QL: Object-oriented Queries on Relational Data

Pavel Avgustinov, Oege de Moor, Michael Peyton Jones, Max Schäfer[1]

1 **Semmle Ltd**
publications@semmle.com

Abstract

This paper describes QL, a query language particularly suited for querying complex, potentially recursive data structures. While QL compiles to Datalog and runs on a standard relational database, it provides familiar-looking object-oriented features such as classes and methods, reinterpreted in logical terms: classes are logical properties describing sets of values, subclassing is implication, and virtual calls are dispatched dynamically by lookup in the most specific class containing the receiver value. Furthermore, QL has a prescriptive type system where types actively influence program evaluation rather than just describing it. In combination, these features enable the development of concise queries based on reusable libraries, which are written in a purely declarative style, yet can be efficiently executed even on very large data sets. In particular, we have used QL to implement static analyses for various programming languages, which scale to millions of lines of code.

1 Introduction

QL is a declarative, object-oriented logic programming language for querying complex, potentially recursive data structures encoded in a relational data model. While it is a general-purpose query language, its strong support for recursively defined predicates and aggregates makes it well suited for implementing static analyses and computing software metrics. Indeed, it is in this area that QL has seen most use so far, and (instead of an abstract motivation) we will introduce the language by a concrete example drawn from it.

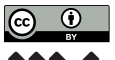
A static analysis implemented in QL is simply a query run on a special database: the database contains a relational representation of the program to analyse (encoding, say, its AST or CFG), from which the query computes a set of result tuples. A bug finding analysis, for instance, could return pairs of source locations and error messages. Since the database describes the program as it was at one particular point in time, we refer to it as a *snapshot database*. A snapshot database is created by a language-specific *extractor*. We have built extractors for various different languages based on existing compiler frontends.

As our first example of a QL query, let us consider an analysis for finding useless expressions in JavaScript programs, i.e., pure (that is, side effect-free) expressions appearing in a void context where their value is immediately discarded. Typically, this indicates a typo, for instance mistyping an assignment "`x = 42;`" as an equality check "`x == 42;`".

To identify such expressions we need to implement a purity analysis and a check to determine whether an expression appears in a void context. Fortunately, the former is already implemented in our standard QL library for JavaScript, so we can concentrate on the latter.

A simple query for finding useless expressions is shown in Listing 1. At a very high level, it breaks down into three sections:

- An `import` statement pulls in the existing QL library `javascript`, which, as its name suggests, provides general support for working with JavaScript snapshot databases.
- A predicate `inVoidContext` is defined to identify expressions in void context.



Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Listing 1 QL query for finding useless expressions in JavaScript

```
import javascript

predicate inVoidContext(Expr e) {
  exists (ExprStmt s | e = s.getExpr()) or
  exists (SeqExpr seq, int i |
    e = seq.getOperand(i) and
    (i < count(Expr op | op = seq.getOperand(_))-1 or
     inVoidContext(seq))
  )
}

from Expr e
where e.isPure() and inVoidContext(e) and not (e instanceof VoidExpr)
select e, "This expression has no effect."
```

- The main `from-where-select` clause defines the analysis itself:
  - the `from` part declares a variable `e` ranging over all expressions in the analysed program;
  - the `where` part imposes three conditions on `e`: it must be pure, appear in a void context, and not be a `void` expression, which explicitly discards the value of its operand;
  - the `select` part specifies the results to return for values of the `from` variables that pass the `where` conditions; in this case, `e` itself is returned with an explanatory message.

Taking a closer look at the definition of `inVoidContext`, it is declared as a unary predicate with a single parameter `e` of type `Expr`. `Expr` and its subtypes model JavaScript expression ASTs: for instance, `BinaryExpr` is a subclass of `Expr` representing all binary expressions, which in turn has a subclass `AddExpr` representing additions; another subclass of `Expr` is `SeqExpr`, representing sequence (or "comma") expressions with two or more operands.[1]

The body of the predicate is a first-order formula with two disjuncts. Its first disjunct says that `e` is in void context if it is the toplevel expression in an expression statement (as in our example above). The second disjunct handles the case where `e` is an operand of a sequence expression: `e` is in void context if it is not the last operand, or if the entire sequence is in void context, as determined by a recursive call to `inVoidContext`.

Judging from this example, it might look like QL is a domain-specific language for querying and navigating ASTs. But this appearance is deceptive: the classes used in this example and the navigation operations available on them are defined entirely in QL, not built into the language. As we shall see, there is nothing about QL that is specific to dealing with ASTs, or even for writing static analyses, but its object-oriented features allow the development of reusable domain-specific libraries (such as the `javascript` library and its cousins for other languages), providing a rich and convenient API for query writers.

Perhaps more surprisingly, there are not even any objects in the traditional sense of structured records with fields and methods. QL programs only work with atomic values; structured data is encoded as relational tables. For example, it is natural at first to think of the formula `e = seq.getOperand(i)` as an operation on an object `seq`, perhaps involving reading the i-th element of one of its fields holding an array of references to other objects, and then storing the result in variable `e`. In fact, though, all three variables `e`, `seq` and `i` range

[1] We use the popular ESTree encoding for JavaScript ASTs format (`https://github.com/estree/estree`), which coalesces nested sequence expressions into a single n-ary expression.

over atomic values: the latter is an integer, and the former two are *entity values*, that is, opaque identifiers representing entities modelled in the database (in this case, expressions).

In JavaScript snapshot databases, the expression AST structure of the program is encoded in a relation `exprs` containing 4-tuples (c, k, p, i), for entity values c and p and integers k and i, stating that expression c is the the i-the child of p in the AST (ignoring k for now). Class `Expr` and its subclasses define an object-oriented view of this relation; for example, `getOperand` is defined such that `e = seq.getOperand(i)` is compiled to $\exists k.$`exprs`$(\texttt{e}, k, \texttt{seq}, \texttt{i})$: no field reads, no assignments, just logic.

In particular, `e` is not an output computed from inputs `seq` and `i`: all three variables are on an equal footing, and there is not even any requirement that `e` is functionally determined by `seq` and `i` (though in this particular case it is). This becomes obvious in our use of the **count** aggregate to determine the number of operands to `seq`: `op = seq.getOperand(_)` holds for any value of `op` that is an operand of `seq` (where "`_`" is the special don't-care variable familiar from other logic languages), so `seq.getOperand(_)` behaves like a multi-valued expression. We use the aggregate to count how many of those values there are to obtain the number of operands of `seq`.

In spite of their unusual semantic underpinnings, QL classes offer very similar features to their Java counterparts. In particular, classes can have member predicates, such as the `isPure` predicate on `Expr`, which is defined in the standard QL library for JavaScript and overridden with different implementations for various subclasses of `Expr`. Calls such as `e.isPure()` are dispatched virtually, looking up the most specific applicable definitions of `isPure` based on the (runtime) value of `e`.

Like in Java and many other languages, all variables in QL have statically declared types, offering the usual benefits of enabling smart IDEs.[2] However, type declarations in QL are not just assertions to be checked by the compiler, but do, in fact, affect program semantics at runtime: the values that a variable can take during execution are restricted to those that conform to the declared type. In particular, the declared types of predicate parameters and quantified variables restrict the set of values they may range over.

As mentioned above, our example query makes use of (a small part of) the standard QL libraries for JavaScript. Just like the query, the libraries are implemented in an entirely declarative style, specifying *what* should be computed rather than *how*. In fact, QL exposes no details at all of the underlying database system on which the queries are run, and it is up to the optimiser to translate the high-level QL code into an efficiently executable query plan.

In this paper, we present the core features of QL:

- We explain the semantics of classes, member predicates and virtual dispatch, first informally (Section 2) and then more formally via a translation from a subset of the language, Core QL, to plain Datalog (Section 3).
- We discuss practical usage of QL in Section 4, and report on a case study in using QL to implement static checks for Java in Section 5.
- We put QL into context in Section 6, exploring in detail how far it matches the principles of object orientation laid down in the literature, and briefly survey related work.

[2] In fact, this is the main motivation for choosing the `from-where-select` query syntax instead of SQL's `select-from-where`: variables are declared upfront, so code completion is available in the `select` part.

2 Overview of QL

The fundamental semantic model of QL is that of Datalog: programs define a set of *intensional predicates*, one of which is a distinguished *query predicate*. They are evaluated on top of an *extensional database* (EDB), which defines a set of *extensional predicates*. While intensional predicates are defined by formulas of first-order logic (possibly involving recursion between predicates), extensional predicates are defined as explicit sets of tuples stored in the database. Unlike Prolog, Datalog does not allow the use of complex terms, so intensional predicates can only refer to values already contained in the database and cannot build up new data structures, such as lists. Like many Datalog dialects, QL somewhat relaxes this restriction by providing support for arithmetic and string operations.

The semantics of a program is the least fixpoint of its intensional predicates, that is, intensional predicates are assigned the smallest sets of tuples that satisfy their recursive definitions. Since such a fixpoint need not exist in general, QL imposes the restriction that (mutual) recursion is only allowed under an even number of negations, which is a variant of the *stratified negation* restriction used in many Datalog systems [30]. Once a fixpoint solution has been found, the set of tuples assigned to the query predicate is returned as the overall result of the program.

