



On the Design, Implementation, and Use of Laziness in R

AVIRAL GOEL, Northeastern University, USA

JAN VITEK, Czech Technical University and Northeastern University, USA

The R programming language has been lazy for over twenty-five years. This paper presents a review of the design and implementation of call-by-need in R, and a data-driven study of how generations of programmers have put laziness to use in their code. We analyze 16,707 packages and observe the creation of 270.9 B promises. Our data suggests that there is little supporting evidence to assert that programmers use laziness to avoid unnecessary computation or to operate over infinite data structures. For the most part R code appears to have been written without reliance on, and in many cases even knowledge of, delayed argument evaluation. The only significant exception is a small number of packages which leverage call-by-need for meta-programming.

CCS Concepts: • **General and reference** → **Empirical studies**; • **Software and its engineering** → **General programming languages**; **Scripting languages**; *Semantics*.

Additional Key Words and Phrases: R language, delayed or lazy evaluation

ACM Reference Format:

Aviral Goel and Jan Vitek. 2019. On the Design, Implementation, and Use of Laziness in R. *Proc. ACM Program. Lang.* 3, OOPSLA, Article 153 (October 2019), 27 pages. <https://doi.org/10.1145/3360579>

1 INTRODUCTION

Since its inception, in 1993, R has had a call-by-need semantics. When a function is invoked its arguments are packaged up into *promises* which are evaluated on demand. The values obtained by evaluating those promises are memoized to avoid the need for recomputation. Thus the following definition when called with `a+b` and `d+d`

```
f <- function(x,y) x + x
```

evaluates `a+b` and does not evaluate `d+d`. With an estimated two million users world-wide [Smith 2011], R is the most widely used lazy functional programming language in existence. The R community has developed an extensive body of reusable, documented, tested and maintained code. The CRAN open source repository, the main source for R packages, hosts over 16,000 packages with an average of 6 new ones added every day [Ligges 2017]. Bioconductor [Gentleman et al. 2004], another open source R package repository, hosts over 1,700 packages for genomic data analysis.

It is fascinating to observe that R's laziness mostly remains secret. The majority of end-users are unaware of the semantics of the language they write code in. Anecdotally, this holds even for colleagues in the programming language community who use R casually. Moreover, we do not know of any studies of the design and efficacy of call-by-need in R. With twenty-five years of practical experience with laziness, some lessons can surely be drawn.

Hudak [1989] defined lazy evaluation as the implementation of normal-order reduction in which recomputation is avoided. He went on to enumerate two key benefits for programmers: (1) Sub-computations are only performed if they are needed for the final result; (2) Unbounded data structures include elements which are never materialized. Haskell is a language designed and implemented to support lazy evaluation, its compiler has optimization passes that remove some of



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2019 Copyright held by the owner/author(s).

2475-1421/2019/10-ART153

<https://doi.org/10.1145/3360579>

the overhead of delayed evaluation, and its type system allows laziness to co-exist with side effects in an orderly manner.

R differs from Haskell in its approach to lazy evaluation. The differences are due in part to the nature of the language and in part to the goals of its designers. As R frequently calls into legacy C and Fortran libraries, performance dictates that the memory layout of R objects be consistent with the expectations of those libraries. For statistical and mathematical codes, this mostly means array of primitive types, integer and floating point, should be laid out contiguously using machine representations for primitives. Interoperability is thus the reason for builtin datatypes being strict, and consequently for mostly giving up on the second benefit of laziness right out of the gate. As for the first benefit, it goes unfulfilled because R tries to only be as lazy as it needs. In numerous places, design choices limit its laziness in favor of a more predictable order of execution; this is compounded by a defensive programming style adopted in many packages where arguments are evaluated to obtain errors early.

Given the above, one may wonder *why bother being lazy?* In particular, when this implies run-time costs that can be significant as R does not optimize laziness. Communications with the creators of R suggests call-by-need was added to support meta-programming and, in particular, user-defined control structures. While R was inspired by Scheme, the latter's macro-based approach to meta-programming was not adopted. Macros were deemed too complex for users. Instead, R offers a combination of call-by-need and reflection. Call-by-need postpones evaluation so that reflection can inspect and modify expressions, either changing their binding environment or their code. This approach was deemed sufficiently expressive for the envisioned use-cases and, unlike macros, it was not limited to compile-time redefinition—an important consideration for an interactive environment. Evaluating how that choice worked out is an interesting but separate question.

This paper describes the design and implementation of call-by-need in the production R virtual machine, and provides data about its use in practice. The latter is obtained by dynamic analysis of a corpus of programs hosted in the CRAN and Bioconductor software repositories. Dynamic analysis is limited to behaviors that it observes, low code coverage may affect the results. Values that appear not to be evaluated (or always evaluated in the same order) may be evaluated (or evaluated in different orders) in control flow branches that were not exercised. We are encouraged by the fact that Krikava and Vitek [2018] reports line-level code coverage of over 60% for a similar corpus. To mitigate this threat to validity, we perform a qualitative analysis of a sample of our corpus. We claim the following contributions:

- We describe the design and implementation of call-by-need in R and give a small-step operational semantics for a subset of the language that includes promises, eval, substitute and delayedAssign.
- We release an open source, carefully optimized, dynamic analysis pipeline consisting of an instrumented R interpreter and data analysis scripts.
- We present results about the strictness of functions, their possible evaluation orders, as well as the life cycle of promises from the analysis of 232,290 scripts exercising code from 16,707 packages.

Our semantics has not been validated by testing. Our results were obtained with version 3.5.0 of GNUR and packages retrieved on August 1st, 2019 from CRAN and Bioconductor. Our software and data was validated as Functional and Reusable and is available in open source from:

<https://doi.org/10.5281/zenodo.3369573>

2 BACKGROUND ON R

The R project is a tool for implementing sophisticated data analysis algorithms. At heart, R is a *vectorized, dynamic, lazy, functional and object-oriented* programming language with a rather unusual combination of features [Morandat et al. 2012], designed to be easy to learn by non-programmers and to enable rapid development of new statistical methods. The language was created in 1993 by Ihaka and Gentleman [1996] as a successor to an earlier language for statistics named S [Becker et al. 1988]. We focus on features relevant to this work.

In R, most data types are vectorized. Values are constructed by the `c(...)` function: `c("hi", "ho")` creates a vector of two strings. The language does not differentiate scalars from vectors, thus `1==c(1)`. In order to enable equational reasoning, R copies most values accessible through multiple variables if they are written to. Consider the `swap` function which exchanges two elements in a vector and returns the modified vector:

```
> swap <- function(x, i, j) { t<-x[i]; x[i]<-x[j]; x[j]<-t; x }
> v <- c(1,2,3)
> swap(v,1,3)
```

The argument vector `v` is shared, as it is aliased by `x` in the function. Thus, when `swap` first writes to `x` at offset `i`, the vector is copied, leaving `v` unchanged. It is the copy that is returned by `swap`. Behind the scenes, a reference count is maintained for all objects. Aliasing a value increases the count. Any update of a value with a count larger than one triggers a copy. One motivation for this design was to allow users to write iterative code that updates vectors in place. A loop that updates all elements of an array will copy at most once.

Every linguistic construct is desugared to a function call, even control flow statements, assignments, and bracketing. Furthermore, all functions can be redefined in packages or user code. This makes R both flexible and challenging to compile [Flückiger et al. 2019]. A function definition can include default expressions for parameters, these can refer to other parameters. R functions are higher-order. The following snippet declares a function `f` which takes a variable number of arguments (triple dots are a `vararg`), whose parameters `x` and `y`, if missing, evaluate to `y` and `3*x`. The function returns a closure.

```
> f <- function(x=y, ..., y=3*x) { function(z) x+y+z }
```

Function `f` can be called with no arguments `f()`, with a single argument `f(3)`, with named arguments `f(y=4, x=2)` and with a variable number of arguments, `f(1, 2, 3, 4, y=5)`.

