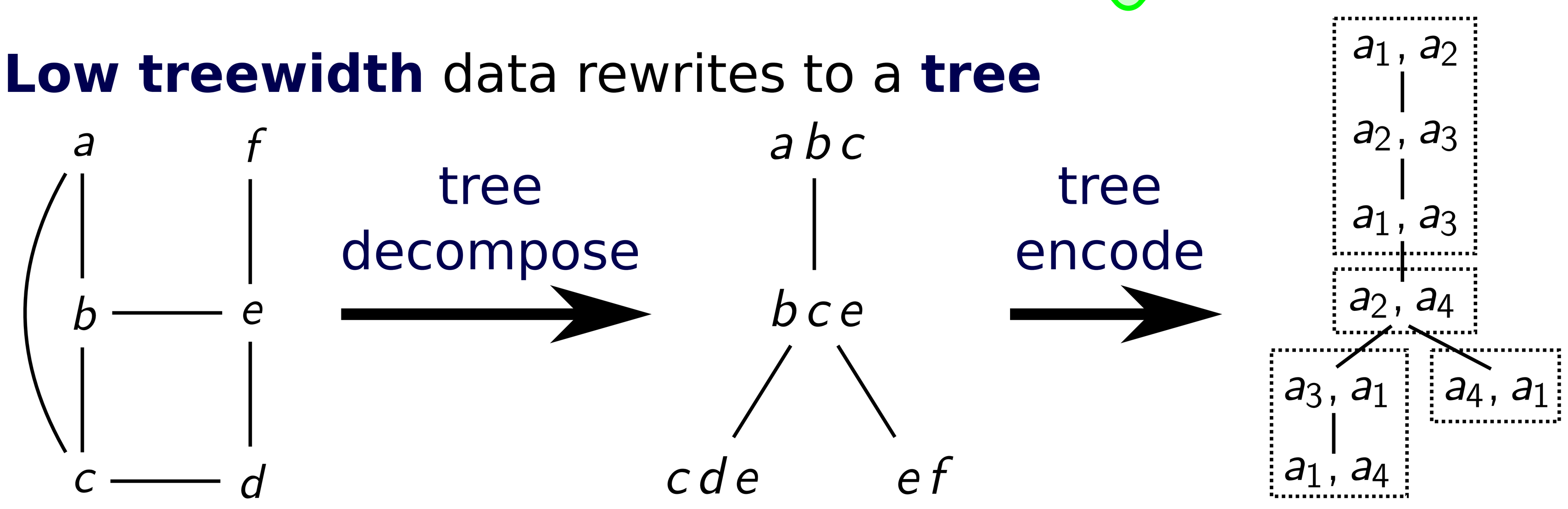
Restricting Instance Structure with Treewidth

Treewidth: Measures how much the data is close to a **tree**

We can decompose **trees** ...and **low treewidth data** too:



Low treewidth data rewrites to a **tree**



Courcelle's theorem: Monadic second-order (MSO) queries can be evaluated by a tree automaton on the tree encoding:
→ Efficient on bounded-treewidth, **non-probabilistic** data

→ This extends to **probabilistic query evaluation** [ABS15]
→ Can we go **beyond** bounded treewidth? (e.g., cliquewidth)

Dichotomy Result:

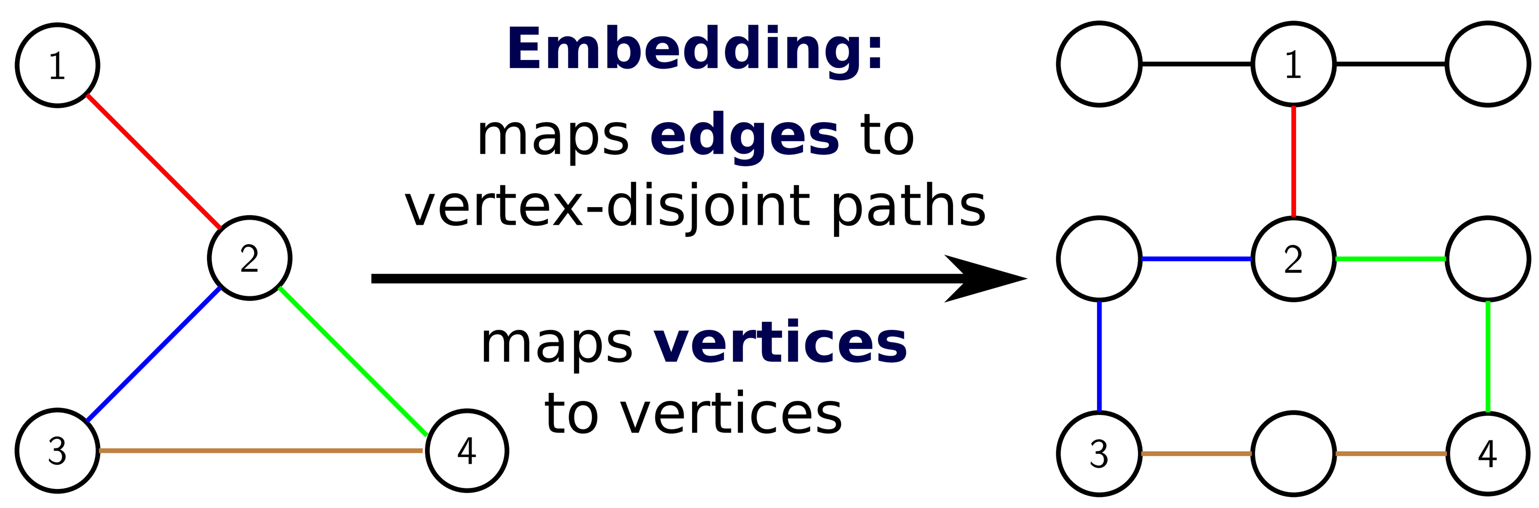
- PQE is **linear-time** (up to arithmetics) for **MSO** on any **bounded-treewidth** TID instance family [ABS15]
- PQE is **#P-hard** (under RP reductions) for **FO** on any **unbounded-treewidth** constructible graph family
  → Similar hardness results with MSO for non-probabilistic evaluation and counting, extending [GHL⁺14]