The grounding of QL's semantics in Datalog is not just an expository device: as explained in Section 4, our implementation compiles QL to plain Datalog, and we shall provide a precise semantics for a core calculus of QL in the next section by formalising the essential parts of that translation.

2.1 Classes

A type in QL represents a set of values, which we will call the *extent* of the type. Classes are types whose extent is defined by a unary intensional predicate called the *characteristic predicate* (or *character* for short) of the class.

There are also two kinds of *base types*, that is, types which are not themselves defined in QL: *primitive types* such as `int` or `string` are built into the language; *entity types* are defined by unary extensional predicates, whose names by convention start with an "@" character. Primitive types always have the same extent, regardless of the content of the EDB, while the extent of entity types and classes may depend on the EDB. For example, snapshot databases representing JavaScript programs defines entity types `@expr` and `@seqexpr` whose extent is, respectively, the set of all expressions and the set of sequence expressions in the represented program.

Subtyping can be thought of as set inclusion of extents: if A is a subtype of B, then the extent of A is a (not necessarily proper) subset of the extent of B.[3] For entity types the subtyping relation is given by the database schema: for instance, the schema for JavaScript snapshot databases declares `@seqexpr` to be a subtype of `@expr`, and it is up to the database system to ensure that this constraint is met at runtime. For classes, direct supertypes are specified as part of their declaration using a Java-like `extends` clause.

While entity types can only be subtypes of other entity types, classes can extend either other classes or base types. For instance, we can define a class `Digit` with the extent $\{0, 1, 2, \ldots, 9\}$:

[3] The reverse direction cannot, in general, hold: since characters are arbitrary predicates, inclusion of extents is undecidable, while our subtyping relation needs to be kept decidable.

```
class Digit extends int {
  Digit() { (int)this in [0..9] }
}
```

The `extends` class makes `Digit` a subtype of the built-in `int` type, and the character (which syntactically looks like a constructor in Java) further restricts the extent of `Digit`. The `x in [a..b]` notation is a convenience for defining ranges (note that an explicit cast is necessary when using variables with a class type in numeric operations).

Characteristic predicates can contain arbitrary QL code. For instance, we can define the class of even digits and the class of prime digits by subclassing `Digit` and performing arithmetic checks on `this`.

```
class Even extends Digit { Even() { (int)this % 2 = 0 } }

class PrimeDigit extends Digit {
  PrimeDigit() {
    count(Digit divisor | (int)this % (int)divisor = 0) = 2
  }
}
```

Observe that the extents of the two classes overlap, yet neither is a subset of the other. This is a natural consequence of defining types by arbitrary characteristic predicates, but it means that not every value has a unique tightest type.

Like Java, QL has an `instanceof` operator, which in QL is really just syntactic sugar for calling the character of a class. For instance, the class of odd digits can be defined like this:

```
class Odd extends Digit { Odd() { not this instanceof Even } }
```

Being intensional predicates, characters can be recursive. For instance, we could define:

```
class Even extends Digit { Even() { this = 0 or (int)this-1 instanceof Odd } }
class Odd  extends Digit { Odd() { (int)this-1 instanceof Even } }
```

However, recursion has to be stratified, so the following is not acceptable, since there is no unique least fixpoint solution to the predicate definitions:

```
  Even() { not this instanceof Odd } // illegal recursion through not
  Odd() { not this instanceof Even } // illegal recursion through not
```

A class may extend multiple supertypes, which simply means that it is a subtype of the their intersection. The (potentially trivial) intersection of all supertypes of a class is called the *domain* of the class. For instance, the class of even prime digits ($\{2\}$) is defined as

```
class EvenPrime extends Even, PrimeDigit {}
```

In fact, since `EvenPrime` imposes no additional constraints on `this` in its character, its extent is exactly equal to its domain. In general, the extent of a class consists of all those values in its domain that satisfy the body of the character; hence, the implicit `this` variable in the character ranges over the domain of the class.

It should be emphasised that the constructor-like syntax for characters is purely superficial: QL has no `new` expression. Like plain Datalog, QL programs can never construct new values or objects, they can only work with primitive values and the values present in the EDB.

2.2 Prescriptive typing

Every variable in QL has a declared type. In most statically typed imperative and functional languages, such declarations are purely compile-time artifacts that describe the set of values

the variable is allowed to take on at runtime; they are checked for consistency by the compiler but play no role at runtime. In contrast to this *descriptive* typing discipline, QL follows a *prescriptive* model, where the syntactic type declaration corresponds to a semantic containment check at runtime.

For instance, consider the predicate `isSmall` that holds for all `Digit`s smaller than five:

```
predicate isSmall(Digit d) { (int)d < 5 }
```

We can use it in a query like the following (which will return the numbers $0, \ldots, 4$):

```
from int i
where isSmall(i)
select i
```

Note that `i` is declared to be of type `int`, but is passed as an argument to `isSmall`, whose parameter is declared to be a `Digit`. Under a descriptive typing discipline, this would be a compile-time type error, but not so in QL: declaring `d` to be a `Digit` simply means that in order for a value to satisfy the predicate `isSmall`, it has to *both* be a `Digit` *and* satisfy the logical conditions imposed by the body of the predicate (namely, being smaller than five).

Another way of looking at it is that type declarations entail an implicit `instanceof` test (which is, in fact, made explicit when translating to plain Datalog), and our definition of `isSmall` is equivalent to

```
predicate isSmall(int d) { d instanceof Digit and d < 5 }
```

The call `isSmall(i)` thus has a perfectly well-defined semantics, regardless of the declared type of `i`, and regardless of what set of values `i` ranges over at runtime. If none of these values happen to be in `Digit`, then `isSmall(i)` will evaluate to an empty set of tuples, as in the following query:

```
from int i
where isSmall(i) and i < 0
select i
```

Of course, a (part of a) query that always evaluates to the empty set is not very useful, and most likely accidental. Our QL compiler tries to detect such empty formulas by running a type inference algorithm similar to the one described by Schäfer et al. [31] on the generated Datalog. In general, emptiness of Datalog (and hence QL) formulas is, unfortunately, undecidable even without arithmetic or string operations, so we can never find all empty formulas. In practice, however, it is quite effective at catching a large number of user errors.

Besides type declarations, `instanceof` tests and casts also restrict the possible values of variables and expressions: as mentioned above, a formula `x instanceof A` restricts `x` to only take on values from the extent of `A`. Similarly, a cast expression `(A)e` evaluates to those values of `e` that are in `A`. Recall that QL expressions are multi-valued, hence casts never fail: if no value of `e` is in the extent of `A`, the cast expression has no values.

2.3 Member predicates

The predicate `isSmall` defined above really describes a property of `Digit`s. In the spirit of object oriented programming, it thus makes sense for it to be part of class `Digit`.

QL allows for *member predicates* to be defined as part of a class. Like characteristic predicates, member predicates have an implicit parameter `this`. Adding `isSmall` as a member predicate to `Digit` we get

```
class Digit extends int {
  Digit() { (int)this in [0..9] }
  predicate isSmall() { (int)this < 5 }
}
```

Member predicates are invoked using a method call-like syntax as in the following query:

```
from Digit d
where d.isSmall()
select d
```

Of course, member predicates can have other parameters besides `this`. For instance, we could add a predicate `divides` to check whether one digit is a divisor of another:

```
class Digit extends int {
  ...
  predicate divides(Digit that) { (int)that % (int)this = 0 }
}
```

There is one important difference between characters and member predicates: in the former, `this` ranges over the domain of the class (that is, the intersection of the extents of its supertypes), while in the latter `this` ranges over the extent of the class itself. This is because the character is what defines the extent of the class in the first place, so by restricting `this` to range over the extent of the class in the character, we would introduce a direct recursive call from the character to itself, which under least fixpoint semantics would mean that the character (and hence the extent of the class) is always empty.

2.4 Multi-valued expressions

Taking the analogy between member predicates and methods further, QL allows treating predicates as multi-valued "functions" with a dedicated `result` parameter. For instance, the member predicate `divides` could instead be written as a multi-valued function returning any of the divisors of a digit:

```
class Digit extends int {
  ...
  Digit getADivisor() { (int)this % (int)result = 0 }
}
```

Note that member predicates using the function syntax have an implicitly declared `result` variable whose type is the declared result type. The results of the predicate are precisely those values that the `result` variable is bound to.

Syntactically, calls to predicates in function syntax are treated like function calls; in particular, they can be chained as in `d.getADivisor().getADivisor()`, which evaluates to all divisors of divisors of `d`.

Semantically, however, such predicates are still relations: there is no requirement that `result` has precisely one value for each value of `this`, or that `result` is "computed from" `this` in some operational sense. In fact, it is quite possible to use `getADivisor()` in reverse to compute all values of `this` yielding a given `result` value, as shown in the following query:

```
from Digit d
where d.getADivisor() = 2
select d  // returns 0, 2, 4, 6, 8
```

When translating to Datalog, predicates using the function syntax are desugared into normal predicates by making the `result` parameter explicit and introducing temporary variables as necessary. For instance, `d.getADivisor()=2` is translated into code of this form:

```
exists (Digit tmp | d.getADivisor(tmp) and tmp = 2)
```

Thus, multi-valued expressions are a purely syntactic, if practically very useful, feature.

