Querying the Deep Web: Back to the Foundations

Andrea Calí[1,4], Davide Martinenghi[2], Igor Razgon[1], and Martín Ugarte[3]

[1]Dept of Comp. Sci. and Inf. Syst.
Birkbeck, Univ. of London, UK

[2]Dip. di Elettr., Informaz. e Bioing.
Politecnico di Milano, Italy

[3]Web and Information Tech. Lab.
Université Libre de Bruxelles

[4]Oxford-Man Inst. of Quantitative Finance
University of Oxford, UK

{andrea,igor}@dcs.bbk.ac.uk
davide.martinenghi@polimi.it
mugartec@ulb.ac.be

Abstract. The Deep Web is the large corpus of data accessible on the Web through forms and presented in dynamically-generated pages, but not indexable as static pages, and therefore invisible to search engines. Deep Web data are usually modelled as relations with so-called access limitations, that is, they can be queried only by selecting certain attributes. In this paper we give some fundamental complexity results on the problem of processing conjunctive (select-project-join) queries on relational data with access limitations.

1 Introduction

The term *Deep Web* (also called *Hidden Web*) [7, 5, 6] refers to the data content that is created dynamically as the result of a specific search on the web. For example, when we query a White Pages website, the generated output consists of one or more pages containing the result of a query posed on an underlying database; these pages cannot be indexed by search engines. When we query whitepages.com through a form, we are forced to fill in some fields of the form, for instance the Name field; the result is then structured as a table. A Deep Web source can be naturally modelled as a relational table (or a set of relational tables) that can be queried only according to so-called *access patterns*, each of which enforces the selection on some of the attributes (which corresponds to filling the input fields in the form with values), which are called *input* attributes. Relational tables accessible through access patterns are said to have *access limitations*.

Processing structured queries over Deep Web sources is the key problem in the integration of such sources. Interestingly, when Deep Web sources are modeled, as mentioned, as relations with access limitations, answering a simple conjunctive (select-project-join) query on such sources requires, in the worst case, the evaluation of a *recursive* Datalog query plan. In such plans, values obtained as output from a source are used as input for other sources; the compatibility of values is established by assigning to each attribute of a relation a so-called

$$\begin{array}{rlrl}
\rho_1 : & q() \leftarrow \hat{r}(X,Y), \hat{s}(Z,Y) & \rho_5 : & dom(Y) \leftarrow \hat{s}(X,Y) \\
\rho_2 : & \hat{r}(X,Y) \leftarrow dom(X), r(X,Y) & \rho_5 : & dom(Y) \leftarrow \hat{r}(X,Y) \\
\rho_3 : & \hat{s}(X,Y) \leftarrow dom(X), s(X,Y) & \rho_6 : & dom(a)
\end{array}$$

Fig. 1. Datalog program for Example 1

abstract domain, which expresses the type of value (e.g. name, address etc.) as opposed to the concrete domain (e.g. string, integer etc.).

In this paper we consider conjunctive queries (CQs) on relational schemata with access limitations and the two traditional problems associated with them: *query answering* and *query containment*. Such problems have been extensively studied in the literature; however, for some cases the problem of determining the complexity is still open. We tackle some of such cases with the following results:

- We show that CQ answering under access limitations is NP-complete with respect to combined complexity; thus, the access limitations do not increase the complexity of classic query answering (without access limitations).
- We consider the problem of CQ containment under access limitations, known to be CO-NEXPTIME-complete in its general form [3, 4] and thought to be EXPTIME-complete in the case of queries without constants [2]. We first address the case of *input-only* predicates; we show that in such a case the problem is Π_2^{p}-complete. As for the hardness, we show that Π_2^{p}-hardness holds under stricter conditions: for predicates of arity $\leqslant 2$ and two abstract domains. Then we address CQ containment for (input-output) binary predicates; we conjecture that this problem is also in Π_2^{p}.

2 Query Answering

We assume the reader is familiar with the notions of relational schema and instance, conjunctive query and Datalog program — otherwise, see for instance the book of Abiteboul et al. [1]. We consider relational schemata whose predicates are annotated so as to express whether each argument/attribute is *input* (needs to be selected) or *output*; for instance, r^{iio}, of arity 3, has the first two attributes as input attributes, and the third as output.

In the presence of access limitations on the sources, queries cannot be usually evaluated as in the traditional case, as we show below. Given a conjunctive query q, a schema with access limitations (implicit), an instance D and a set I of initial constants, the answers to q, denoted $\mathrm{ans}(q, I, D)$, are obtained starting from the constants in I and extracting all possible tuples (by using the constants as input in all possible ways); with the newly obtained constants again all possible tuples are extracted, and so on, until no new tuple is extracted – see e.g. [5].

Example 1. Consider a schema with predicates r^{io} and s^{io} (which contain the facts of D), a set of initial constants $I = \{a\}$, and the Boolean CQ $q() \leftarrow r(X,Y), s(Z,Y)$. Assume there is a single abstract domain, represented by the unary predicate *dom*, associated with all attributes (arguments). The Datalog program Π_q for q is shown in Figure 1 (facts of D omitted). The query is rewritten over the *cache* relations $\hat{r}, \hat{s}$ (rule ρ_1) defined in the cache rules ρ_2 and ρ_3, which contain the facts extracted according to rules ρ_2 and ρ_3. ■

We now come to our result on the decision problem of CQ answering under access limitations; w.l.o.g., we consider Boolean CQs.