Values can be tagged by user-defined attributes. For instance, one can attach the attribute `dim` to the value `x<-c(1,2,3,4)` by evaluating `attr(x,"dim")<-c(2,2)`. Once done, arithmetic functions will treat `x` as a 2×2 matrix. Another attribute is `class` which can be bound to a list of names. For instance, `class(x)<-"human"`, sets the class of `x` to `human`. Attributes are used for object-oriented dispatch. The “S3 object system” supports single dispatch on the class of the first argument of a function, whereas the “S4 object system” allows dispatch on all arguments. These names refer to the version of the S language which introduced them. Popular data types, such as data frames, also leverage attributes. A data frame is a list of vectors with `class` and `colname` attributes.

R supports reflection and meta-programming. The `substitute(e, envir)` function yields the parse tree of the expression `e` after performing substitutions defined by the bindings in `envir`.

```
> substitute(expression(a + b), list(a = 1))
expression(1 + b)
```

R allows programmatic manipulation of parse trees, which are themselves first class objects. They are evaluated using the `eval(e, envir)` function. Environment on the call stack can also be accessed.

3 CALL-BY-NEED IN R

The combination of side effects, frequent interaction with C, and absence of types has pushed R to be more eager than other lazy languages. This section reviews the design and implementation of laziness. In R, arguments to a user-defined function are bundled into a thunk called a *promise*. Logically, a promise combines an expression's code, its environment, and its value. To access the value of a promise, one must *force* it. Forcing a promise triggers evaluation and the computed value is captured for future reference. The following snippet defines a function *f* that takes argument *x* and returns *x*+*x*. When called with an argument that has the side effect of printing to the console, the side effect is performed once as the second access to the promise is cached.

```
> f <- function(x) x+x
> f( {print("Hi!");2} )
"Hi!"
4
```

Promises associated to parameters' default values have access to all variables in scope, including other parameters. Promises cannot be forced recursively, that is an error. Promises are mostly encapsulated and hidden from user code. R only provides a small interface for operating on promises:

- **delayedAssign(x, exp, eenv, aenv)**: create a promise with body *exp* and binds it to variable *x* (where *x* is a symbol). Environment *eenv* is used to evaluate the promise, and *aenv* is used to perform the assignment.
- **substitute(e, env)**: substitutes variables in *e* with their values found in environment *env*, returns an expression (a parse tree).
- **force(x)**: forces the promise *x*. This replaces a common programming idiom, *x*<-*x*, which forces *x* by assigning it to itself.
- **forceAndCall(n, f, ...)**: call *f* with the arguments specified in the varargs of which the first *n* are forced before the call.

While R does not provide built-in lazy data structures, they can be encoded. Figure 1 shows a lazy list that uses environments as structs; environments have reference semantics. R provides syntactic sugar for looking up variables (*\$*), functions for creating environment out of lists (*list2env*) and for capturing the current environment (*environment*). The singleton *nil* has a tag that is tested in the empty function. A new list is created by *cons*; it returns its environment in which *h* and *t* are bound to promises. The *head* and *tail* functions retrieve the contents of *h* and *t* respectively. This example also illustrates how promises can be returned from their creation environment, namely by being protected by an environment.

```
nil <- list2env(list(tag="nil"))
empty <- function(l) l$tag=="nil"
cons <- function(h,t) environment()
head <- function(l) l$h
tail <- function(l) l$t
```

Fig. 1. Lazy list in R

R evaluates promises aggressively. The sequencing operator *a*;*b* will evaluate both *a* and *b*, assignment *x*<-*a* evaluates *a*, and **return** also triggers evaluation. In addition, many core functions are strict. R has two kinds of functions that are treated specially:

- *Builtins*: There are 680 builtins, typically written in C, providing efficient implementations of numerical methods and other mathematical functions. The argument lists of builtins are evaluated eagerly.
- *Specials*: There are 46 specials used to implement core language features such as loops, conditionals, bracketing, etc. These functions take expressions (parse trees) which are evaluated in the calling environment or in a specially constructed environment.

Builtins and specials are exposed as functions to the surface language either directly or through wrappers which perform pre-processing of arguments before passing them to these functions.

We would be remiss if we did not mention context-sensitive lookup, one of the unusual features of R. When looking up a variable x , in head position of a function call, e.g., $x(\dots)$, R finds the first definition of x in a lexically enclosing environment, if x is bound to a closure, that closure is returned. Otherwise, lookup continues in the enclosing scope. A corollary of function lookup is that it forces promises encountered along the way.

3.1 Implementing Promises

A promise has four slots: *exp*, *env*, *val* and *forced*. The *exp* slot contains a reference to the code of the promise, the *env* refers to environment in which the promise was originally created. The *val* slot holds the result of evaluating the *exp*. The *forced* flag is used to avoid recursion. When a promise is accessed, the *val* slot is inspected first. If it is not empty, that value is returned. Otherwise, *forced* is checked, and if it is set an exception is thrown. The *forced* flag is updated and the expression is evaluated in the specified environment. Once the evaluation finishes, the *val* slot is bound to the result, the *env* slot is cleared to allow the environment to be reclaimed, and the *forced* flag is unset. The implementation does little to optimize promises. In some cases, a promise can be created pre-forced with a value pre-assigned. The GNUR implementation recently added a bytecode compiler, this compiler eliminates promises when they contain a literal [Tierney 2019].

4 SEMANTICS

This section describes a small-step operational semantics in the style of [Wright and Felleisen 1992] for a core R language with promises. Our goal is to provide an easy to follow—the entire semantics fits on a page—but precise—unlike the above prose description—account of R’s call-by-need semantics. We build upon the semantics of Core R [Morandat et al. 2012], but omit vectors and out-of-scope assignments. Instead, we add delayed assignment, default values for arguments, *substitute* and *eval*. To support these features we add strings as a base type and the ability to capture the current environment.

Figure 2 gives the syntax and semantics of our calculus. The surface syntax includes terms for strings, variables, string concatenation, assignment, function declaration, function invocation (one and zero argument functions), environment capture, substitution, *eval*, and delayed assignment. The syntax is extended with additional terms used during reduction where variables and expressions can be replaced by values ranged over by meta-variable v which can be one of a string (s), a closure ($\lambda x = e.e, E$), an environment ($\text{env}(l)$), or a promise ($\text{prom}(l)$). Mutable values are heap allocated and ranged over by meta-variable l . We use the \perp value to denote an invalid reference.

The reduction relation is of the form $S E \rightarrow S'; E'$ where the stack S is a collection of expression-environment pairs ($e E$) and the heap maps variables to values. Frames are mutable and can be shared between closures, so they are stack allocated. Promises are quadruples (v, e, E, R) where v is the cached result of evaluating the body e in environment E . R is a status flag where \downarrow indicates we have started evaluating the promise. Evaluation contexts C are deterministic. Following R, some builtin operations such as string concatenation are strict. But function calls are lazy in their