2.5 Overriding and virtual dispatch

Given that we have classes that contain member predicates and that may extend each other, it is natural to ask whether there is a notion of overriding and virtual dispatch, and indeed there is: intuitively, at runtime a call `x.p(...)` is dispatched to the definition of `p` belonging to the tightest class containing `x`, i.e., the most specific applicable definition of `p`.

There are two sources of ambiguity: first, `x` may, in general, have multiple values; this is solved by dispatching the call separately for each individual value. Second, classes may overlap, so even for a single value of `x` there can be multiple most specific definitions of `p`; this is solved by dispatching to each definition separately and unioning the results.

More formally, let us represent member predicates by *relation specifiers* of the form $C.p/n$, where C is the name of the class in which the predicate is declared, p is the name of the predicate itself, and n is the predicate's arity, not including the result parameter.[4] We say that a predicate $C.p/n$ *overrides* a predicate $C'.p/n$ if C is a transitive subtype of C'; in this case, we also say that $C.p/n$ is *more specific* than $C'.p/n$.

A member predicate is a *root definition* (or *rootdef* for short) if it does not override any other predicate. The set of rootdefs of a predicate is the set of all rootdefs that it overrides, or the predicate itself if it is already a rootdef. Note that due to multiple inheritance a predicate can have more than one rootdef, but every predicate has at least one.

The *static target* of a member predicate call `x.p(...)`, where the declared type of `x` is a class C, is the most specific predicate $D.p/n$ such that C is a reflexive, transitive subtype of D and n is the number of arguments in the call. In a valid QL program, every predicate call must have a unique static target. The *dispatch candidates* of `x.p(...)` are all the rootdefs of the static target, as well as any predicates that override at least one of the rootdefs.

At runtime, for every value v of `x`, the *applicable targets* of the call are those dispatch candidates $D.p/n$ for which v is in the extent of D, and the *actual targets* are the most specific applicable targets. The call is dispatched to all actual targets for each value of `x`.

In summary, dispatch for a call `x.p(...)` occurs in two stages, one static and one dynamic. At compile-time we compute the set of dispatch candidates, which contains all rootdefs of `p` above the declared type of `x` (that is, member predicates with the same signature that do not themselves override another definition) and all methods that override them. At runtime, each of these candidates applies only if the value of `x` is contained in the corresponding class, and there is no more specific class that also contains `x`.

For example, assume we add a member predicate `kind` to class `Digit` like this:

```
class Digit extends int {
  ...
  string kind() { result = "digit" }
}
```

We override `kind` in the subclasses of `Digit` to result in `"even"` for `Even`, `"odd"` for `Odd` and `"even prime"` for `EvenPrime`. Now consider this query:

[4] QL allows overloading, so there may in fact be multiple member predicates with the same arity as long as they have different parameter types. Like in Java, overloading is resolved entirely statically based on declared types, and hence plays no role in virtual dispatch, and we shall ignore it for simplicity.

```
from Even e select e, e.kind()
```

The static target of the call `e.kind()` is `Even.kind/0`, whose (unique) root definition is `Digit.kind/0`. The dispatch candidates are `Digit.kind/0`, `Even.kind/0`, `Odd.kind/0` and `EvenPrime.kind/0`.

Since `e` is declared to be an `Even`, it ranges over the set $\{0, 2, 4, 6, 8\}$. For the runtime values 0, 4, 6 and 8, the applicable targets of `e.kind()` are `Digit.kind/0` and `Even.kind/0`, and the (unique) actual target is `Even.kind/0`. For the value 2, the applicable targets are `Digit.kind/0`, `Even.kind/0` and `EvenPrime.kind/0`, and the actual target is `EvenPrime.kind/0`.

Hence, the query returns the following set of tuples:

```
{ (0, "even"), (2, "even prime"), (4, "even"), (6, "even"), (8, "even") }
```

Now consider what happens if we add a new class

```
class Two extends Digit {
  Two() { this = 2 }
  string kind() { result = "2" }
}
```

`Two.kind/0` has `Digit.kind/0` as its root definition, so it is now also a dispatch candidate for `e.kind()`, and it is an applicable target for $e = 2$. We now have *two* applicable targets in this case, neither of which is more specific than the other. Hence they will *both* be called, so the query additionally returns the tuple `(2, "two")`.

If, on the other hand, we define `Two` to extend `int` instead of `Digit`, its extent does not change, but `Digit.kind/0` is no longer a root definition of `Two.kind/0`, which hence is no longer a dispatch candidate for `e.kind()`.

We could have defined dispatch in QL in a completely dynamic fashion by considering all predicates with matching name and arity as dispatch candidates, ignoring static type declarations and the class hierarchy. However, as the last example shows this means that we would end up considering predicates from classes that are completely unrelated to the declared type of the receiver in the inheritance hierarchy, which seems undesirable in practice.

Instead, we take a rootdef and all its overriding methods to be alternative implementations of the same operation. To perform virtual dispatch, we identify (at compile time) all operations the static call target implements, and invoke (at runtime) all possible implementations.

2.6 Abstract classes

QL classes as we have described them so far lend themselves quite well to top-down modelling: starting from a general superclass representing a large set of values, we carve out individual subclasses representing more restricted sets of values. In particular, the extent of a class is always defined by filtering its domain through the body of its character.

A classic example where this approach is useful is when modelling ASTs: the node types of an AST form a natural inheritance hierarchy, where, for example, there is a class `Expr` representing all expression nodes, with many different subclasses for different categories of expressions; there might, for instance, be a class `ArithmeticExpr` representing arithmetic expressions, which in turn could have classes `AddExpr` and `SubExpr`.

In other cases, however, we might prefer to instead think of a class as being the union of its subclasses. Here, the superclass exists purely as an interface that provides certain member predicates, with subclasses filling in concrete implementations.

QL supports a notion of abstract classes that allow us to do exactly this: like a concrete class, an abstract class has one or more superclasses and a characteristic predicate. However,

the extent of an abstract class is not the set of values that satisfies its character, but rather the union of the extents of all its subclasses. In particular, an abstract class without subclasses has an empty extent. We will present a practical example of an abstract class in Section 4.

2.7 Miscellanea

QL has various other language features that are important in practice but are either not semantically fundamental, or have direct counterparts in other Datalog dialects. We briefly describe a few of the most important ones.

As we mentioned at the beginning of this section, QL predicates may be recursive, and our program analysis queries make heavy use of this feature. A particularly common kind of recursion is transitive closure, for which QL offers a syntactic shorthand: for a binary predicate `p`, `p+` denotes its transitive closure, and `p*` its reflexive transitive closure. Obviously, this is purely syntactic sugar that is easily translated into plain recursion.

In addition to virtual calls, QL also provides statically dispatched `super` calls. Like their syntactic counterpart in Java, the static target of such a call is looked up in the superclass of the enclosing class, and no virtual dispatch takes place: the actual target is simply the static target. Most often, `super` calls are used in overriding predicate definitions to delegate to the definition that is being overridden.

In order to facilitate reuse and modularity, QL programs can be split over multiple files, and an `import` statement can be used to make definitions from one file available in another. For instance, the QL standard library for JavaScript is split into 40 individual files, which are all imported into a single file `javascript.qll`. Since `import` statements are transitive, QL queries for JavaScript can simply import `javascript.qll` (as shown in Listing 1) to gain access to the entire library. As in Java, implementation hiding is facilitated by access modifiers: member predicates may be marked `private`, meaning that they cannot be called from outside the enclosing class.

QL supports aggregates to perform arithmetic operations such as sum or average on (multi-)sets of values. While very useful in practice, aggregates are really a feature of the Datalog dialect into which QL is compiled, and they do not interact with the language's object oriented features, hence we will not further discuss them.

As a syntactic convenience, casts may be written in a postfix form as `x.(A)`, similarly to type assertions in Go. Semantically, this is entirely equivalent to `(A)x`, but it saves parentheses in chained calls involving casts.

Finally, member predicates of abstract classes may themselves be abstract, meaning that they do not have a body, and the QL compiler checks that each subclass provides an overriding definition of the predicate. Observe that our definition of virtual dispatch guarantees that an abstract member predicate is never the actual target of a call: since the extent of the abstract class is the union of the extents of its subclasses and since each of those subclasses overrides the abstract predicate, there must always be at least one more specific applicable target. Thus, abstract predicates are not semantically fundamental, and in particular have no deep semantic connection with abstract classes.

3 Semantics of Core QL

To formally describe the semantics of QL, we concentrate on a subset dubbed *Core QL* that captures the object-oriented features of QL, while omitting other features that are either purely syntactic or are semantically orthogonal. The semantics of Core QL will be described by a translation to plain Datalog.

