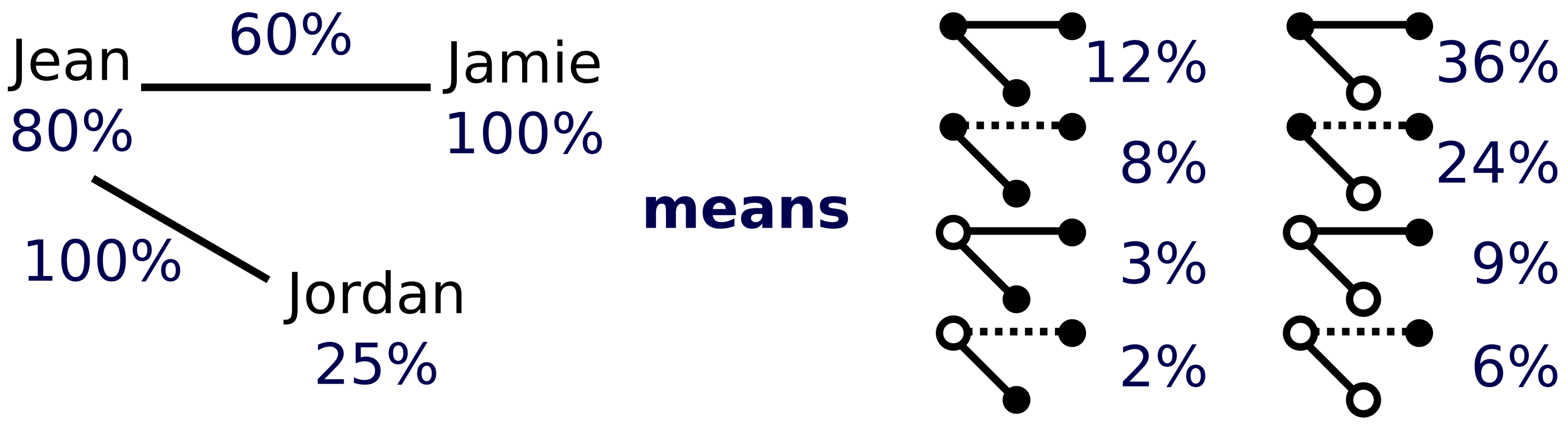
Leveraging the Structure of Uncertain Data

Antoine Amarilli
joint work with Pierre Bourhis and Pierre Senellart
Télécom ParisTech, CNRS LTCI, Université Paris-Saclay

Probabilistic Databases

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



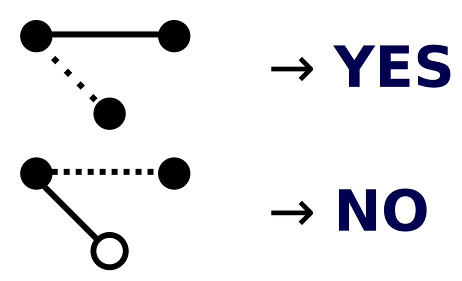
We assume **independence** across all these facts

→ Cannot represent **correlations**, like "*Monoamorous users are flirtatious with* **≤1** *person*"

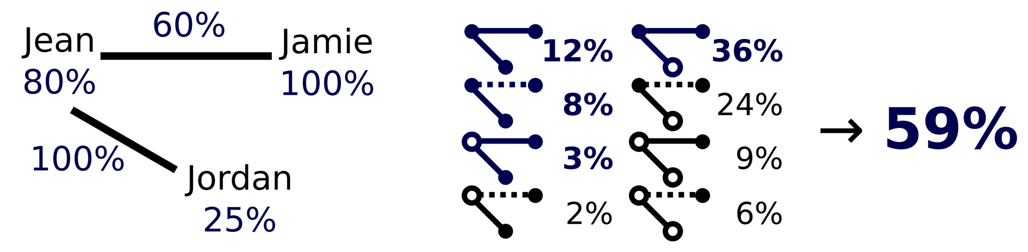
Query Evaluation

Query: are there **active users** engaged in **flirtatious** chat?

On **deterministic** data, this is **easy**:



On **probabilistic** data, evaluation becomes **much harder!**



Can we compute **efficiently** the probability of a **query**?