$e ::= s \mid x \mid e \# e \mid x \leftarrow e \mid \text{fun}(x=e)e \mid x(e) \mid x() \mid \text{env} \mid \text{subst}(x) \mid \text{eval}(e, e) \mid \text{delay}(x, e, e)$	
$\mathbb{C} ::= [] \mid \mathbb{C} \# e \mid v \# \mathbb{C} \mid x \leftarrow \mathbb{C} \mid \mathbb{C}(e) \mid \mathbb{C}() \mid \text{eval}(\mathbb{C}, e) \mid \text{eval}(v, \mathbb{C}) \mid \text{delay}(x, e, \mathbb{C})$	
$F ::= \epsilon \mid F[x \mapsto l]$	frames
$E ::= \epsilon \mid l \cdot E$	environments
$S ::= \epsilon \mid e E \cdot S$	stacks
$H ::= \epsilon \mid H[l \mapsto v] \mid H[l \mapsto F] \mid H[l \mapsto (v, e, E, R)]$	heap
$R ::= \uparrow \mid \downarrow$	forced flag
<hr/>	
$\text{Fun} \quad \frac{v = (\lambda x = e.e', E)}{\mathbb{C}[\text{fun}(x=e)e'] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$	$\text{Concat} \quad \frac{}{\mathbb{C}[s \# s'] E \cdot S; H \rightarrow \mathbb{C}[ss'] E \cdot S; H'}$
$\text{Assign} \quad \frac{E = l' \cdot E' \quad H(l') = F \quad F' = F[x \mapsto v] \quad H' = H[l' \mapsto F']}{\mathbb{C}[x \leftarrow v] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$	$\text{Delay} \quad \frac{H(l) = l' \cdot E' \quad H(l') = F \quad \text{fresh } l'' \quad F' = F[x \mapsto \text{prom}(l'')] \quad H' = H[l'' \mapsto (\perp, e, E, \downarrow)][l' \mapsto F']}{\mathbb{C}[\text{delay}(x, e, \text{env}(l))] E \cdot S; H \rightarrow \mathbb{C}[\text{env}(l)] E \cdot S; H'}$
$\text{Env} \quad \frac{\text{fresh } l \quad H' = H[l \mapsto E]}{\mathbb{C}[\text{env}] E \cdot S; H \rightarrow \mathbb{C}[\text{env}(l)] E \cdot S; H'}$	$\text{Subst} \quad \frac{\text{get}(H, E, x) = v \quad v = (_, e, _, _) \quad \text{string}(e) = s}{\mathbb{C}[\text{subst}(x)] E \cdot S; H \rightarrow \mathbb{C}[s] E \cdot S; H'}$
$\text{Eval} \quad \frac{e = \text{eval}(s, \text{env}(l)) \quad \text{parse}(s) = e' \quad H(l) = E'}{\mathbb{C}[e] E \cdot S; H \rightarrow e' E' \cdot \mathbb{C}[e] E \cdot S; H'}$	$\text{EvalRet} \quad \frac{e = \text{eval}(s, \text{env}(l'))}{v E' \cdot \mathbb{C}[e] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$
$\text{Invk1} \quad \frac{v = (\lambda x = e'.e'', E') \quad \text{fresh } l, l' \quad E'' = l \cdot E' \quad H' = H[l \mapsto F] \quad F = [x \mapsto l'] \quad H'' = H'[l' \mapsto (\perp, e, E, \downarrow)]}{\mathbb{C}[v(e)] E \cdot S; H \rightarrow e'' E'' \cdot \mathbb{C}[v(e)] E \cdot S; H''}$	$\text{Invk0} \quad \frac{v = (\lambda x = e'.e'', E') \quad \text{fresh } l, l' \quad E'' = l \cdot E' \quad H' = H[l \mapsto F] \quad F = [x \mapsto l'] \quad H'' = H'[l' \mapsto (\perp, e', E'', \downarrow)]}{\mathbb{C}[v()] E \cdot S; H \rightarrow e'' E'' \cdot \mathbb{C}[v()] E \cdot S; H''}$
$\text{Ret1} \quad \frac{}{v E' \cdot \mathbb{C}[v'(e)] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$	$\text{Ret0} \quad \frac{}{v E' \cdot \mathbb{C}[v'()] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$
$\text{Lookup} \quad \frac{\text{get}(H, E, x) = v \quad v \neq \text{prom}(l)}{\mathbb{C}[x] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$	$\text{Lookup2} \quad \frac{\text{get}(H, E, x) = \text{prom}(l)}{\mathbb{C}[x] E \cdot S; H \rightarrow \text{prom}(l) E \cdot \mathbb{C}[x] E \cdot S; H'}$
$\text{Force} \quad \frac{H(l) = (\perp, e, E', \downarrow) \quad H' = H[l \mapsto (\perp, e, E', \uparrow)]}{\text{prom}(l) E \cdot S; H \rightarrow e E' \cdot \text{prom}(l) E \cdot S; H'}$	$\text{ReadVal} \quad \frac{H(l) = (v, e, \epsilon, \downarrow)}{\text{prom}(l) E \cdot S; H \rightarrow v E \cdot S; H'}$
$\text{Memo} \quad \frac{v \neq \text{prom}(l'') \quad H(l) = (\perp, e, E', \uparrow) \quad H' = H[l \mapsto (v, e, \epsilon, \downarrow)]}{v E' \cdot \text{prom}(l) E \cdot S; H \rightarrow v E \cdot S; H'}$	$\text{RetProm} \quad \frac{v \neq \text{prom}(l)}{v E' \cdot \mathbb{C}[x] E \cdot S; H \rightarrow \mathbb{C}[v] E \cdot S; H'}$
<hr/>	
$\frac{E = l \cdot E' \quad H(l) = F \quad F(x) = v}{\text{get}(H, E, x) = v}$	$\frac{E = l \cdot E' \quad H(l) = F \quad x \notin \text{dom}(F)}{\text{get}(H, E, x) = \text{get}(H, E', x)}$
<hr/>	

Fig. 2. Syntax and Semantics

argument, the expression in x (e) remains untouched by the context. We omit the definition of the functions *parse* and *string* which, respectively, turn strings into expressions and vice-versa. The expression *fresh* is used to obtain new heap references.

The semantics is given by the following rules. Rules *Fun* and *Concat* deal with creating a closure in the current environment and concatenating string values. Rule *Assign* will add a mapping from variable x to value v in the current frame. As frames are allocated in the heap, this updates the heap. Rule *Delay* performs a delayed assignment, that is to say, an assignment that does not force the right-hand side. It takes a variable, an expression and an environment in which this expression will be evaluated. The rule creates a new unevaluated promise and binds it to x . Rule *Env* grabs the current environment and returns it as a value. Rule *Subst* looks up the variable given as argument, obtains the promise bound to it, extracts its body and deparses it into a string. Rule *Eval* takes a string and an environment, parses that string into an expression and schedules it for execution. Rule *EvalRet* takes the result of that evaluation and replaces the call to *eval* with it. Rule *Invk1* and *Invk0* handle user-defined function calls. Both rules expect v to be a closure. They allocate a new promise for x . They differ on the body of that closure and the environment. *Invk0* has no argument and will use the default expression e' specified in the function declaration, it uses the environment where x is bound for evaluation. Rules *Ret1* and *Ret0* are the corresponding returns rules that replace the call with the computed value. Rules *Lookup* and *Lookup1* are used to read variables from the current environment. If the result is a promise it is scheduled for execution by *Lookup1*. Rule *Force* will actually evaluate a promise that has been pushed on the stack, if that promise has not yet been evaluated. It sets the flag to avoid recursive evaluation. Rule *ReadVal* retrieves the value of an already evaluated promise. Rule *Memo* stores the result of evaluation in the promise and discard its environment. Finally, rule *RetProm* returns from evaluating a promise by replacing the variable looked up with the result.

4.1 Takeaways

R is rather strict for a lazy language. This manifests itself in the definition of evaluation contexts. Intuitively, any position where \mathbb{C} appears is evaluated strictly in left-to-right order. The key place where R differs from other lazy languages is that the right-hand side of assignments is strict. Our semantics does not show data structures, in R they are all strict. Strictness also shows up in the *Ret1* and *Ret0* rules which force the evaluation of the return expression. Lastly, strictness is enforced in the *Lookup2* rule which does function lookup. If a promise is returned, it must be evaluated. Another property that the semantics ensures is that promises are stored in environments and, whenever they are accessed they are forced. The only way for a promise to outlive the frame that created it is to be returned as part of an environment or closure. Following R, it is possible to create a cycle in promise evaluation, the expression $(\text{fun}(x = x) x)()$ when evaluated creates a closure and invokes it. The function's body triggers evaluation of the promise bound to x . Since no argument was provided, the default expression is evaluated causing a cycle. Like in R, this results in a stuck state in the semantics.

5 ANALYSIS INFRASTRUCTURE

We now explain the infrastructure that assembles the corpus and collects and analyzes traces. Our analysis pipeline starts with scripts to download, extract and install open source R packages. Next, an instrumented R virtual machine generates events from program runs. This is followed by an analyzer that processes the execution traces to generate tabular data files in a custom binary format. Other scripts post-process the data, compute statistics, and generate graphs. The entire pipeline is managed by a Makefile that invokes an R script to extract runnable code snippets from installed packages and runs the other steps in parallel. Parallelization is achieved using GNU Parallel.






