$prog$	::=	$\overline{cd}\ \overline{pd}$	program
cd	::=	$\textbf{abstract}^?\ \textbf{class}\ C\ \textbf{extends}\ \overline{T}\ \{C()\ \{f\}\ \overline{pd}\}$	class definition
pd	::=	$\textbf{predicate}\ p(\overline{T\ x})\ \{f\}$	predicate definition
f, g	::=	$p(\overline{x})$	formula
	|	$x.p(\overline{y})$	
	|	$C.\textbf{super}.p(\overline{x})$	
	|	$\textbf{not}\ f$	
	|	$f\ \textbf{and}\ g$	
	|	$f\ \textbf{or}\ g$	
	|	$\textbf{exists}(T\ x \mid f)$	
S, T	::=	C	type reference
	|	$@b$	
	|	$C.\texttt{domain}$	

Figure 1 Syntax of Core QL; $\overline{\cdot}$ denotes (possibly empty) sequences, $\cdot^?$ optional elements

3.1 Core QL

Figure 1 presents the syntax of Core QL. Like full QL, Core QL programs consist of toplevel predicates and (concrete and abstract) classes with a characteristic predicate and member predicates. We do not model QL's `from-where-select` query syntax, but simply consider queries as special toplevel predicates.

Predicates can declare parameters, and their bodies are first-order formulas with predicate calls as atomic formulas. As in full QL, there are calls to toplevel predicates and to member predicates, and the latter may be either virtual calls or **super** calls. Unlike full QL, **super** calls always have to be explicitly annotated with the class they refer to.

Type references appearing in **extends** clauses, parameter declarations or existential quantifiers are either class names C or base type names $@b$. We assume the latter to be defined by an underlying database schema. Core QL also has a syntax for *domain types* of the form $C.\texttt{domain}$ for a class name C; these cannot appear at the source level but play a crucial role in the semantics of classes.

Among the QL features omitted from Core QL are overloading, the function syntax for predicates, expressions, the **forall** quantifier, casts and **instanceof**: these can all be desugared into Core QL features. Other QL features such as primitive types and aggregates have no counterpart in Core QL, but their semantics is largely orthogonal to the object-oriented features of the language, which are the focus of our presentation.

3.2 Datalog

The target language for our translation is an untyped variant of Datalog. A Datalog program consists of a series of intensional predicate definitions of the form $p(\overline{x}) \leftarrow \varphi.$, where p is a predicate name, $\overline{x}$ is a possibly empty sequence of variable names, and φ is a formula of first-order logic with the usual logical connectives. The free variables of φ must be exactly $\overline{x}$. The atoms of φ are calls of the form $r(\overline{y})$, where r is either the name of an intensional predicate defined in the same program, or the name of an extensional predicate.

We say that an intensional predicate p calls a predicate q, written $p \to q$, if the body of p contains a call to q. As usual, $\to^*$ denotes the reflexive transitive closure of this relation.

$p \stackrel{-}{\rightarrow} q$ means that one of the calls to q in p occurs under an odd number of negations.

We require all Datalog programs to be *stratified*, that is, recursive call chains of the form $p \rightarrow^* q \stackrel{-}{\rightarrow} r \rightarrow^* p$ are not allowed. Any stratified Datalog program has a least fixpoint semantics, that is, given an interpretation of the extensional predicates each intensional predicate has a unique minimal interpretation that satisfies the predicate's definition.

3.3 Valid Core QL

In order to be meaningfully translatable to Datalog, a Core QL program has to fulfil a set of syntactic requirements and pass some static semantic checks. There is one additional check that is easiest to perform on the generated Datalog and will be discussed later.

The syntactic requirements are entirely standard and mostly naming related:

▶ **Definition 1** (Syntactic validity)**.** In order for a Core QL program to be *syntactically valid*, the following conditions have to be satisfied:

- No two classes and no two toplevel predicates with the same arity may have the same name; no two member predicates of the same class with the same arity, and no two parameters of the same predicate may have the same name.
- Every **extends** clause must list at least one type.
- Every characteristic predicate must have the same name as its enclosing class.
- No predicate parameter may have the name **this**.
- For every variable name appearing in a formula, there must either be an enclosing **exists** declaring a variable of that name, or the enclosing predicate must have a parameter of that name, or the variable name is **this** and it appears in a member predicate or character. In particular, every variable name can be associated with a declared type.
- Similarly, for every class name appearing in a type reference there must be a class of the same name, and for every predicate name appearing in a call to a toplevel predicate, there must be a toplevel predicate of that name with the appropriate arity.
- **super** calls may only appear in member predicates.

To formulate the static semantic checks, we first introduce some terminology.

▶ **Definition 2** (Relation specifiers)**.** A *relation specifier* $C.p/n$ consists of a class name C and a pair p/n, where p is a predicate name and n a natural number.

Unless otherwise specified, we require relation specifiers to be *valid*, that is, C must be the name of a class defined in the program, and C must declare a member predicate p with n parameters. We abbreviate $C.p/n$ as $C.p$ where n is not important or obvious from context.

▶ **Definition 3** (Subtyping)**.** The *subtyping* relation $S <: T$ is the smallest relation such that for every class C

- $C <: C.\texttt{domain}$,
- if C extends T, then $C.\texttt{domain} <: T$.

As usual, $S <:^+ T$ denotes the transitive closure of this relation.

▶ **Definition 4** (Overriding)**.** $C.p/n$ overrides $D.p/n$, written $C.p/n \prec D.p/n$, if $C <:^+ D$. We write $C.p/n \preceq D.p/n$ to mean that either $C = D$ or $C.p/n \prec D.p/n$. If $D.p/n$ overrides no other member relation, it is a *rootdef*. We write $\rho(C.p/n)$ for the set of all rootdefs $D.p/n$ such that $C.p/n \preceq D.p/n$.

▶ **Definition 5** (Member predicate lookup)**.** We define a lookup function $\lambda(S, p, n)$ that looks up a member predicate in a type given a name and its arity and returns a set of candidates:

$$\lambda(S,p,n) = \begin{cases} \{C.p/n\} & \text{if } S = C \text{ and } C.p/n \text{ is valid} \\ \bigcup_{S<:T} \lambda(T,p,n) & \text{otherwise} \end{cases}$$

The static semantic checks guarantee that a program can be translated to Datalog:

▶ **Definition 6** (Translatability)**.** A syntactically valid Core QL program is *translatable* if the following conditions are met:

- It is not the case that $T <:^+ T$ for some type T; that is, the subtyping relation is acyclic.
- For every member predicate call $x.p(\overline{y})$ where x has type T there is exactly one static call target, i.e., $|\lambda(T, p, |\overline{y}|)| = 1$. In practice, this means that classes must override ambiguously inherited predicates.
- Similarly, for every call $D.\mathbf{super}.p(\overline{x})$ in a member predicate of a class C, we must have $C <:^+ D$ and $|\lambda(D, p, |\overline{x}|)| = 1$.

3.4 Translation to Datalog

The translation from (translatable) Core QL to Datalog is presented in Figure 2 as a family of structurally recursive translation functions:

- $\mathcal{T}_c$ translates Core QL class definitions into sequences of Datalog predicates, using an auxiliary function $\mathcal{K}$ to generate the extent predicate as explained below;
- $\mathcal{T}_m$ translates Core QL member predicates into Datalog predicates; it takes the declaring class of the member predicate as an additional argument;
- $\mathcal{T}_p$ translates toplevel Core QL predicates into Datalog predicates;
- $\mathcal{T}_b$ translates Core QL predicate and character bodies into Datalog formulas; it takes a type environment as an additional argument;
- $\mathcal{T}_f$ translates Core QL formulas into Datalog formulas; it takes a type environment as an additional argument.

The type environments Γ used by $\mathcal{T}_b$ and $\mathcal{T}_f$ are partial functions from variable names to type references. We write $\langle x := T \rangle$ to denote the type environment that maps x to T, and contains no other mappings. As usual, $\Gamma[x := T]$ is a type environment that is identical to Γ except that it maps x to T.

We use $\bigvee_{i \in I} \varphi_i$ and $\bigwedge_{i \in I} \varphi_i$ to denote disjunctions and conjunctions of families of formulas indexed by a set I. For empty index sets we define $\bigvee_{i \in \emptyset} \varphi_i := \bot$ and $\bigwedge_{i \in \emptyset} \varphi_i := \top$, i.e., empty disjunctions are false and empty conjunctions are true.

We now discuss the individual translation functions in greater detail.

Classes

For every Core QL class C, we generate a definition for its domain predicate C.`domain`, its characteristic predicate $C.C$, and its extent predicate C. Additionally, each member predicate pd_i is translated recursively using $\mathcal{T}_m$.

The domain predicate is defined as the intersection of the characteristic predicates of all supertypes of C. The characteristic predicate is generated from its Core QL definition by $\mathcal{T}_b$, which delegates to $\mathcal{T}_f$, but additionally enforces prescriptive typing, which is not a feature of plain Datalog.

