

# MSc Internship

## What make probabilistic data efficiently queriable?

### Topic description

Probabilistic databases are compact representations of probability distributions over regular databases. A number of models have been proposed for probabilistic data, both in the relational [1] and the XML [2] settings. Evaluating a Boolean query over a probabilistic database amounts to computing the probability that this Boolean query matches in the probability distribution. One crucial question is whether query evaluation remains tractable on probabilistic databases.

A number of research works have looked at characteristics of queries that may make them tractable: thus, queries without self-joins are tractable over tuple-independent databases if and only if they are hierarchical [3], while tree-pattern queries with a single join are tractable if and only if they are equivalent to a join-free query [4].

The objective of this internship is to take the problem from the other side: identifying classes of data for which queries are tractable. One direction is to look at bounding the treewidth [5] of the data; another is to try finding join patterns that make querying easy [6].

### Supervision

This 6 month Master's internship will be supervised by Pierre Senellart at Télécom ParisTech. The research takes part in Serge Abiteboul's Webdam project on the *Foundations of Web Data Management*.

### References

- [1] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.
- [2] Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity, September 2011. Preprint available at <http://pierre.senellart.com/publications/kimelfeld2012probabilistic.pdf>.
- [3] Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 16(4), 2007.
- [4] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Value joins are expensive over (probabilistic) XML. In *LID*, 2011.
- [5] Neil Robertson and P. D. Seymour. Graph minors. III. Planar tree-width. *Journal of Combinatorial Theory, Series B*, 36(1), 1984.
- [6] Dan Olteanu, Jiewen Huang, and Christoph Koch. Approximate confidence computation in probabilistic databases. In *ICDE*, pages 145–156, 2010.