Corpus	Trace	Reduce	Combine	Merge	Summarize	Report
 10h	 2d 4h	 1d 15h	 3h 30m	 1h 24m	 2h 18m	 37 s
 16,707	 2.8 M	 7.4 M	 842	 36	 77	 15
 232,290	 5.1 TB	 76 GB	 20.4 GB	 21 GB	 1.5 GB	 268

Fig. 3. Tracing Pipeline

Figure 3 shows the main stages of our pipeline. We give corpus size, number of files generated, size of data and time taken by each step. The remainder of this section details the various stages.

5.1 Instrumented R

The instrumented R Virtual Machine is based on GNU-R version 3.5.0. Its goal is to produce program execution traces with all the events required to answer our research questions. On the face of it, this is not a difficult task. And, in the end, we only need to add 1,886 lines of C code to expose an event data structure with fields to describe a variety of execution events that capture the internal interpreter state. The challenge was identifying where to insert those 1,886 lines in an interpreter whose code is 542,809 LOC written over twenty-five years by many developers and outside contributors. The system has grown in complexity with an eclectic mix of ad hoc features designed to support growing user requirements. For instance, the code to manage environments and variable bindings in `main/envir.c` is over 2,864 LOC with 131 functions with a large number of identical code fragments for managing these data structures duplicated in various files. We succeeded by a lengthy trial and error process. The events recorded by the instrumented virtual machine are:

- **Call, Return:** at each function call, records function's type, arguments, environment and return value.
- **S3Dispatch:** at each S3 dispatch, records method name and first argument.
- **S4Dispatch:** at each S4 dispatch, records method name, definition, environment and dispatch arguments.
- **Eval:** at each eval, records the evaluated expression and its environment.
- **Substitute:** at each substitute, records the arguments.
- **ArgListEnter, ArgListExit:** records expressions that are being promised.
- **CtxtEnter, CtxtExit:** records the address of a stack frame.
- **CtxtJump:** records the popped stack frames at each non-local return.
- **PromEnter, PromExit:** records evaluated promise and when evaluation terminates.
- **PromRead:** when a promise's value is read, records the promise and its value.
- **PromSubst:** generated when a promise's expression is read.
- **FunLoadStart, FunLoadEnd:** generated when looking up a function.

- **GC**: generated at each garbage collection cycle.
- **Alloc**: generated when memory is allocated.
- **Free**: generated when memory is reclaimed.
- **Deserialize**: generated when an object is deserialized.
- **VarDef**: when a variable is defined, record the symbol, value and environment.
- **VarWrite**: when a variable is updated, record a the symbol, value and environment.
- **VarRem**: when a variable is deleted, records the symbol and environment.
- **VarRead**: when a variable is read, record the symbol, value and environment.

Events can be disabled to ignore implementation details of the virtual machine and also to avoid recursion. R objects captured in events are protected from the garbage collector to prevent them from being reclaimed during analysis. Table 1 shows the number of times events are triggered.

Table 1. Events

Call	831 B	CtxtEnter	143 B	Eval	1.7 T	PromEnter	223 B	VarDef	365 B
Return	807.4 B	CtxtExit	143 B	ArgListEnter	111 B	PromExit	223 B	VarWrite	140 B
GC	4.8 M	CtxtJump	6 B	ArgListExit	111 B	PromRead	102 B	VarRem	8.4 B
Alloc	2.4 T	S3Dispatch	2.6 B	FunLoadStart	78 B	PromSubst	1.9 B	VarRead	1.6 T
Free	2.3 T	S4Dispatch	936 M	FunLoadEnd	78 B	Deserialize	109 B	Substitute	1.7 B

5.2 Tracer

The tracer is a small R package (73 LOC) that calls into a larger C++ library (6,080 LOC). It is loaded in the instrumented R virtual machine and, during program execution, it maintains objects that model various aspects of the program such as functions, calls, promises, variables, environments, stacks and stack frames. As events are generated, the tracer updates its model of the state. The tracer is able to process 803.1 K events per second on our benchmark machine.

Some design decisions allowed the tracer to scale. Firstly, copying model objects is avoided as much as possible. They are created by a singleton factory that caches them in a global table. This optimization pays off as model objects are large and costly to copy. But keeping these objects alive too long will increase footprint and hinder any attempt at running multiple tracers on the same machine in parallel. To reduce tracing footprint, the R garbage collector was modified so that model objects can be deallocated as soon as the R object they represent is freed. One slightly surprising design choice is to link all model objects together. This pays off when an event triggers a cascade of changes to model objects. This comes at a price of course, as lists of model objects are circular it is necessary to perform reference counting to reclaim them. One last implementation trick is the use of a shadow stack that mirrors the stack maintained by the R virtual machine. The shadow stack is used to look up data after a longjump.

The tracer generates large amounts of data. Our first prototype used Sqlite to store the generated data. However, we found the approach limiting. During development we kept running into errors because the database schema and the tracer were out of sync. Due to the iterative nature of data analysis, we were modifying the schema frequently and this became a pain point. Furthermore, our database was normalized, thus requiring join operations in the analysis. At our scale, these joins were expensive, causing database operations to run for days. Lastly, insertions ended being a bottleneck; we could trace fewer than 1000 packages per day and filled up a 1 TB disk. In the end, we implemented a custom format. As the event stream has substantial amounts of redundancy, we applied streaming compression on the fly. Compression yields an average 10x saving in space and 12x improvement in loading time.

5.3 Execution

For each package to be analyzed our infrastructure must extract executable code from that package. Extraction invokes an R API which locates executable code snippets in the documentation and RMarkdown files. Files in the test directory are copied as is. All snippets and tests set up the tracer and initialize it with paths to input and output data before execution. For each package, the tracer generates 12 data files and 4 status files. These files denote the different possible states of the tracing. They allow the infrastructure to discard data from failing programs. The R scripts responsible for generating traces do extensive logging of intermediate steps for debugging purposes.

5.4 Post-processing

This part of the pipeline analyzes the raw data. It is 4K lines of R code. Scale was our major challenge. We faced difficulties both due to execution time and data size. In the 232,290 programs that were traced, the tracer observed 1.7 T expression evaluations, 831 B calls to 698.4K functions and 270.9 B promises. The raw data generated by the tracer is 5.1 TB but the reduced data is just 76 GB. In hindsight, it appears that incorporating analysis in the tracer, i.e., pre-summarizing data in C++ would have been beneficial. However, this would require knowing ahead of time all the analysis that we would perform. Part of the challenge lies in the fact that the event of interest and attached analyses were not fixed ahead of time. Pre-summarized data makes it harder to pose new questions. It also makes it harder to detect bugs because summarized data resists correlation with actual code.

The pipeline steps are detailed next.

- *Prescan*: Scan the raw data directory and output a list of all the subdirectories that contain raw data. There is a directory per package and multiple files in each directory.
- *Reduce*: Given a list of directories, this step uses GNU Parallel to partially summarize the raw data. This is the most expensive step in terms of size and speed. Since the data files are large, we must limit the degree of parallelism drastically to avoid running out of memory.
- *Scan*: Create a list of all the files successfully reduced.
- *Combine*: Combine information from all the programs into a single data table per analysis question.
- *Summarize*: Compute summaries of the merged data for: (1) event frequency, (2) object frequency, (3) functions with their definitions, (4) argument information, (5) escaped arguments, (6) information about parameters, (7) information about promises.
- *Report*: Generate graphs and tables from an RMarkdown notebook as well as \LaTeX macros for inclusion in the paper.

5.5 Threats to Validity

We have mentioned in the introduction that code coverage is a worry for any dynamic approach. There are two additional points to consider. Firstly, C and C++ functions can bypass the R extension API and directly modify R objects' internals. For example, set a promise's value without going through the API thus obviating our hooks. Such behavior breaks the R semantics and is error prone as the R internals do change. We have not observed this behavior in practice, but given the large number of packages, it may happen. Secondly, we disable the bytecode compiler for this study. Since the compiler eliminates promises for literal arguments, we observe more promises than actually created by the GNUR Virtual Machine under its default settings.