Classes $(cd \equiv \textbf{abstract}^? \textbf{ class } C \textbf{ extends } \overline{T}\ \{C()\ \{f\}\ \overline{pd}\})$

$$\mathcal{T}_c(cd) := \begin{array}{lll} C.\texttt{domain}(\textbf{this}) & \leftarrow & \bigwedge_{C<:B} B.B(\textbf{this}) \wedge \bigwedge_{C<:@b} @b(\textbf{this}). \\ C.C(\textbf{this}) & \leftarrow & \mathcal{T}_b(f, \langle \textbf{this} := C.\texttt{domain} \rangle). \\ C(\textbf{this}) & \leftarrow & \mathcal{K}(cd). \\ \overline{\mathcal{T}_m(pd_i, C)} & & \end{array}$$

$$\begin{array}{lll} \mathcal{K}(cd) & := & \bigvee_{D<:C} D(\textbf{this}) \quad \text{if } cd \text{ is abstract} \\ \mathcal{K}(cd) & := & C.C(\textbf{this}) \quad \text{if } cd \text{ is concrete} \end{array}$$

Toplevel predicates $(pd \equiv \textbf{predicate}\ p(\overline{T\ x})\ \{f\})$

$$\mathcal{T}_p(pd) := p(\overline{x}) \leftarrow \mathcal{T}_b(f, \langle \overline{x_i := T_i} \rangle).$$

Member predicates $(pd \equiv \textbf{predicate}\ p(\overline{T\ x})\ \{f\})$

$$\mathcal{T}_m(pd, C) := \begin{array}{lll} C.p(\textbf{this}, \overline{x}) & \leftarrow & \mathcal{T}_b(f, \langle \textbf{this} := C, \overline{x_i := T_i} \rangle). \\ C.p^{\text{disp}}(\textbf{this}, \overline{x}) & \leftarrow & (\bigwedge_{D.p \prec C.p} \neg D(\textbf{this})) \wedge C.p(\textbf{this}, \overline{x}). \end{array}$$

Predicate/character bodies

$$\mathcal{T}_b(f, \Gamma) := (\bigwedge_{(x,S)\in\Gamma} S(x)) \wedge \mathcal{T}_f(f, \Gamma)$$

Predicate calls

$$\begin{array}{lll} \mathcal{T}_f(p(\overline{x}), \Gamma) & := & p(\overline{x}) \\ \mathcal{T}_f(x.p(\overline{y}), \Gamma) & := & \bigvee_{R.p\in\rho(D.p)} (\bigvee_{B.p \preceq^* R.p} B.p^{\text{disp}}(x, \overline{y})) \quad \text{where } D.p := \lambda(\Gamma(x), p, |\overline{y}|) \\ \mathcal{T}_f(C.\textbf{super}.p(\overline{x}), \Gamma) & := & D.p(\textbf{this}, \overline{x}) \quad \text{where } D.p := \lambda(C, p, |\overline{x}|) \end{array}$$

Other formulas

$$\begin{array}{lll} \mathcal{T}_f(\textbf{not } f, \Gamma) & := & \neg \mathcal{T}_f(f, \Gamma) \\ \mathcal{T}_f(f \textbf{ and } g, \Gamma) & := & \mathcal{T}_f(f, \Gamma) \wedge \mathcal{T}_f(g, \Gamma) \\ \mathcal{T}_f(f \textbf{ or } g, \Gamma) & := & \mathcal{T}_f(f, \Gamma) \vee \mathcal{T}_f(g, \Gamma) \\ \mathcal{T}_f(\textbf{exists}(C\ x \mid f), \Gamma) & := & \exists x.\, (C(x) \wedge \mathcal{T}_f(f, \Gamma[x := C])) \end{array}$$

Figure 2 Translation from Core QL to Datalog

The extent predicate, finally, is the Datalog predicate that actually defines the extent of the class. For concrete classes, this is the same as the characteristic predicate. For abstract classes, however, their extent is instead defined as the union of the extents of their subclasses.

The distinction between these three predicates is subtle, but crucial. C.`domain` is mainly needed to circumscribe the type of **this** inside the characteristic predicate of C. To see why it cannot have type C, consider what the definitions of the characteristic predicate and the extent predicate would look like if it did:

$$\begin{array}{lcl} C.C(\textbf{this}) & \leftarrow & C(\textbf{this}) \wedge \ldots. \\ C(\textbf{this}) & \leftarrow & C.C(\textbf{this}). \end{array}$$

Note the recursion between $C.C$ and C, which is resolved by computing least fixpoints. Clearly, both rules are satisfied if $C.C$ and C are empty, and this is also the least fixpoint. In other words, typing **this** as C in the character would render every concrete class empty. Using type C.`domain` instead breaks the recursion, and both predicates are now interpreted as the subset of the extent of C.`domain` that satisfies the body of the character, as expected.

The distinction between the characteristic predicate $C.C$ and the extent predicate C is only relevant for abstract classes (and the two are indeed equal for concrete classes): the extent of an abstract class is not the extent of its characteristic predicate, but rather the union of the extents of its subclasses, and this is precisely how C is defined.

Predicates

Toplevel Core QL predicates are translated directly into Datalog predicates of the same name, using $\mathcal{T}_b$ to translate the body and enforce prescriptive typing for all parameters.

Member predicates $C.p$ are translated into two Datalog predicates: an implementation predicate of the same name, and a dispatch predicate $C.p^{\text{disp}}$. The latter is used during dispatch translation as explained below.

Formulas

Most Core QL formulas are straightforward to translate into their Datalog counterparts, except that for quantifiers we again enforce prescriptive typing. The two most interesting cases are **super** calls and virtual calls. For the former, we simply use the λ function to look up the member predicate to invoke, explicitly passing in **this** as the first argument.

For virtual calls, we need to implement dispatch. Recall that for a call with static target $D.p$, the dispatch candidates are all member predicates that override a rootdef for $D.p$. Hence the call is translated into two nested disjunctions: the outer disjunction is over all rootdefs $R.p$ of the static target $D.p$, while the inner is over all methods overriding the rootdef. For each dispatch candidate $B.p$ identified in this way, we emit a call to $B.p^{disp}$, which in turn invokes $B.p$, but only if it is a most specific implementation of p for the given parameter.

The syntactic and static semantic checks ensure that the result of the translation is a valid Datalog program, except that they do not yet ensure stratification. While it would be possible to devise a QL-level check for this, it is conceptually simpler to check stratification of the generated Datalog, and map any violations of this condition back to the QL code it originated from. This is also how we implement the stratification check in our QL compiler.

3.5 Example

As a concrete example of the translation, we show the Datalog definitions generated for classes `Digit` and `Even` from Section 2, including definitions for the method `kind` and a

query predicate that computes all even digits `e` and their kinds `k`.[5]

```
Digit.domain(this) ← int(this).
Digit.Digit(this) ← Digit.domain(this) ∧ range(this, 0, 9).
Digit(this)       ← Digit.Digit(this).
Digit.kind(this, result) ← Digit(this) ∧ string(result) ∧ result = "digit".
Digit.kind^disp(this, result) ← ¬Even(this) ∧ Digit.kind(this, result).

Even.domain(this) ← Digit.Digit(this).
Even.Even(this) ← Even.domain(this) ∧ mod(this, 2, 0).
Even(this) ← Even.Even(this).
Even.kind(this, result) ← Even(this) ∧ string(result) ∧ result = "even".
Even.kind^disp(this, result) ← Even.kind(this, result).

query(e, k) ← Even(e) ∧ string(k) ∧ (Digit.kind^disp(e, k) ∨ Even.kind^disp(e, k)).
```

4 QL in Practice

The previous two sections have presented the semantics of QL in some detail, using toy examples for simplicity. We now show how these concepts apply in a more realistic setting.

4.1 Databases and schemata

Recall that QL programs (or rather, the Datalog programs into which they are translated) are run on a relational database. In practice, we use our own custom database system, but in principle QL programs could just as well be run on an off-the-shelf system.[6]

As in any relational database, data is represented in terms of tuples (rows), grouped into relations (tables) such that all tuples in a relation have the same arity (number of columns). A table may have a distinguished *primary key* column, meaning that no two rows in the table have the same value in this column; in other words, each value in the primary key column uniquely identifies a row. A table t_1 may also have *foreign key* columns referring to the primary key column of a table t_2 ($t_1 = t_2$ is allowed), meaning that any value occurring in the foreign key column must also be present in the corresponding primary key column; in other words, each row of t_1 references a unique row of t_2.

Keys can be used to model hierarchical data structures using flat tables. Suppose, for instance, that we want to build a snapshot database representing a JavaScript program. Among other data, we may want to represent the abstract syntax tree of source files, including, in particular, all expressions. Simplifying somewhat, we could introduce a table `exprs` with four columns: a primary key column `id` with a unique ID for each expression; a column `kind` indicating what kind of expression we are dealing with, encoded as an integer; a foreign key column `parent` referencing the ID of the parent expression in the AST; and an integer column `idx` recording the ordering among children of the same parent.[7] Since `id` is a primary

[5] Core QL does not include primitive types or arithmetic operations, so for the purposes of this example we have treated `int` and `string` like entity types, and assumed EDB relations `range(a,b,c)` and `mod(x,y,z)` corresponding to QL's range operator `a in [b..c]` and the modulo operator `x%y=z`, respectively.

[6] In fact, early versions of our compiler translated QL to SQL, using a third-party database system as our backend. We found, however, that the generated queries performed quite badly, mostly due to our liberal use of recursion, which is not well supported on most SQL systems.

[7] Alternatively, each expression could keep references to their child expressions, but as different kinds of expressions have different numbers of children, this would require tuples with different arities, which would have to be stored in different tables.

key, every expression is guaranteed to have a unique ID, and since `parent` is a foreign key, it is a well-defined reference to another expression in the same table.