Other PQE Results:

- Tractability of MSO on **bounded-treewidth** instances can be explained by **lineages** (d-DNNFs)
- Tractability of some **safe CQs** (inversion-free) can be explained via **instance rewritings**

Lower Bounds

- **#P-hard problem:** count **matchings** of graph G
  → Embed G as a **topological minor** in the family



Theorem [CC14]: There is a constant c such that
for any **planar** graph G of size n with max **degree** 3,
for any graph H of **treewidth** $\geq n^c$,
we can **embed** G as topological minor of H in **RP time**

- Pick H from the **unbounded-treewidth** graph family
  Constructibility: we can build H in time Poly(n^c)
  → Set **probabilities** on H to give a **subdivision** of G

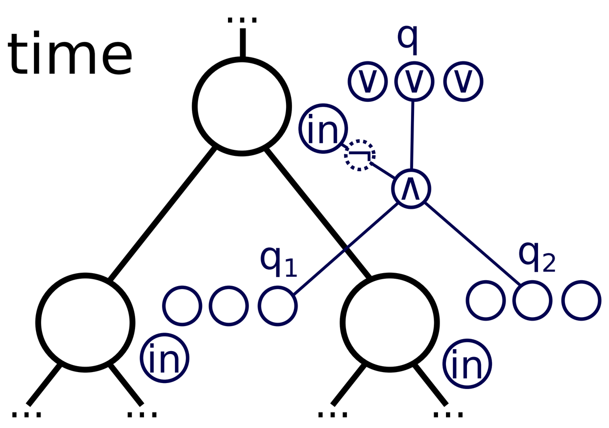
- Code hard problem as an **FO query** to reduce to PQE

Lower bounds also for **UCQs with inequalities:**
inexistence of concise **OBDD** representations

Lineages

Lineage φ of a Boolean **query** Q on **instance** $\mathcal{I}$:
φ is a **Boolean function** whose variables are the facts of $\mathcal{I}$
such that $\mathcal{I}' \subseteq \mathcal{I}$ satisfies Q **iff** φ holds for the valuation for $\mathcal{I}'$
→ The **probability** of Q on $\mathcal{I}$ is the **probability** of φ

Lemma [ABS15]: We can build in linear time
a **circuit** that captures the **lineage**
of a **tree automaton** on an input tree
→ Extends to **MSO** on **treelike data**



Further: efficient PQE using the circuit
→ The circuit is **bounded-treewidth:** use message passing [ABS15]
→ The circuit is a **d-DNNF**: exclusive OR, independent AND

Safe Queries

A CQ is **inversion-free** [JS13] if:

- It is **hierarchical**: under $\exists x$, each atom contains x
  Example: $\exists x\ \mathbf{R}(x, y)\ \exists y\ \mathbf{S}(y)$ but not $\exists xy\ \mathbf{R}(x)\ \mathbf{S}(x, y)\ \mathbf{T}(y)$

- Each relation **R** has an **order** on its attributes s.t.
  all variables of each **R**-atom were **quantified** in that order
  Example: $\exists xy\ \mathbf{R}(x, y)\ \mathbf{S}(y, x)$ but not $\exists xy\ \mathbf{R}(x, y)\ \mathbf{R}(y, x)$

Theorem [JS13]: PQE for **inversion-free** CQs is in **PTIME**

→ We explain this using **lineage-preserving rewritings**:
$\mathcal{I}$ **rewrites** to $\mathcal{I}'$ for Q if Q has **same provenance** on $\mathcal{I}$ and $\mathcal{I}'$

Theorem: Up to ranking, for any **inversion-free** CQ Q,
any instance $\mathcal{I}$ rewrites to some $\mathcal{I}'$ of **bounded treewidth**

References

[ABS15]: A. Amarilli, P. Bourhis, P. Senellart
Provenance Circuits for Trees and Treelike Instances, ICALP'15
[CC14]: C. Chekuri, J. Chuzhoy
Polynomial Bounds for the Grid-Minor Theorem, STOC'14
[DS12]: N. Dalvi, D. Suciu
The Dichotomy of Probabilistic Inference for Unions of Conjunctive Queries. JACM, 2012
[GHL⁺14]: R. Ganian, P. Hliněný, A. Langer, J. Obdržálek, P. Rossmanith, S. Sikdar
Lower Bounds on the Complexity of MSO1 Model-Checking. JCSS, 2014
[JS13]: A. Jha, D. Suciu
Knowledge Compilation Meets Database Theory: Compiling Queries to Decision Diagrams. TCS, 2013