L-rigid Databases and the Expressibility of Incomplete Relational Query Languages

José María Turull Torres

Universidad Nacional de San Luis
turull@iamba.edu.ar, turull@unsl.edu.ar

Advisor: A. Mendelzon (University of Toronto)

Abstract

The class of *Computable Queries* ($\mathcal{CQ}$) was defined by Chandra and Harel in 1980, as functions on structures rather than functions on numbers (as recursive functions). With this formulation of the notion of a query to a relational database (db), the field of Finite Model Theory became a suitable theoretical framework for relational databases. In this framework the notion of *local expressibility* of a given logic is the class of queries which can be expressed in that logic. On the other hand, in 1991, Abiteboul and Vianu defined the *Generic Machine*, denoted as GM^{loose}, which they proved to be strictly included in $\mathcal{CQ}$. They also proved that this class of machines "behaves" as complete w.r.t. the whole class $\mathcal{CQ}$ when working on classes of *ordered* db. If we consider the structures as db instances, and if we are using a GM^{loose} machine, it means that we will not be able to compute queries such as "*give me the names of the salesmen who sell to an even number of clients*" unless our db is ordered. In the present Thesis, we study some properties of relational db which can increase the expressive power of relational languages or formalisms which are incomplete in the general case, when working on classes of db which satisfy that properties. The formalism which we first consider is the GM^{loose} machine. And the properties which we study for classes of db are *rigidity*, $\mathcal{L}$ *rigidity*, *partial rigidity*, $\mathcal{L}$ *partial rigidity* and *almost rigidity*, for different fragments $\mathcal{L}$ of First Order Logic (FO). Recall that a structure is *rigid* if its only automorphism is identity. The fragments of FO which we consider are defined by stating a bound in the number of different variables which can be used in a formula. We denote by FO^k the fragment of FO with at most k different variables, possibly reused. The reason why we use FO^k as the target logic is that Abiteboul and Vianu proved that the expressive power of the GM^{loose} model is not altered if we use dynamically

generated queries in FO^k. We prove that there is a wider set of classes of finite structures for which GM^{loose} is complete than just those in which all structures are ordered, and these are at least what we define later as *strongly FO^k rigid* classes. Roughly, this means that for every element in every structure in such a class, there is a formula in FO^k which defines that particular element in the structure. So that our first result means that what is relevant as to expressibility of queries is rigidity and not order. This was independently noted by Abiteboul and Vianu, and by Dawar. Then we consider the existence of classes of db which are rigid, but which are not FO^k rigid for any k. So, we generalize the notion of strong rigidity by allowing the bound on the amount of variables in the defining FO formulae to be an arbitrary sub-linear function on the size of the db, and we define a new class of machines which lies between GM^{loose} and $\mathcal{CQ}$, and which we call $GM^{f(n)}$. This model was independently defined in a work of Abiteboul, Papadimitriou and Vianu, in 1994, and they proved that it is not complete when $f(n)$ is sub-linear. Then we prove that, for every sub-linear function $f(n)$ the machine $GM^{f(n)}$ "behaves" as complete when working on strongly $FO^{f(n)}$ rigid classes of db. Then we conclude that strong rigidity is always a nice property, in the sense that for any given strongly rigid class of db there will always be a machine (or a language) which is not complete, but which will work as if it were on any db of that class. We consider two examples of classes of rigid graphs which we build, and for which we exhibit two different sets of properties that define their elements. One set is in FO^2 while the other (which seems to be naturally induced by the definition of the graphs) cannot be bounded by any constant. On the other hand, we explore some aspects of query computability in FO^k. We build a family of classes of graphs, which we call *Clique Intersection graphs* (CI), for which *no boolean query* can be computed in FO^k, unless it is trivial. Up to now there were only two known examples of non trivial classes with this property: the class of all structures which satisfy the k-extension axiom and the Paley graphs. However, the CI graphs have a much simpler and concrete structure. We prove this result by means of a bounded Back and Forth system of partial isomorphisms which we define in terms of a relation over a special class of induced sub-hypergraphs. We use a novel technique by defining these sub-hypergraphs as companion structures of the sub-graphs, which carry information regarding the different ways in which the sub-graphs can be expanded. Another interesting property of CI graphs is that they can be easily expanded to graphs which are rigid. Then we face a Model Theoretic issue, which is related to our main subject and which is a long standing open problem, with a different and novel approach: the definition of classes of rigid structures which are not FO^k rigid for any k, and which are "constructible". By this it is meant, in an informal sense, that one can have an intuitive image of the structures which form the class. Though it is well known that "many" classes of such structures are not FO^k rigid for any k, no constructible class is known at present. There are two independent constructions (Gurevich and Shelah in 1995, and Andréka et al in 1995) but they use either probabilistic methods or existential proofs, rather than constructive proofs, so that we cannot have an intuitive image of the structures. Regarding our Rigid CI graphs, we exhibit a set of properties with an unbounded number of variables and we *conjecture* that the class is not FO^k rigid for any k. But we consider here another approach. We wonder whether by relaxing a little bit the hypothesis, we could find such a class. For this sake we define two new properties: a structure is *partially rigid* (*pr*) if it has a non empty subset of definable elements, and a class of *pr* structures is *almost rigid* if the limit of the quotient between the number of definable elements and the size of the structure tends to 1, as the number of definable elements tends to infinity. So, we define here classes of structures, which are *constructible* (*Clique Intersection Tree structures*), and we prove that they are *strongly* $FO^{f(n)}$ *pr*, with $f(n) = n^{2/(2^{\hat{k}}-1)}$. Then we show that these classes are not strongly FO^k *pr* for any k. Moreover, these classes are *almost rigid*. We also build a hierarchy of these classes as to the bound $f(n)$. And we *conjecture* that this hierarchy is strict. This means that, for these classes of structures, there is no GM^{loose} machine which can compute the automorphism types of all the definable elements of every structure in the class while having that structure as input. Then, to prove the stated result, we first give a sufficient condition for the non FO^k *pr* for any k, of a class of *pr* structures, in terms of the equivalence of structures under FO^k. We prove the equivalence of pairs of non isomorphic CIT structures under FO^k by means of a bounded Back and Forth system of partial isomorphisms, and by using the previously stated result. We also use here auxiliary structures. Finally, we study the property of *pr* as a means to strictly increase the class of queries which can be computed or expressed by an incomplete machine or language on a given class of structures. For this sake we define a notion of *relative* completeness of a formal machine which involves two different classes of structures. We achieve completeness in the class of "small" db by evaluating queries in the class of "big" db. Then we prove that if $\mathcal{C}$ is a class of structures which is strongly $FO^{g(n)}$ *pr*, with $g(n)$ sub-linear, and if $\mathcal{C}'$ is the class of the restrictions of the structures in $\mathcal{C}$ to their respective rigid sub-domains, then $GM^{g(n)}$ is complete on $\mathcal{C}$ w.r.t. $\mathcal{C}'$. We prove also other relative completeness results with different ways of defining the substructures. This means that we can achieve completeness with a class of superstructures built in a way that adds as few elements as possible to every structure in the given class. And in this sense the property of *almost rigidity* is also important: we add "almost no" new elements to the structures.