→ The task is **intractable** (#P-hard) for many queries even when the query is fixed (i.e., in **data complexity**)

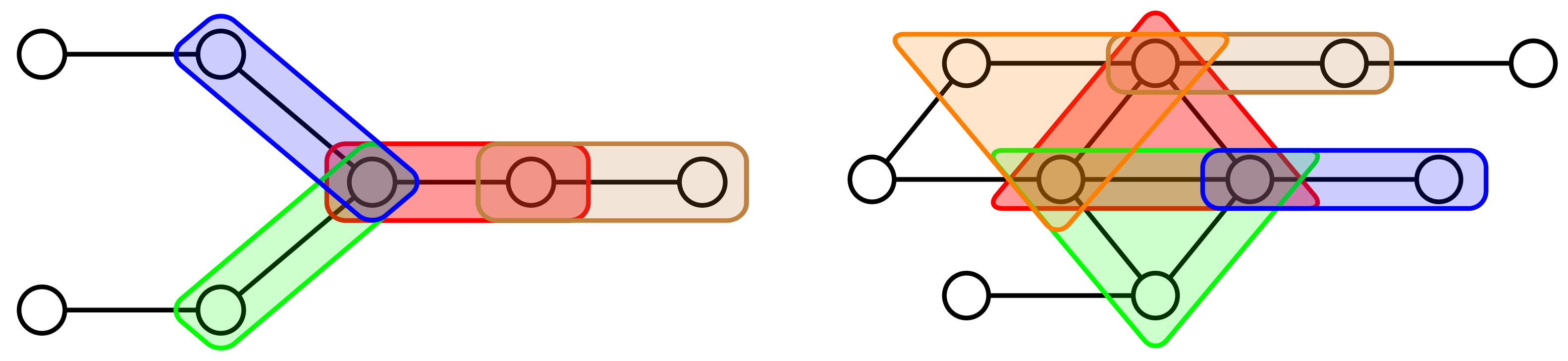
→ **What can we do?** (especially on **simple**, realistic data?)

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

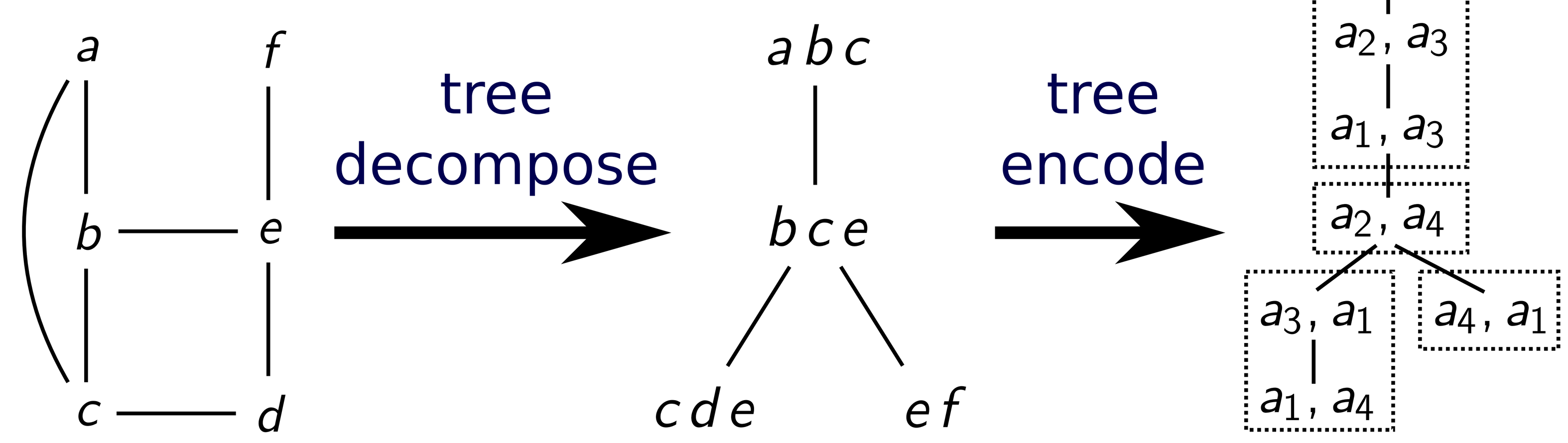
Treewidth

A **measure** of how much the data is similar to a **tree**

We can decompose **trees** ...and **treelike data** too:



Low treewidth data rewrites to a **tree**



Courcelle's theorem: Monadic second-order queries are **tractable** to evaluate on **non-probabilistic data** using **tree automata** on the tree encoding

Results

Our **main result**:

> *For any monadic second-order* ***query*** *q and integer k*
> *for any input TID* ***database*** *D of treewidth ≤ k*
> *we can compute in O(|D|) the* ***probability*** *of q on D*
> *(up to polynomial arithmetic costs)*

Extensions:

→ Also for low-treewidth **correlations** (e.g. **mutually exclusive** database facts)

→ Also for expressive **lineages** ($\mathbb{N}[X]$-provenance)