6 CORPUS OF R PROGRAMS

The corpus used in this study was assembled on August 1st, 2019 from the two main code repositories, namely the *Comprehensive R Archive Network* (CRAN) [Ligges 2017] and *Bioconductor* [Gentleman et al. 2004]. Both are curated repositories; to be admitted packages must conform to some well-formedness rules. In particular, they must contain use-cases and tests along with the data needed to run them. We believe this corpus is representative of sophisticated uses of the R language. Anecdotal evidence suggests that the majority of R code written is made up of small scripts, straight-line sequences of package calls, that read data, apply some models to it and then visualize the results. Most end-users neither define functions nor write loops, their code is simple. Without a source of end-user code it is not possible to validate this hypothesis, but if true then our corpus is representative of the interesting R code. R is also used in many industrial settings that do not publish their code to open source repositories. We have no information on those use-cases.

Our snapshot of CRAN includes 14,762 packages, and for Bioconductor, 3,087 packages. Bioconductor is also used to store data, 1,741 packages contain software, 1,319 contain data and 27 are so-called workflows. Starting with 17,849 software packages, our scripts downloaded and successfully installed 17,479 of them. The reasons some packages did not install were varied, they included missing dependencies and compiler errors.

These install errors may be fixable but automating those fixes would be hard. We chose to discard the packages that could not be installed. Out of the installed packages, we were able to successfully record execution traces for 16,707 packages. Some packages did not trace owing to run-time failures. Again, we discard the failing packages on the grounds of having a sufficient number of running ones

Table 2. Corpus

	Tests	Examples	Vignettes	
Scripts	44.1 K	220 K	9.8 K	<i>Install</i>
	23.3 K	202 K	6.6 K	<i>Trace</i>
LOC	2.7 M	1.6 M	614 K	<i>Install</i>
	1.3 M	1.6 M	327 K	<i>Trace</i>

For each package, our scripts gathered runnable code from three different sources: test cases, examples and vignettes. Test cases are typically unit tests written to exercise individual functions, while examples and vignettes demonstrate the expected end-user usage of the particular package. These use-cases may load other packages and access data shipped with the package or obtained from the internet. Table 2 gives the number of scripts of each kind that could be installed and the number of scripts that were successfully traced. In terms of lines of code, we exercised 25.6 M lines of R and 10.4 M lines of C. The total size of our database after analysis is 5.2 TB.

We observed 831 B calls to 698.4 K functions. 26.7% of the calls are made to 157 builtin functions, and 60.3% are made to 33 special functions. The remainder are calls to R functions. Of these, 2% are calls to 34.1 K different S3 methods and 0.8% are calls to 60.2 K S4 methods. There are 415.3 K plain R functions.

The number of times `delayedAssign` is called is 82 M, `force` is called 101.8 M times, `forceAndCall` is invoked 1.3 B times and `substitute` 1.7 B times. Functions exit due to `longjumps` 23.7 B times (this marks explicit use of the `return` function).

Figure 4 shows how many functions were exercised in each test package. More than 9 K (59.2%) of the packages had over 10 functions called. On the other hand 2.7 K of the packages had a single function invoked (7.0%). These may either be small packages, or, more likely, the provided tests have low coverage.

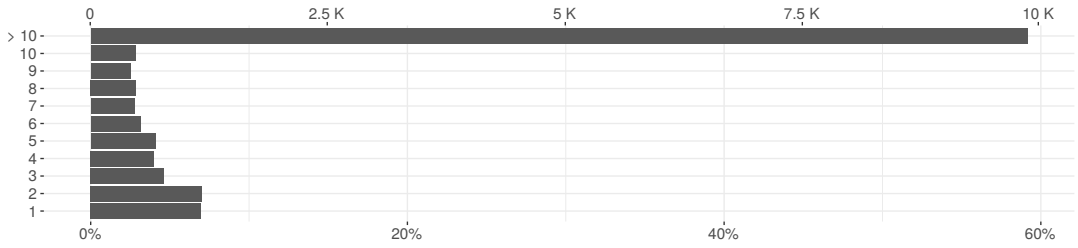


Fig. 4. Functions per package

Figure 5 shows how many times functions were called. 45.0% of the exercised functions were called more than ten times and 18.9% of the functions were called only once. Clearly functions that are called only once may lack coverage.

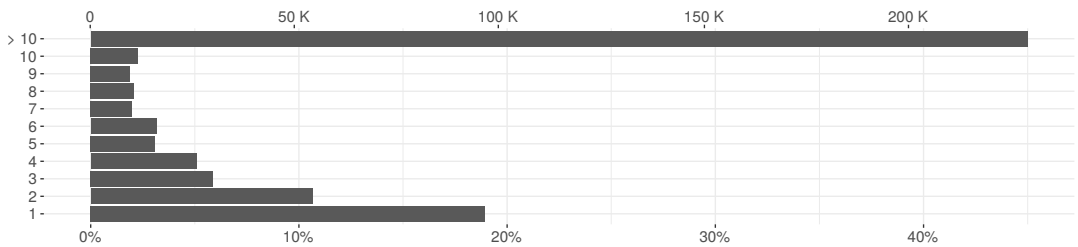


Fig. 5. Function calls

Figure 6 shows the number of parameters per function; 43.7% have only one, while 9.8% have more than five. The function `meta::forest.meta` has the most with 199 parameters. There are 2.1 M distinct parameters.

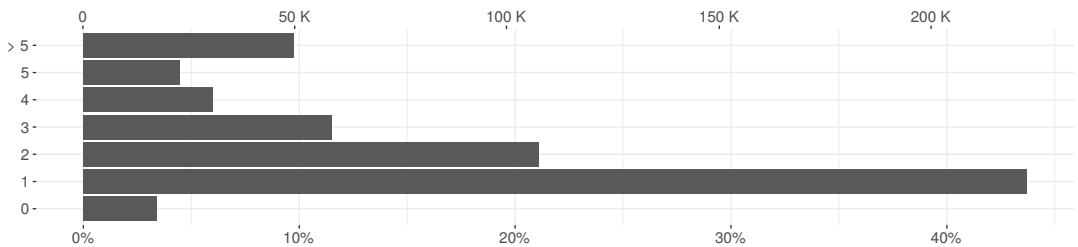


Fig. 6. Formal Parameters

Of the 261 B arguments that were passed at run-time, 25.1% were default arguments, 72.1% were non-default arguments and the remaining 2.8% were missing arguments.

7 ANALYZING LAZINESS USAGE PATTERNS

This section presents the results of our empirical study of call-by-need in the R language.

7.1 Life Cycle of Promises

The first research question we address is how promises are created and used in the wild.

RQ1: *What is the life cycle of R promises in the corpus?*

In our corpus, and likely, all R programs, promises are the most frequently allocated object. We observed the creation of 270.9 B promises. For context, Figure 7 shows the distribution of application-level objects; most are vector of characters, logical, integers, and doubles, in that order. Lists are often used in package code and internal functions. Raw values hold uninterpreted byte strings. Closures represent functions and environments map names to value, symbols are language-level names and S4 are instance of classes. Environments are frequently observed because one is created for each function call, but they are also, albeit rarely, used as hashmaps in user code.

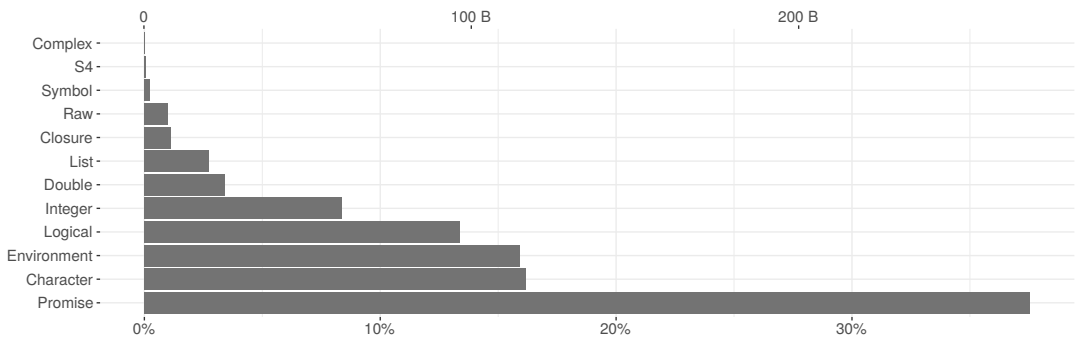


Fig. 7. Object counts

Where are promises created? Argument lists account for 94.3% of promises. The remainder are used for lazy loading of functions, are created by calls to `delayedAssign`, or in internal functions of the R virtual machine. One could expect that there would be more promises than values since every operation in R is a function call. This is not the case. Some values are composite of multiple simpler elements (e.g. data frames) and these are wrapped in a single promise. Values can be returned without being bound to a promise. Lastly, calls to internal (builtin and special) functions do not pack arguments in promises.