For example, assume we want to represent a comparison expression `x === 1`; its two children are the variable reference `x` and the integer literal `1`. Assume further that we assign them the IDs 0, 1 and 2, and encode "equality expression" as kind 2, "variable reference" as kind 1, and "integer literal" as kind 0. The variable reference `x` thus has `id` 1, `kind` 1, `parent` 0, and `idx` 0, corresponding to the tuple `exprs(1, 1, 0, 0)`, while "`1`" gives rise to `exprs(2, 0, 0, 1)`. In practice, we would additionally store the name of the referenced variable and the value of the integer literal in separate tables, which we elide for simplicity.

At the storage level, all four columns of the `exprs` table look the same: they are just integers. QL, on the other hand, espouses a strongly typed view where keys are treated as opaque values and annotated with an entity type. Primary key columns define an entity type whose extent is the set of values occurring in that column. For instance, the `id` column of `exprs` could be annotated with the entity type `@expr`, meaning that it defines the extent of entity type `@expr`. Foreign key columns are also annotated with entity types, and the database system ensures that they only contain values drawn from the extent of their type.

This information about tables, the types of their columns, and the entity types they define is described by a *schema*, which thus defines the interface between a QL program and the database on which it is run.

4.2 Data abstraction

Given a snapshot database with a schema as described above, we could now write our analysis queries in plain Datalog, directly accessing the information stored in the tables. However, this can become quite cumbersome since we need to remember which column contains which piece of information. If, at some point, we want to change the database schema, a large-scale refactoring may be necessary to ensure that the analysis works with the updated schema.

QL classes provide a convenient way of abstracting away from the specifics of how data is stored in tables and providing a higher-level interface, thereby acting like abstract datatypes. For instance, we could implement a QL class `Expr` to provide an abstract view of the `exprs` table discussed above:

```
class Expr extends @expr {
  Expr getParent() { exprs(this, _, result, _) }

  Expr getChildExpr(int i) { exprs(result, _, this, i) }

  string toString() { result = "expr" }
}
```

Since `Expr` should contain *all* expressions represented in the database, it has a trivial character, and hence the same extent as `@expr`. The member predicate `getParent` provides access to the `parent` column of the `exprs` table, while `getChildExpr` enables navigation in the other direction. Note that we do not need to check that the index `i` is in range: if there is no i-th child, the predicate will simply fail to hold. QL also requires each class to define (or inherit) a `toString` member predicate, for which we provide a dummy implementation.

This interface allows us to navigate the program AST as a graph without being exposed to the details of its relational representation. For instance, `e.getParent+() = f` expresses the property that expression `e` is nested within expression `f` (using QL's "+" syntax for transitive closure).

If all client analyses use `Expr` instead of directly accessing the EDB, we can easily change our data representation later on. For instance, it may not be desirable to record the parent

expression directly in the `exprs` table, since top-level expressions do not have a parent expression. Instead, the parent-child relation could be stored in a separate three-column table `expr_nesting(child,parent,idx)`. The first two columns are foreign keys, so they must refer to properly defined `@expr` values, but there is no requirement that every `@expr` value appears in the first column (or, for that matter, the second column), so expressions without parents can now be modelled.

If we want to switch to this representation, we can simply update the definitions of `getParent` and `getChildExpr` without affecting any client analyses:

```
class Expr extends @expr {
  Expr getParent() { expr_nesting(this, result, _) }
  Expr getChildExpr(int i) { expr_nesting(result, this, i) }
  ...
```

4.3 Inheritance

Class `Expr` abstracts away from the details of the relational encoding of the AST and is useful for implementing generic syntax tree traversal, but if we want a richer semantic interface we have to implement subclasses of `Expr`. For instance, we could implement a class `EqExpr` to exclusively represent equality checks (and no other expressions):

```
class EqExpr extends Expr {
  EqExpr() { exprs(this, 2, _, _) }
  Expr getLeftOperand() { result = this.getChildExpr(0) }
  Expr getRightOperand() { result = this.getChildExpr(1) }
  string toString() { result = "===" }
}
```

The characteristic predicate filters out those expressions that do not have kind 2 (which, in our example encoding, represents equality). We provide getter predicates for the two operands of the equality, further abstracting away from the details of our AST representation, and we override the `toString` predicate to provide a more specialised string representation. Similar classes can be implemented for variable references, literals, and other expressions.

QL classes thus allow us to impose an abstract data type representation on relational data. Since classes can freely overlap, we can even implement multiple representations for the same data. For instance, we could overlay a control flow graph structure on top of the AST by defining a class `CFGNode` that also extends `@expr`, but presents it under a different interface, offering, say, a method `getASuccessor()` to compute call graph successors.

4.4 Overriding

As a practical example of overriding, consider implementing the member predicate `Expr.isPure` used in Listing 1. Its default implementation in class `Expr` is `none()`, which is a built-in predicate that always fails. In other words, we conservatively assume that all expressions are impure, and override it in subclasses:

```
class Expr extends @expr {
  predicate isPure() { none() }
  ...
```

In class `Literal`, we override `isPure` as `any()`, a built-in predicate that always succeeds:

```
class Literal extends Expr {
  predicate isPure() { any() }
  ...
```

As another example, equality checks are pure if all of their children are:

```
class EqExpr extends Expr {
  predicate isPure() { forall (Expr c | c = this.getChildExpr(_) | c.isPure()) }
  ...
```

4.5 Interface vs Implementation

Abstract classes support decoupling interface and implementation even further: while `Expr` implements an interface in terms of one particular set of EDB relations, abstract classes specify only an interface, which may be implemented in multiple different ways.

As an example, assume we want to implement an analysis for JavaScript to find comparisons between expressions with incompatible (dynamic) types, which will always evaluate to `false` at runtime. Assume further that we have implemented a binary predicate `incompatTypes(e, f)` that infers possible types of `e` and `f` and checks whether they are compatible. Using class `EqExpr` defined above, we could implement our analysis as follows:

```
from EqExpr eq, Expr l, Expr r
where l = eq.getLeftOperand() and r = eq.getRightOperand() and incompatTypes(l, r)
select eq, "Operands have incompatible types."
```

Other JavaScript language constructs that compare values in the same way include, e.g., the `switch` statement. If we want to consider them in our query, we could add another disjunct to the `where` part, but this would make it less readable, and we would need to keep extending it for any other equality tests we want to support. Instead, we introduce an abstract class capturing the common interface for all equality tests:

```
abstract class EqualityTest extends ASTNode {
  abstract Expr getALeftOperand();
  abstract Expr getARightOperand();
}
```

Like all classes, `EqualityTest` needs a superclass: we choose `ASTNode`, which is a common superclass of `Expr` and `Stmt` defined in the JavaScript QL libraries. The interface defined by `EqualityTest` consists of member predicates to access the left and right operands of the comparison. We allow a single equality test to have multiple left or right operands; e.g., in a `switch`, every `case` is viewed as a right operand of the comparison.

We can implement this interface on `EqExpr` and `SwitchStmt` by introducing new classes that have the same extent as `EqExpr` and `SwitchStmt`, respectively, but extend `EqualityTest`:

```
class EqExprEqualityTest extends EqExpr, EqualityTest {
  Expr getALeftOperand() { result = this.getLeftOperand() }
  Expr getARightOperand() { result = this.getRightOperand() }
}

class SwitchEqualityTest extends SwitchStmt, EqualityTest {
  Expr getALeftOperand() { result = this.getExpr() }
  Expr getARightOperand() { result = this.getACase().getExpr() }
}
```

The extent of `EqualityTest` now contains all equality expressions and all switch statements, under a convenient interface for our query:

```
from EqualityTest eq, Expr l, Expr r
where l = eq.getALeftOperand() and r = eq.getARightOperand() and incompatTypes(l, r)
select eq, "Operands have incompatible types."
```

To add support for other kinds of equality tests, all we need to do is to define new subclasses of `EqualityTest`; the query need no longer be changed.

4.6 Optimisation

In practice, the translation shown in Figure 2 can produce very inefficient Datalog, particularly when translating virtual calls: the disjunction over all candidates can be quite large, and in many cases the context restricts the receiver variable in such a way that some disjuncts end up always being false, which would lead to a lot of wasted computation if evaluated naively. For example, in the translation shown in Section 3.5, the `query` predicate restricts `e` to `Even`, so the dispatch disjunct $\texttt{Digit.kind}^{\text{disp}}(\texttt{e}, \texttt{k})$ can never apply. Another source of inefficiency are superfluous type guards for variables that are already restricted sufficiently by their uses. For instance, the conjunct `string(result)` in `Even.kind` is implied by `result="even"` and hence unnecessary.

One could devise a more sophisticated compilation scheme that does not generate useless dispatch disjuncts or type tests, but we choose to instead perform these optimisations at the Datalog level: eliminating infeasible dispatch disjuncts is a special case of the more general problem of detecting formulas that logically contradict other formulas in their context and hence are empty in context; similarly, unnecessary type tests are a special case of formulas that are logically implied by other formulas in their context and hence are redundant. We use a type inference-based approach to identifying such empty or redundant formulas based on the work of de Moor et al. [14] and Schäfer et al. [31]. We also apply various standard optimisations such as inlining, join ordering and the magic sets transformation [6]; the latter two rely on (compile-time) estimation of (run-time) relation sizes, for which we use an approach similar to the one described by Sereni et al. [32].