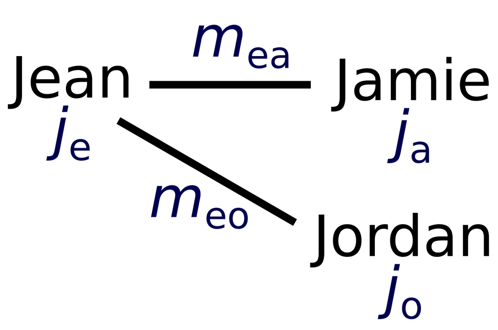
Lower bound: Query evaluation on TID is **intractable** if the treewidth is not bounded (under some technical conditions)

References

Antoine Amarilli, Pierre Bourhis, Pierre Senellart:
- *Provenance Circuits for Trees and Treelike Instances*, ICALP'15
- *Tractable Lineages on Treelike Instances: Limits and Extensions*, PODS'16

Lineages

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

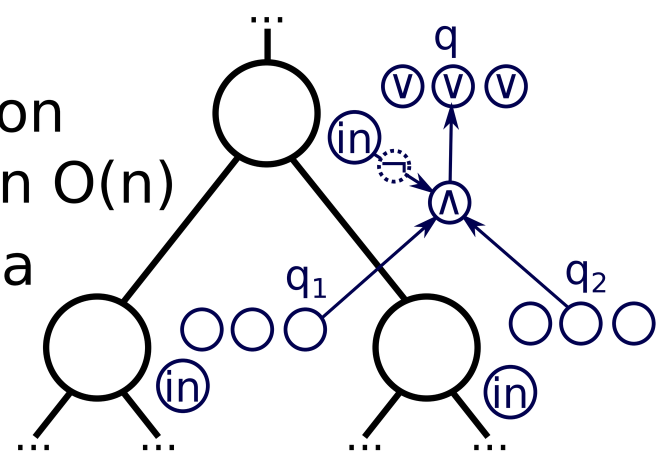


Is there a flirtatious pair of active users?

Lineage: $j_e \wedge ((m_{ea} \wedge j_a) \vee (m_{eo} \wedge j_o))$

Our main technical results:

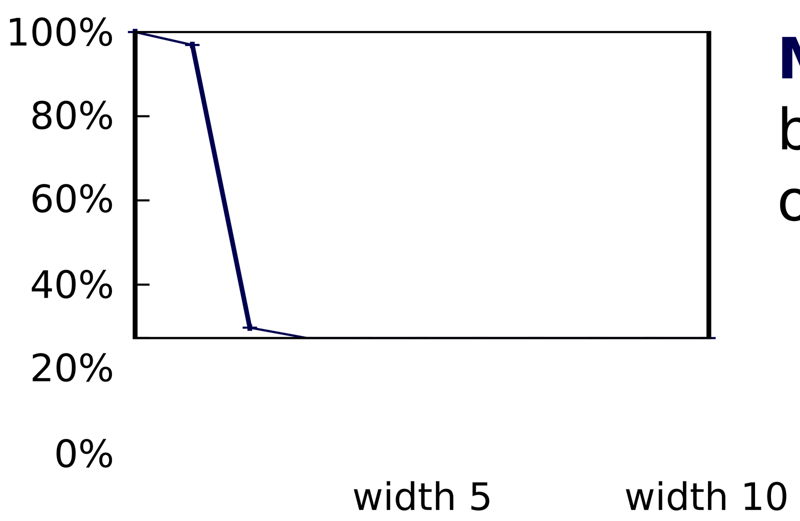
- Lineage circuits for **tree automata** on **uncertain trees** can be computed in O(n)
- Extends to **bounded treewidth** data
- The circuit has **low treewidth** itself so probability computation is 1



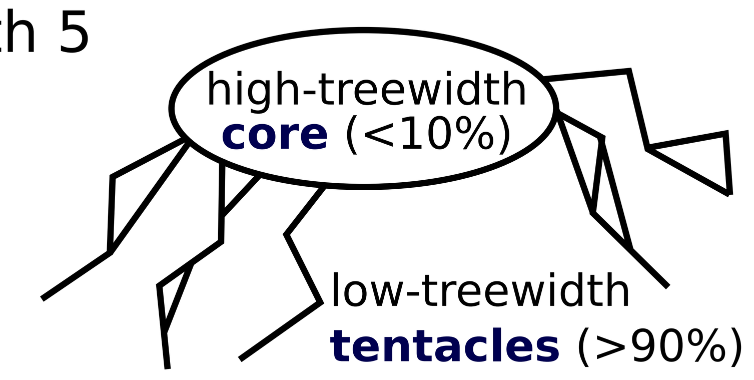
Ongoing work with Mikaël Monet, Silviu Maniu, and Pierre Senellart

- 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



→ Answer queries with **uncertainty** (e.g., RER trip time) with **tree automata** on the tentacles and **sampling** on the core

- 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?