What do promises yield? Figure 8 shows the contents of promises that were forced. The most common types are character vector, logical vector, environment, closure, integer vector, double vector, null and list, in that order. The presence of null values is explained by the fact that many default parameter values are set to null and these default values are promised. S4 objects are rarely used in R programs outside of Bioconductor packages. Symbol, complex vector and raw vector are also quite rare in practice.

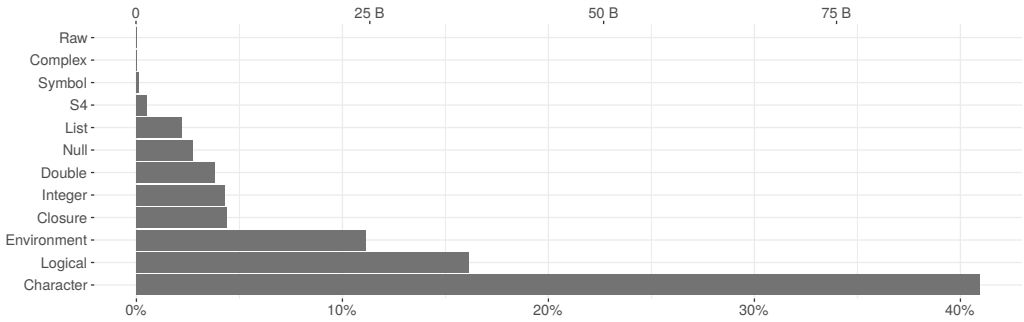


Fig. 8. Promise results

Figure 9 shows the content of the expression slot of promises, i.e. their code. Only 17.3% of promises contain a function call, e.g. $1+2$ or $f(z)$. The majority contain a single symbol to be looked up in the promise's defining environment, e.g. x . That symbol may be bound to another promise, in which case forcing the promise will be recursive. Promises can also hold inlined scalar constants, such as a single double 1.1 .

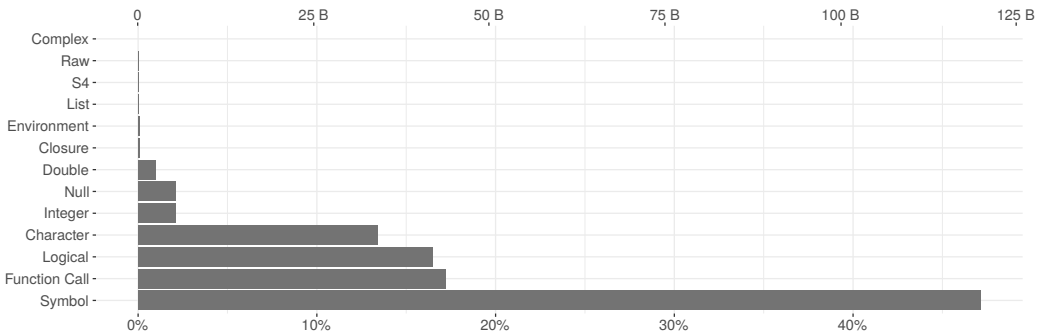


Fig. 9. Promise expressions

How often are promises accessed? 87.3% of argument promises are forced; the remaining went unused. Unused arguments are not unusual, some functions have over twenty parameters, and many of these are only needed in special circumstances. Figure 10 shows, for each individual promise, the number of times its value was read. Most promises are used once (forced), 9.6% are accessed twice, and 3% are accessed three times.

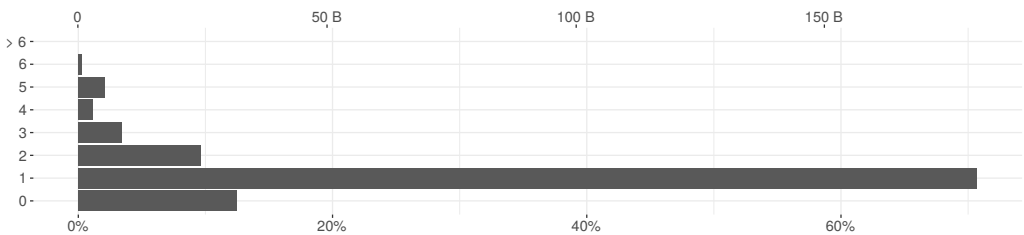


Fig. 10. Reads

How far are promises forced from their creation? Promises can be passed from one function to the next, traveling down the call stack. Regardless of the distance from the frame that created them, promises evaluate in their creation environment. But the farther from creation, the harder it is for a compiler to optimize them. Figure 11 shows the distance between promise creation and forcing. 79.7% promises are evaluated in the callee, 17.1% promises are forced two level down, and the remaining 3.2% are evaluated deeper.

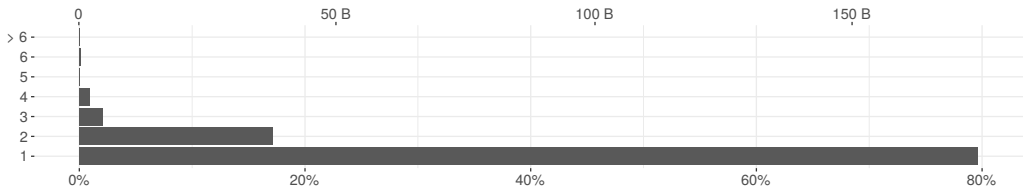


Fig. 11. Force depth

How long do promises live? Promises are short-lived, over 99.5% do not survive one garbage collection cycle. This confirms the folklore that most objects die young. It also means that promises are exerting pressure on the memory subsystem of the virtual machine. Only 0.5% of promises survive multiple cycles. Of those, 76.7% are non-argument promises and 0.08% are escaped arguments. As mentioned earlier, non-argument promises are created explicitly through `delayedAssign` and implicitly through lazy-loading of package code; both are expected to be long-lived. Escaped promises are promises that outlive their defining function. They may be long-lived as well. Of the long-lived promises, 23.2% are argument promises; their longevity is likely due to long running functions.

What are promise life cycles? If we characterize the life of a promise by the events that affect it, promise life cycles can be summarized by sequences of events. Ignoring creation and reclamation, the events of interest are Force, Read, Meta-program, Escape, Assign, and deSerialize. Note that forcing a promise is an implicit read. We observed 28.1 K unique life cycles. Table 3 shows the most frequent sequences for (a) argument promises (19.2 K unique sequences), (b) escaped argument promises (7.3 K sequences) and (c) non-

Table 3. Promise life cycle

F	70.5%	EF	70.7%	A	67.5%
–	11.9%	FRER	5.4%	–	16.5%
FR	9.5%	FER	5%	AR	11.2%
FRR	3%	FRRER	4%	AA	2%
FRRRR	2%	EFR	3%	F	2%
FRRR	1%	FRRERR	1%	S	0.2%
M	0.5%	EFRR	1%		
a. Argument		b. Escaped		c. Non-Argument	

argument promises (6.2 K sequences). For argument promises the two most common sequence are F (a promise that is forced) and the empty sequence (unused promise). The next sequences are forces followed by a growing number of reads. Meta-programming occurs only infrequently. For escaped promises the most frequent sequence is EF, the promise escapes and is forced later. The second most frequent sequence is FRER, the promise is forced, read, escapes and is read again. Lastly, non-argument promises are most often created and assigned a value in the C code. About 0.2% of these promises are obtained from deserialization (S) owing to lazy-loading of packages.

