Master thesis subject: Computational complexity of graph data exchange

Equipe BD, Liris, University of Lyon 1

Responsable HDR Angela Bonifati (Liris)

Encadrants Angela Bonifati (Liris), Radu Ciucanu (University of Oxford), Emmanuel Coquery (Liris) and Romuald Thion (Liris)

Emails pour contact angela.bonifati@univ-lyon1.fr, radu.ciucanu@cs.ox.ac.uk, emmanuel.coquery@univ-lyon1.fr, romuald.thion@univ-lyon1.fr

Context Data exchange is the task of transforming data structured under a source schema S into data structured under a target schema T in such a way that all constraints in a fixed set of source-to-target constraints $\mathcal{M}_{st}$ and in a fixed set of target constraints $\mathcal{M}_t$ are satisfied. A multifaceted investigation of data exchange has been carried out in the past decade [4]. The two key problems in data exchange are: (i) the existence of solutions problem, i.e. given a setting $\Omega = (S, T, \mathcal{M}_{st}, \mathcal{M}_t)$ and an instance I of S, decide whether there exists a solution for I under Ω, and (ii) the query answering problem consists of deciding whether a given tuple of constants belongs to $certain_\Omega(Q, I)$ or not, where the *certain answers of* Q *w.r.t.* I *under* Ω, denoted $certain_\Omega(Q, I)$, are the answers that hold for all solutions i.e., the set $\bigcap\{Q(J) \mid J \in Sol_\Omega(I)\}$.

All the complexity results in data exchange deal mainly with data complexity, i.e., assume that both Ω and Q are fixed.

The existence of solutions problem is undecidable for schema mappings in which $\mathcal{M}_t$ are arbitrary tuple-generating dependencies (tgds) and equality-generating dependencies (egds) and is in PTIME for schema mappings in which $\mathcal{M}_t$ is the union of a weakly acyclic set of tgds and a set of egds.

Finding certain answers involves computing the intersection of a (potentially) infinite number of sets. This strongly suggests that computing certain answers for arbitrary FO queries is an undecidable problem. This does not preclude, however, the existence of interesting classes of queries for which the problem of computing certain answers is decidable, and even tractable. Let Ω = (S, T, $\mathcal{M}_{st}$, $\mathcal{M}_t$) be a data exchange setting, such that $\mathcal{M}_t$ consists of a set of egds and a weakly acyclic set of tgds, and let Q be a union of conjunctive queries.

Then, the problem of computing certain answers for Q under Ω can be solved in polynomial time.

The Problem In our previous work [3], we have investigated the two problems of interest in the case of relational-to-graph data exchange with target constraints. In such a case, the data exchange setting $\Omega = (S, \Sigma, \mathcal{M}_{st}, \mathcal{M}_t)$ consists of a source schema S, a target graph database that is simply an alphabet of symbols Σ and $\mathcal{M}_{st}$ and $\mathcal{M}_t$ that are a fixed set of source-to-target constraints and a fixed set of target constraints, respectively. We have focused on the query complexity existence of solutions and query answering instead of considering the data complexity, thus by focusing on the case of a fixed source instance I (hence, with a query Q and a setting Ω that are not fixed).

We have shown the intractability of the problems of interest and we proved NP lower bounds for the problems studied under particular cases (NRE - nested regular expressions [5] and subsets thereof in the various constraints). Our setting is novel with respect to graph-to-graph data exchange [2] in which no target constraints $\mathcal{M}_t$ were considered.

Expected Work In this master work, we would like to pursue the analysis of the complexity of the problems of interest further by investigating the following problems:

1. can upper bounds be found for the query complexity of the above problems of interest?

2. we also would like to study the computational complexity of the problems of interest in terms of data complexity rather than query complexity, as done in the data exchange literature. One important problem is the data complexity of existence of solutions problem, in which the mapping Ω is considered to be fixed. It is well known that the data complexity is polynomial in the relational case. Does it remain polynomial in the relational-to-graph case?

3. in the case of [2], they provided a definition of a universal representative, which is a graph pattern - produced by applying the chase algorithm - that possesses an homomorphism to all possible solutions to the graph data exchange problem. In [3], we have shown that no suitable universal representative as defined in [2] can be found in the presence of target constraints. We need a new definition of universal representatives that takes into account such target constraints and, correspondingly, a new chase variant to produce such universal representatives.

4. we would like to further study the complexity of the above problems for novel classes of constraints: source-to-target constraints with less expressive regular expressions (like unnested RPQs as in [1]) and target constraints of the form of *sameAs*[1] target tgds, inspired by the mappings

[1] `http://www.w3.org/wiki/WebSchemas/sameAs`

between RDF[1] datasets in the Semantic Web (indicating that two nodes in the graph with different identity are actually the same identity).

Remark We foresee a follow-up as a PhD thesis on the same subject. Both the stage and the PhD will be funded by the ANR (Projet de Recherche Collaborative 2016-2020 DataCert: Coq Deep Specification of Security Aware Data Integration - Partners: Liris, LRI (Paris Sud) and CRIStAL (Lille)).

References

[1] Guillaume Bagan, Angela Bonifati, and Benoît Groz. A trichotomy for regular simple path queries on graphs. In *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*, pages 261–272, 2013.

[2] P. Barceló, J. Pérez, and J. L. Reutter. Schema mappings and data exchange for graph databases. In *ICDT*, pages 189–200, 2013.

[3] Iovka Boneva, Angela Bonifati, and Radu Ciucanu. Graph data exchange with target constraints. In *Proceedings of the Workshops of the EDBT/ICDT 2015 Joint Conference (EDBT/ICDT), Brussels, Belgium, March 27th, 2015.*, pages 171–176, 2015.

[4] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theor. Comput. Sci.*, 336(1):89–124, 2005.

[5] J. Pérez, M. Arenas, and C. Gutierrez. nSPARQL: a navigational language for RDF. *J. Web Sem.*, 8(4):255–270, 2010.

[1] `http://www.w3.org/RDF/`