M2 Internship Proposal:
Lower bounds on provenance and counting for recursive queries

Antoine Amarilli

One central question in database theory is *query evaluation*: designing efficient algorithms to evaluate queries on databases, and understanding the computational complexity of this problem. Traditionally, database theory investigated query languages inspired by the relational algebra, but recent research studies *recursive queries*, in particular *reachability queries* and *regular path queries*. Such queries can be posed in the setting of *graph databases*, where recursive navigation is crucial. For instance, in a knowledge graph representing relationships between companies, subsidiaries, and brands, we can write a regular path query to efficiently determine whether two given brands ultimately belong to the same parent company.

In some cases, beyond computing the query answer, we may need to compute extended information that explains how the answer was obtained. This is the notion of *provenance* [GKT07]. The provenance of a query explains how the answer depends on the input data; it can typically be represented as a Boolean circuit [DMRT14]. In some cases, provenance can even be represented into restricted circuit classes from the field of *knowledge compilation* [AC24], which may ensure the tractability of more tasks. This can include *probabilistic reasoning* (e.g., how likely is it that two companies share a parent company, if the ownership links are uncertain?), *minimal witness problems* (e.g., how many ownership links at minimum are needed to witness a given answer?), *resilience problems* [AGMM25], etc.

The topic of this internship is to study the computation of provenance for recursive queries over graph databases and its representation in circuit classes from knowledge compilation. Specifically, the project will focus on the very general class of *homomorphism-closed queries*, i.e., queries that are preserved under homomorphisms. Indeed, for homomorphism-closed queries that are not expressible without recursion, it was recently shown that probabilistic reasoning is generally intractable [AC22]. The goal of this internship is to study whether these techniques can be used to show lower bounds on the size of provenance representations in tractable circuit formalisms. In particular, the internship will focus on the following directions:

- Can the results of [AC22] be generalized to show lower bounds on the size of *structured* circuit representations of the provenance of unbounded homomorphism-closed queries?
- Can such bounds be extended in the case of arbitrary-arity data, beyond the arity-two databases studied in [AC22]?
- Can some lower bounds be shown without assuming structuredness, e.g., for so-called *DNNF circuits*? indeed, such bounds have recently been shown for some specific regular path queries [AvBGM25], but it is unclear whether such techniques can be generalized.

- Are there connections to minimal witness problems, e.g., those recently studied for modular walks in [AGW25]? or to resilience problems [AGMM25]?

Supervision and environment. This proposal is for a master internship (M2 level, for a duration of 4–6 months), expected to start in Spring 2025. The internship will be supervised by Antoine Amarilli[1] (Advanced Research Position at Inria). The internship will take place in the LINKS team of the Inria center at University of Lille, in the North of France. The LINKS team focuses on logics, algorithms, formal language theory, and database theory, and offers a dynamic environment for research on these topics. Applications should be sent by email to: `a3nm@a3nm.net`.

References

[AC22] Antoine Amarilli and İsmail İlkan Ceylan. The dichotomy of evaluating homomorphism-closed queries on probabilistic graphs. *LMCS*, 2022.

[AC24] Antoine Amarilli and Florent Capelli. Tractable circuits in database theory. *SIGMOD Rec.*, 53, 2024.

[AGMM25] Antoine Amarilli, Wolfgang Gatterbauer, Neha Makhija, and Mikaël Monet. Resilience for regular path queries: Towards a complexity classification. Under review, 2025.

[AGW25] Antoine Amarilli, Benoît Groz, and Nicole Wein. Edge-minimum walk of modular length in polynomial time. In *ITCS*, 2025. To appear.

[AvBGM25] Antoine Amarilli, Timothy van Bremen, Octave Gaspard, and Kuldeep S. Meel. Approximating queries on probabilistic graphs. Under review, 2025.

[DMRT14] Daniel Deutch, Tova Milo, Sudeepa Roy, and Val Tannen. Circuits for Datalog provenance. In *ICDT*, volume 3, 2014.

[GKT07] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, 2007.

[1] `https://a3nm.net/`