Does context sensitive lookup force promises? Looking up a function name such as `f()` may force a promise if that name is bound to one in the environment. If the promise yields a closure, that closure will be invoked, otherwise lookup continues. This allows “harmless” shadowing of function names as seen in Figure 12 where `addToGList.grob` from package `grid` defines a parameter `gList` that shadows a function of the same name defined by the same package.

```
addToGList.grob <- function(x, gList)
  if (is.null(gList)) gList(x) else { gList[[length(gList)+1L]]<-x; return(gList) }
```

Fig. 12. Shadowing

We observed 651.7 K function lookups (out of a total of 831 B lookups) which caused a promise to be forced. Out of those, 86.3% yielded a closure. The 13.7% that did not yield a closure are cases where a function name was shadowed. Table 4 shows the 30 most commonly shadowed function names and the number of functions in which shadowing happens. Many of those names correspond to common R functions such as `c`, `names`, `print` and `max`.

Table 4. Context sensitive lookup

plot	971	formula	204	order	83	ncol	71	list	53
log	821	scale	188	drop	82	dim	69	rug	49
c	461	t	184	which	82	max	67	matrix	48
legend	334	names	182	grid	76	format	66	clean	47
file	221	round	130	class	74	sort	60	start	45
length	209	title	129	print	74	nrow	57	unique	43

Do promises force each other? A parameter list can include parameters with default values that refer to each other. This is a semantic quirk that can lead to promises forcing one another. Consider Figure 13 and function `sample.int` from the base package. Parameter `useH` has a default value that depends on the other four arguments and `s` depends on `n`. Function `rmslash` has a recursive dependency between `center` and `Scatter`. The function expects at least one of those to be provided at each call site, if this is not the case an error will be reported.

```
sample.int <- function(n,s=n,r=F,p=NULL,useH=(!r && is.null(p) && s<=n/2 && n>1e+07))
  if (useH) .Internal(sample2(n,s)) else .Internal(sample(n,s,r,p))

rmslash <- function(center=rep(0,nrow(Scatter)), Scatter=diag(length(center))) {
  if (length(center) != nrow(Scatter)) stop("<error-message>")
  ...
```

Fig. 13. Argument lists

We found 4.8 K default value expressions in 3.9 K functions. At run-time, there were 336.9 M default expressions forcing other parameters. This is 0.2% of the forced promises.

Does method dispatch force promises? The two widely used object systems, S3 and S4, have an impact on promises. In order to find which method to invoke, one or more arguments must be forced. Overall, only 1% of promises participate in method dispatch. 671.5 M promises are forced due to S3 dispatch and 536.2 M are forced due to S4 dispatch. As these numbers are small, we ignore dispatch for the remainder of the study.

How often does non-local return happen? The `return` function pops the call stack until it arrives at the frame where it originated from. A call such as `f(return(1))` will, when `f`'s argument is forced, return from `f` and its caller. So, when a promise containing `return` is forced, the current function stops executing and the stack is unwound; this is a non-local return. Only 297.4 M arguments to 16 functions performed non-local returns. One of the most common causes is the `base::tryCatch` function. The function sequentially attaches handlers specified in its `vararg` list and executes `exp` by wrapping it in a call to `return`. The return causes the control flow to exit `doTryCatch` and `tryCatch`.

```
tryCatch <- function(exp, ..., fin) {
  ...
  doTryCatch <- function(e, nm, penv, handler) {
    .Internal(.addCondHands(nm, handler), penv, environment(), F))
    e
  }
  value <- doTryCatch(return(exp), nm, penv, handler)
  ...
}
```

Takeways. Promises dominate the memory profile of R programs. They are short lived, 80% are evaluated in the called function and over 99% do not survive a single GC cycle. The vast majority of promises contain a value or a variable. Only 17.3% contain code that needs to be evaluated. Of those expression-carrying promises 80.7% are unused, 7.0% are evaluated in the called function and 12.3% are evaluated down the call-stack or meta-programmed. Overall most promises lead a rather mundane life that one would hope a compiler could optimize out of existence.

7.2 Strictness

Our second research question concerns strictness. A function is said to be *strict* if it evaluates all of its arguments in a single pre-ordained order (e.g., left to right).

RQ2: *What proportion of R functions are strict?*

To answer this question we start with individual parameters. We only consider plain R functions that are called more than once, have at least one parameter and have not abruptly stopped executing owing to non-local returns. There 2.1 M distinct parameters to such functions. For a given parameter and a given function, we aggregate all calls and all uses of that parameter into three categories: parameters that are *Always* evaluated, parameters that are *Never* evaluated, and *Sometimes* evaluated. Figure 14 summarizes this analysis. 87.6% of parameters are always evaluated, 6.0% parameters are evaluated in some calls and not others, 6.4% parameters are never used.

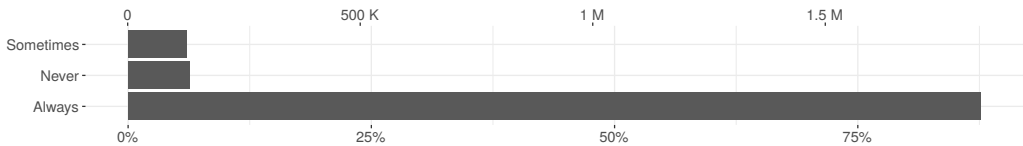


Fig. 14. Parameter strictness

A function is strict if its parameters are always evaluated in the same order. Most functions, 93.6% to be precise, have a single order of evaluation. This means they are candidate for being strict. Functions with multiple orders of evaluation for their arguments are summarized in Figure 15. About 4.7% of the functions have two force orders, and very few functions have more.

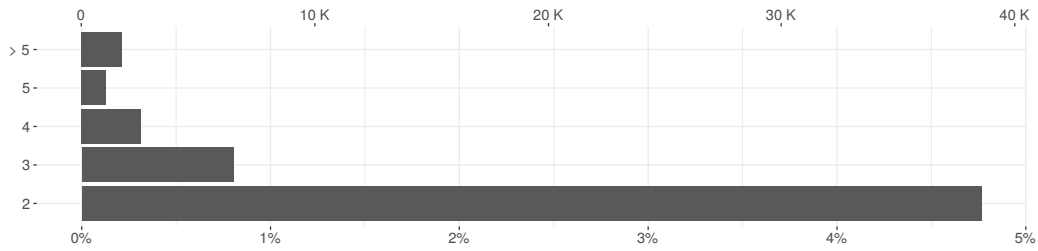


Fig. 15. Function force orders

Based on the above data, out of a total of 388.3 K functions, 83.7% are strict. Figure 16 gives a histogram of function strictness ratios per package. The majority of packages contain only strict functions. The packages that are less than 75% strict account for only 2.6 K packages (16.9% of all packages) and 81.7 K functions (21.0% of all functions).

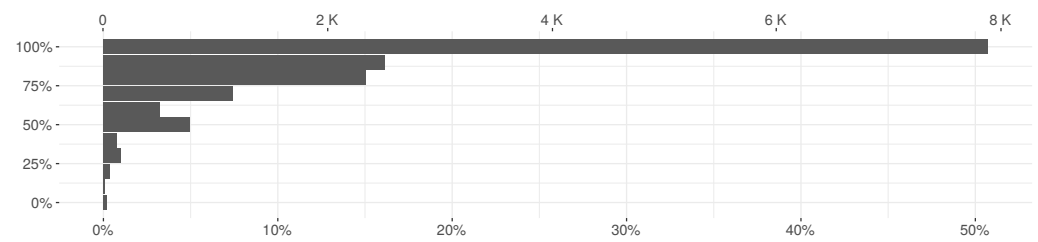


Fig. 16. Strictness per package (x-axis=packages; y-axis=strict function ratio)

One could consider relaxing strictness to allow multiple orders of evaluation if those orders of evaluation could be shown to be semantically equivalent. Some features of R may help. First, all vectors have a copy-on-write semantics, thus many side effects are hidden from view. Moreover, as Figure 17 shows only 25% of *Sometimes* promises perform any computation. In our experience most of those computations are side effect free.