Ultimately, the example from Section 3.5 is simplified by our optimiser to

```
Even(this) ← range(this, 0, 9) ∧ mod(this, 2, 0).
query(e, k) ← Even(e) ∧ k = "even".
```

5 Case Study

To demonstrate the benefits of QL in implementing static checks, we performed an informal case study, reimplementing the Error Prone [19] checks in QL. Error Prone is a static checker for Java that integrates with the compiler and checks Java source code for common mistakes, reporting them as compiler errors and suggesting possible fixes. As of version 2.0.4, there are 101 checks, which we reimplemented in QL, ensuring that they pass all the unit tests of the original. This required about one man-month of effort by an experienced QL programmer.

The original Java implementation of the checks and the fix suggestions comprises about 10,500 lines of code, not including supporting libraries such as the `javac` Compiler Tree API.[8] Our reimplementation, by contrast, is slightly less than 2,000 lines of code, not including the QL standard library for Java. However, our implementation only covers the checks themselves, not the suggested fixes. Manual inspection suggests that the latter account for about 1,100 lines of code in the Java implementation, leaving 9,400 lines of analysis code.

Java is famously verbose, which explains part of the size difference, although QL is overall syntactically quite similar to Java. If we exclude Java `package` declarations and `@Override` annotations (which have no QL counterparts) and `import` statements (the number of which is largely determined by the organisation of the supporting libraries), we can subtract a further 2,800 lines from Error Prone and 100 lines from our implementation. In other words, the Java implementation is 3.5x the size of the QL implementation.

[8] That is, counting only files in the `src/main/java/com/google/errorprone/bugpatterns` directory.

Listing 2 QL code for detecting nested null checks in Java

```
// a "==" test where one operand is a null literal
class NullCheck extends EQExpr { NullCheck() { getAnOperand() instanceof NullLiteral } }

// "inner" is nested inside the then-branch of "outer", and both check nullness of "v"
predicate nestedNullCheck(IfStmt outer, IfStmt inner, Variable v) {
  inner.getParent+() = outer.getThen() and
  outer.getCondition().(NullCheck).getAnOperand() = v.getAnAccess() and
  inner.getCondition().(NullCheck).getAnOperand() = v.getAnAccess()
}
```

This is mostly because AST traversal and filtering, which require lots of boilerplate code in Java, can be expressed very concisely using recursion and prescriptive typing in QL. For example, Listing 2 shows part of a query for finding incorrect uses of double-checked locking: `NullCheck` picks out comparisons to `null`, and `nestedNullCheck` identifies nested `if` statements that check the same variable for nullness. Recursion is used to check the nesting condition. The casts to `NullCheck` would fail in a descriptive interpretation, since `s.getCondition()` is not a `NullCheck` for most `if` statements `s`. In QL, they act as filters, restricting `outer` and `inner` to those `if` statements that do, in fact, check nullness.

In spite of their conciseness, the QL queries scale well: on a 1.5 MLoC Java code base, 64 out of 101 queries finish in one second, a further 19 in five seconds or less. Only five queries take longer than ten seconds, finishing in 12, 12, 13, 15 and 31 seconds, respectively.[9]

In summary, this case study shows that QL allows us to quickly implement static analysis checks as concise and scalable queries. While most of the Error Prone checks are quite syntactic in nature, QL has also been used to implement deep, yet scalable semantic analyses, such as a points-to analysis for Python and flow-based security analyses for Java and C++.

6 Discussion and Related Work

In this section, we will discuss QL's object-oriented features in the light of popular definitions of object orientation in the literature, and then proceed to survey related work.

6.1 Discussion

Looking at Section 2, one may question whether QL is really object-oriented at all. It may seem like we have forcibly applied familiar syntax for methods and classes to completely different concepts. However, these examples were deliberately designed as a minimalist illustration of QL semantics; in practice, QL programs rely on an underlying database as described in Section 4. Here, the parallels with traditional object-oriented programming become quite striking: tuples in a table are records; primary keys uniquely identify tuples, and hence play the role of addresses; foreign keys uniquely reference other tuples, thus acting like references to other records. Hence, a database can be viewed as a strongly typed heap containing a collection of objects with fields that may contain primitive values or references to other objects, where objects with the same layout are collected into tables.

A QL class like `Expr` whose extent is (a subset of) the primary key column of a table thus describes a set of records; its member predicates can access record elements, using

[9] Timings obtained on an Intel Core i7-4900MQ laptop, with 1GB of heap allocated to the analysis.

the primary key for lookup. Subclasses inherit member definitions and can override them, allowing different implementations of the same operation for different objects, the appropriate implementation being chosen at runtime based on the dynamic type of the object.

Of course, QL classes are not restricted to this particular setting, since classes can be sets of arbitrary values, not just primary keys, and EDB tables can be arbitrary relations, for which there is no parallel in other object-oriented languages.

We now discuss QL in the light of several well-known definitions of object orientation.

Wegner [36] proposes the definition "object-oriented = objects + classes + inheritance". In his account, an object is characterised by a set of operations and a state, where the value returned by an operation on an object may depend on the object's state as well as its arguments. In QL, any value can assume the role of an object when viewed as member of a class: its operations are the member predicates defined by the class, and its state is the value itself and other values associated with it in the database, e.g., as part of the same tuple. In QL, this state is immutable. Wegner further defines a class as a template from which objects may be created; this does not match QL's worldview, in which no new values can be constructed. QL classes do, however, fulfil all other characteristics Wegner ascribes to classes: objects of the same class have common operations specified in one or more interfaces, and the class body specifies code for implementing operations in the interface. Inheritance in QL, finally, also closely matches Wegner's definition, which simply stipulates that operations are inherited by subclasses.

Similarly, a folklore definition of object orientation considers its two crucial features to be data abstraction and inheritance. We have shown in Section 4 how QL classes achieve the former by abstracting from the concrete layout of the EDB tables. Member predicate inheritance in QL is entirely conventional, and, as usual, can be used both for implementing an interface and for code reuse. Hence QL fits this definition.

A more recent definition is due to Cook [10, 9], who defined on object-oriented language to be a language that supports the dynamic creation and use of objects. An object, in turn, is defined as "a first-class, dynamically dispatched behaviour", where a behaviour is "a collection of named operations", and dynamic dispatch means that "different objects can implement the same operations in different ways". QL clearly satisfies the second half of this definition: classes associate named operations (member predicates) with arbitrary values, both entity values and primitive values like numbers and strings, and subtyping and virtual dispatch allow implementing the same operations differently for different values.

The first half of Cook's definition is *not* satisfied: QL programs cannot create new values, and can only work with the fixed contents of the EDB and primitive values. This is, in our view, a natural limitation for a query language; QL's concepts of classes and dispatch would continue to make sense in a setting where tables can be extended with new tuples.

QL also does not support mutable state (that is, updating existing EDB tuples), and has no notion of object identity as opposed to value identity. These are sometimes considered essential properties of objects, though Cook argues against this viewpoint.

Ullman [35] argued that query languages cannot be "seriously logical and seriously object-oriented at the same time". His argument is framed in the context of a radical reinterpretation of the relational model in purely object-oriented terms, where tuples and relations are understood as objects with relational operators as methods. In combination with a strong notion of object identity, this would mean that each new tuple created during evaluation would be distinct from any previous tuple, which is incompatible with the traditional view of relations as sets of tuples and hence least fixpoint semantics. Furthermore, a very powerful type system would be needed to capture the type of generic relational operators such union

or join in such a setting. Our approach to object orientation differs significantly from the model discussed by Ullman: QL classes range over values, not tuples, sidestepping his first point. Some values may, of course, happen to be EDB keys uniquely identifying tuples, but QL's semantics is entirely oblivious to this fact. Since tuples are not objects, relational operators are built-in language constructs, not methods, and hence not typed. In Ullman's view, this approach would disbar QL from being "seriously" object-oriented, but we have argued that it nevertheless provides the usual benefits of object orientation.

Several papers have addressed the problem of integrating object identity into logic programming, with proposals ranging from giving the programmer full control over managing object identities [38] to sophisticated extensions of the underlying logic [23]. Our experience with QL suggests that object-oriented programming is quite possible without relying on object identity, so we consider this an orthogonal issue.

In summary, QL supports data abstraction and inheritance with dynamic dispatch, widely considered to be essential features of object-oriented languages. Going beyond traditional object orientation, QL supports overlapping classes, and member predicates that are true relations, both of which are natural in the context of logic programming. On the other hand, object creation and mutation are less natural for a query language, and hence not supported.

6.2 Related Work

Encoding hierarchical data in relational form is conceptually quite straightforward, but querying the encoded data in a traditional relational language is cumbersome. This has been termed the *object-relational impedance mismatch*, and led to calls for replacing the relational model with models directly supporting structured data [5]. We will not discuss the literature on this topic, since QL is based on a completely conventional relational data model. Our approach agrees in spirit, if not in detail, with the so-called Third Manifesto [12], which argues that object orientation should be built on, rather than supplant, the relational model.