Theorem 1. *CQ answering under access limitations is* NP*-complete with respect to combined complexity.*

Proof (sketch). For membership we exhibit a non-deterministic algorithm that performs $\leqslant |D|$ steps; at each step guesses one of the $\leqslant |D|^{W \cdot |\mathcal{R}|}$ possible accesses to relations. Hardness follows from CQ answering without access limitations.

3 Query Containment

Definition 1. *Consider two CQs q_1, q_2 over a schema with access limitations, as well as a set I of initial constants such that* $\mathrm{const}(q_1) \cup \mathrm{const}(q_2) \subseteq I \subseteq \Delta$ *(*$\mathrm{const}(q)$ *denotes the constants in a query q, while Δ denotes the infinite domain of constants); we say that q_1 is* contained *in q_2 under access limitations with respect to I, denoted $q_1 \subseteq_I q_2$, if, for every database D for $\mathcal{R}$, we have* $\mathrm{ans}(q_1, I, D) \subseteq \mathrm{ans}(q_2, I, D)$.

Checking containment amounts to checking containment between two recursive Datalog programs in the special form presented in Section 2. W.l.o.g., we consider Boolean CQs as in Section 2. We first consider the case of *input-only* predicates. An input-only n-predicate r is accessed, in an instance D, with an n-tuple $\langle t \rangle$ of constants of the appropriate domain, and tells (with a Boolean result) whether $r(\langle t \rangle) \in D$. Evidently, this restricts the definition of containment to instances composed solely of constants of the initial set I. The following lemma has a rather straightforward proof. A tight hardness result follows.

Lemma 1. *CQ containment under access limitations with input-only predicates is in* Π_2^{p}.

Theorem 2. *CQ containment under access limitations with input-only predicates of arity $\leqslant 2$ and two abstract domains is* Π_2^{p}*-hard.*

Proof (sketch). The proof is by reduction from a tighter version of Generalised-Graph-Colouring [8], which is Σ_2^{p}-complete and is defined as follows: given a graph F and a positive integer k, is there a two-colouring of the vertices of F that does not contain a monochromatic (on vertices) clique of k vertices? We reduce Generalised-Graph-Colouring to non-containment under the above stated restrictions using a predicate $e/2$ for edges and a predicate $col/2$ to indicate by $col(v, c)$ that the vertex v has colour c.

As a corollary we get tight bounds for the input-only case.

Corollary 1. *CQ containment under access limitations with input-only predicates is* Π_2^{p}*-complete.*

Finally, we studied the binary case with both input and output predicates.

Theorem 3. *CQ containment under access limitations with binary predicates is in* Π_2^{p}.

Proof (sketch). The proof uses the *crayfish-chase* technique of [4] to check $q_1 \subseteq_I q_2$ in the binary case. Relying on the fact that q_2 is "blind" to pairs of atoms that are more than $|q_2|$ steps apart in a join graph, to check the existence of a counterexample for containment we guess, by means of the crayfish-chase, a polynomially bound set of atoms representing a fragment of instance that makes q_1 true; then we check whether no homomorphism maps q_2 onto such fragment.

4 Discussion

We have presented some results on our ongoing study of the fundamentals of the complexity of CQ answering and containment under access limitations. Interestingly, some of the fundamental problems have been overlooked in the literature, for instance the complexity of CQ answering under access limitations, for which we gave a tight bound. We also presented results for the input-only case, employing techniques that, we believe, pave the way to future investigations. The binary case is interesting as most knowledge representation formalisms rely on binary relations; we plan to find a tight bound for its complexity, proving our conjecture. Finally, we shall study CQ answering and containment under access limitations as well as integrity constraints expressed as ontological rules; this has applications in the intersection between the Semantic Web and the Deep Web.

Acknowledgments. Martn Ugarte and Andrea Calí acknowledge support by the EU COST action IC1302 "Keystone". Andrea Calí acknowledges partial support by the EPSRC project "Logic-based Integration and Querying of Unindexed Data" (EP/E010865/1)".

References

1. Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases.* Addison-Wesley, 1995.
2. Michael Benedikt. Personal communication, 2017.
3. Michael Benedikt, Georg Gottlob, and Pierre Senellart. Determining relevance of accesses at runtime. In *Proc. of PODS*, pages 211–222, 2011.
4. Andrea Calì and Davide Martinenghi. Conjunctive Query Containment under Access Limitations. In *Proc. of ER*, pages 326–340, 2008.
5. Andrea Calì and Davide Martinenghi. Querying data under access limitations. In *Proc. of ICDE*, pages 50–59, 2008.
6. Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In *Proc. of CIDR*, pages 44–55, 2005.
7. Jayant Madhavan, Loredana Afanasiev, Lyublena Antova, and Alon Y. Halevy. Harnessing the deep web: Present and future. In *Proc. of CIDR*, 2009.
8. Vladislav Rutenburg. Complexity of generalized graph coloring. In *Proc. of MFCS*, pages 573–581, 1986.