Object-oriented extensions of Datalog have been investigated in the literature before. Abiteboul et al. [2] consider a language where individual rules for predicates may be associated with classes. They consider three variants of overriding, one of which, termed *static inheritance*, is somewhat similar to QL's approach, although they define overriding at the level of individual rules, not entire predicates. Since their language has no static types they have no concept of rootdefs, instead considering all methods with the right name and arity as dispatch candidates; as we have argued, this makes dispatch highly non-local and brittle in the presence of overlapping classes. They also do not consider multiple inheritance. While for the most part they assume that classes are defined directly by the EDB, they also briefly consider "virtual" classes (first proposed in [1]), which, like concrete QL classes, are defined by a characteristic predicate. It is unclear if their proposal supports recursive characters.

Extensions of Prolog with subtyping and inheritance have also been proposed [3, 34]. These approaches focus on types for structured terms and on performing unification modulo subtyping between term constructors, which are not available in Datalog.

The idea of representing programs in a database and implementing analyses as queries is an old one, going back at least to Linton [25]. Much work has gone into proposals for new query languages specifically designed with program analysis in mind [11, 21, 22, 26, 18, 8], while others have explored the use of standard query languages such as relational algebra [29], Prolog [13], and plain Datalog [20]. An earlier version of QL was also previously described informally in this context [16, 15] without, however, discussing its precise semantics.

Recently, Datalog has been used to specify and implement highly scalable flow analyses [7, 33] on the LogicBlox platform [4], which implements a dialect of Datalog with language

extensions including support for state updates and a principled mechanism for creating fresh values through existentially quantified head variables. It would be interesting to investigate whether these concepts could be fruitfully combined with QL's object-oriented features.

Virtual dispatch in QL is a form of predicate dispatch [17], with class characters as guards. While exhaustiveness is ensured (that is, each call has at least one dispatch candidate), we make no attempt to prevent ambiguity: calls with more than one actual target are fully supported. Also, our translation from QL to Datalog is non-modular and requires reasoning about the entire program, unlike more recent work on predicate dispatch [28].

Prescriptive type systems have been studied in the context of Prolog [37, 24]. As pointed out by Meyer [27], prescriptive type annotations are essentially runtime type checks; hence, in spite of its static type declarations QL is, in some sense, dynamically typed.

7 Conclusion

We have presented QL, a declarative object-oriented query language based on Datalog. We have described QL's concepts of classes, subtyping and dynamic dispatch both informally and by means of a translation to plain Datalog: classes are unary predicates representing sets of values; subtyping is set inclusion; and dynamic dispatch resolves calls in the smallest class including the receiver value. As a typical application of QL, we have shown its use in implementing static checks, and presented a case study highlighting its advantages in this domain over Java. Finally, we have discussed QL's merits as an object-oriented language: while it is missing facilities for creating new objects or mutating object state, QL does offer the twin features of abstraction and dynamic dispatch, usually considered to be at the heart of object-oriented programming, without relying on objects in the traditional sense. Apart from QL's practical usefulness, its model of object orientation is an interesting contribution in itself, which, we hope, will spur further discussion of and investigation into the nature of object oriented programming.

References

1 Serge Abiteboul and Anthony J. Bonner. Objects and Views. In *SIGMOD*, 1991.
2 Serge Abiteboul, Georg Lausen, Heinz Uphoff, and Emmanuel Waller. Methods and Rules. In *SIGMOD*, 1993.
3 Hassan Aït-Kaci and Roger Nasr. LOGIN: A Logic Programming Language with Built-In Inheritance. *JLP*, 3(3), 1986.
4 Molham Aref, Balder ten Cate, Todd J. Green, Benny Kimelfeld, Dan Olteanu, Emir Pasalic, Todd L. Veldhuizen, and Geoffrey Washburn. Design and Implementation of the LogicBlox System. In *SIGMOD*, 2015.
5 Malcolm Atkinson, François Bancilhon, David DeWitt, Klaus Dittrich, David Maier, and Stanley Zdonik. The Object-Oriented Database System Manifesto. In *DOOD*, 1989.
6 François Bancilhon, David Maier, Yehoshua Sagiv, and Jeffrey D. Ullman. Magic Sets and Other Strange Ways to Implement Logic Programs. In *PODS*, 1986.
7 Martin Bravenboer and Yannis Smaragdakis. Strictly Declarative Specification of Sophisticated Points-to Analyses. In *OOPSLA*, 2009.
8 Mariano Consens, Alberto Mendelzon, and Arthur Ryman. Visualizing and Querying Software Structures. In *ICSE*, 1992.
9 William R. Cook. On Understanding Data Abstraction, Revisited. In *OOPSLA*, 2009.
10 William R. Cook. A Proposal for Simplified, Modern Definitions of "Object" and "Object Oriented". `http://wcook.blogspot.co.uk/2012/07/proposal-for-simplified-modern.html`, 2012.
11 Roger Crew. ASTLOG: A Language for Examining Abstract Syntax Trees. In *DSL*, 1997.

12 Hugh Darwen and C. J. Date. The Third Manifesto. *SIGMOD Records*, 24(1), 1995.

13 Stephen Dawson, C. R. Ramakrishnan, and David S. Warren. Practical Program Analysis Using General Purpose Logic Programming Systems. In *PLDI*, 1996.

14 Oege de Moor, Damien Sereni, Pavel Avgustinov, and Mathieu Verbaere. Type Inference for Datalog and Its Application to Query Optimisation. In *PODS*, 2008.

15 Oege de Moor, Damien Sereni, Mathieu Verbaere, Elnar Hajiyev, Pavel Avgustinov, Torbjörn Ekman, Neil Ongkingco, and Julian Tibble. .QL: Object-Oriented Queries Made Easy. In *GTTSE*, 2007.

16 Oege de Moor, Mathieu Verbaere, Elnar Hajiyev, Pavel Avgustinov, Torbjörn Ekman, Neil Ongkingco, Damien Sereni, and Julian Tibble. Keynote Address: .QL for Source Code Analysis. In *SCAM*, 2007.

17 Michael D. Ernst, Craig S. Kaplan, and Craig Chambers. Predicate Dispatching: A Unified Theory of Dispatch. In *ECOOP*, 1998.

18 Simon Goldsmith, Robert O'Callahan, and Alexander Aiken. Relational Queries over Program Traces. In *OOPSLA*, 2005.

19 Google. Error Prone. `http://errorprone.info/`, 2015.

20 Elnar Hajiyev, Mathieu Verbaere, and Oege de Moor. *CodeQuest:* Scalable Source Code Queries with Datalog. In *ECOOP*, 2006.

21 Doug Janzen and Kris De Volder. Navigating and Querying Code Without Getting Lost. In *AOSD*, 2003.

22 Stan Jarzabek. Design of Flexible Static Program Analyzers with PQL. *TSE*, 24(3), 1998.

23 Michael Kifer and James Wu. A Logic for Object-oriented Logic Programming. In *PODS*, 1989.

24 T. L. Lakshman and Uday S. Reddy. Typed Prolog: A Semantic Reconstruction of the Mycroft-O'Keefe Type System. In *ISLP*, 1991.

25 Mark Linton. Implementing Relational Views of Programs. In *SDE*, 1984.

26 Michael C. Martin, V. Benjamin Livshits, and Monica S. Lam. Finding Application Errors and Security Flaws Using PQL: A Program Query Language. In *OOPSLA*, 2005.

27 Gregor Meyer. On Types and Type Consistency in Logic Programming. Technical Report Informatik Berichte 199, FernUniversität Hagen, 1996.

28 Todd D. Millstein, Christopher Frost, Jason Ryder, and Alessandro Warth. Expressive and Modular Predicate Dispatch for Java. *TOPLAS*, 31(2), 2009.

29 Santanu Paul and Atul Prakash. A Query Algebra for Program Databases. *TSE*, 22(3), 1996.

30 Teodor C. Przymusinski. On the Declarative Semantics of Deductive Databases and Logic Programs. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*. Morgan Kaufmann Publishers Inc., 1988.

31 Max Schäfer and Oege de Moor. Type Inference for Datalog with Complex Type Hierarchies. In *POPL*, 2010.

32 Damien Sereni, Pavel Avgustinov, and Oege de Moor. Adding Magic to an Optimising Datalog Compiler. In *SIGMOD*, 2008.

33 Yannis Smaragdakis and Martin Bravenboer. Using Datalog for Fast and Easy Program Analysis. In *Datalog Reloaded*, 2010.

34 Gert Smolka and Hassan Aït-Kaci. Inheritance Hierarchies: Semantics and Unification. *JSC*, 7(3/4), 1989.

35 Jeffrey D. Ullman. A Comparison between Deductive and Object-Oriented Database Systems. In *DOOD*, 1991.

36 Peter Wegner. Dimensions of Object-Based Language Design. In *OOPSLA*, 1987.

37 Jiyang Xu and David S. Warren. Semantics of Types in Logic Programming. Technical Report DPS-102, ECRC, Munich, 1990.

38 Carlo Zaniolo. Object Identity and Inheritance in Deductive Databases—an Evolutionary Approach. In *DOOD*, 1989.