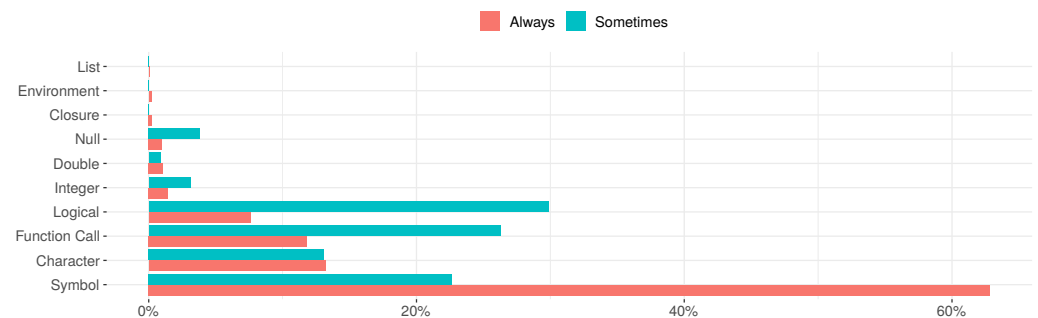


Fig. 17. Promise expressions

We performed an additional analysis to get an upper bound on the side effects performed during promise evaluation. We considered only variable reads and writes, external side effects such as filesystem operations were not taken into account. A meager 16.5 M promises (out of 270.9 B) perform any side-effecting computation. There are several cases to consider, the simplest when the promise performs a side effect to its local environment. For example, consider the `stats::power` function:

```
power <- function() {
  linkinv <- function(eta) pmax(eta^(1/lambda), .Machine$double.eps)
  mu <- linkinv(eta <- eta + offset)
  ...
}
```

The call to `linkinv` takes, as argument, an expression that performs a side effect to the local variable `eta`. This could have been avoided by moving the assignment above the call to `linkinv` but the programmer likely wanted to save one line. This kind of local side effect can affect other promises coming from the same environment (which is not the case here). Of the side-effecting promises, 41.2% are local. Non-local effects can be performed, e.g., using the `<-` operator to assign to a variable in the lexically enclosing environment. The following is snippet from a test script in `cliapp` package. The argument to `capt0` modifies `id` using `<-`.

```
test_that("auto_closing", {
  id <- ""
  f <- function() capt0(id <- cli_par(class = "xx"))
  capt0(f())
  ...
})
```

Out of the side-effecting promises, 0.5% affect a parent environment. Finally, 58.3% promises perform side effects in other environments. The `methods::callNextMethod` function is among the most common sources of mutation of other environments.

```
callNextMethod <- function (...) {
  ...
  callEnv <- parent.frame(1)
  assign(".nextMethod", nextMethod, envir = callEnv)
  ...
}
```

We further count how many of those side effects are performed directly by assignments occurring textually in the promise. Of the 6.8 M promises performing local side effects, 98% perform side effects directly. Of the 9.6 M promises performing side effects in other environments, 0.1% perform side effects directly. Of the 80.2 K promises performing side effects in the lexical parent environment, 0.08% perform side effects directly.

Qualitative analysis. We inspected 100 randomly selected functions that our dynamic analysis marked as strict. Out of those, 82 were indeed strict. The remaining 18 were not, but we did not observe their laziness. The majority (16 out of 18) of incorrectly labeled functions were not using all of their parameters, using them along some execution paths or returning early. We found a single function that was lazy because it called another function that was itself lazy in that particular argument and one function for which an argument escaped. The functions that do not evaluate all parameters could be cases where computational effort is saved if the arguments passed are complex expressions. We also found occurrences of explicit argument forcing. Programmers write code such as `x<-x` or `force(x)` to ensure that argument are values. An example of such code is the `scales::viridis_pal` function; it returns a closure, but forces all of its arguments to avoid capturing their environments:

```
function(alpha=1, begin=0, end=1, dir=1) {
  force_all(alpha,begin,end,dir)
  function(n) viridis(n,alpha,begin,end,dir)
}
```

The authors of higher-order functions such as `apply` or `reduce` often enforce strictness after receiving bug reports from end-users around unwanted lazy evaluation and variable capture interactions. The `forceAndCall` function has been introduced to mitigate this issue by forcing the arguments of a function prior to calling it. Looking for uses of `forceAndCall` revealed additional packages that enforce strictness for higher-order functions. The function is invoked 1.3 B times. The `force` function is also widely used to enforce strictness; it is called 101.8 M times and used in 60% of the packages we inspected. In summary, manual inspection suggests that we overestimate strictness. Improving precision of the analysis would require increasing code coverage. We also found numerous occasions where programmers require strictness to control side effects.

7.3 Meta-programming

The next research question pertains to the use of call-by-need to enable meta-programming.

RQ3: *How frequently are promises used for meta-programming?*

For the purposes of this discussion we define meta-programming as the manipulation of code through calls to `substitute` which lets programmers extract an abstract syntax tree from the body of a promise, modify it, and evaluate with `eval`. We observed 1.7 B calls (2% of all calls) to `substitute`. Figure 18 shows the number of promises that were meta-programmed. The graph has four categories: promises that were created but never used, promises that were meta-programmed, promises that were both meta-programmed and accessed, and lastly, promises that were only accessed. The data shows that 0.5% of promises were used purely for meta-programming purposes, while 0.2% were both forced to obtain a value and used for meta-programming.

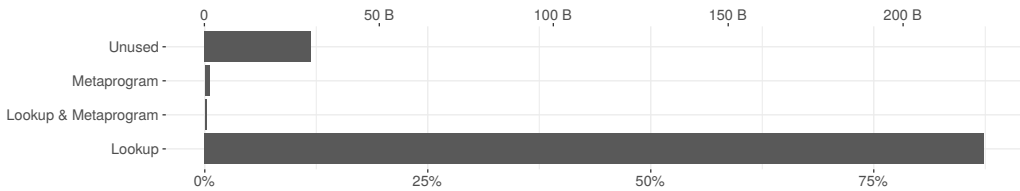


Fig. 18. Meta-programmed promises

Meta-programming is widespread with 2 K (11.9%) packages using it. One feature of `substitute` is that programmers can specify the environment in which to resolve names occurring in its argument. It is also possible to access an environment up the call chain and invoke `substitute` on it. Over 99% of calls to `substitute` use the current environment. Only 0.7% of calls use a custom replacement list or a new environment, and 209.9 K use a parent frame. These are almost entirely due to `deparse`, `eval` and `do.call` which allow specifying their arguments' evaluation environment. For example, the `envnames::get_env_names` function uses `substitute` in different contexts to extract user-defined names of environments. The first call to `substitute` runs in the second frame from the top of the stack, the next call in the first frame, and the third call in the environment in which it is called.

```
get_env_names <- function(envir=NULL, include_functions=FALSE) {
  get_informative_environment_name <- function(envir)
    if (...) envir_name <- deparse(substitute(envir, parent.frame(n = 2)))
    else   envir_name <- deparse(substitute(envir, parent.frame(n = 1)))
  if (!is.null(envir) && !is.environment(envir))
    { error_NotValidEnvironment(deparse(substitute(envir))); return(NULL) }
  envir_name <- get_informative_environment_name(envir)
  ...
}
```

The `glmTMB::makeOp` function constructs ASTs from replacement lists instead of an environment. Arguments `x`, `y` and `op` are evaluated and bound to `X`, `Y` and `op` respectively in the `list` which is then used for replacement in the AST by `substitute`.

```
makeOp <- function(x, y, op=NULL) {
  if (is.null(op) || missing(y)) {
    if (is.null(op)) substitute(OP(X), list(X = x, OP = y))
    else substitute(OP(X), list(X = x, OP = op))
  } else substitute(OP(X, Y), list(X = x, OP = op, Y = y))
}
```

Qualitative analysis. We manually inspected 100 functions that meta-program their arguments, and classified them based on the usage patterns. One common pattern is to extract the source text of an argument. This is used by various plotting functions to give default names to the axes of a graph if none are provided. In the following definition `xAxis` and `yAxis` are parameters used that way. The call to `substitute` returns the AST of the arguments, and `deparse` turns those into text.

```
function(design, xAxis, yAxis) {
  designName <- deparse(substitute(design))
  xAxisName <- deparse(substitute(xAxis))
  yAxisName <- deparse(substitute(yAxis))
  plot(1, type = "n", main = designName, xlab = xAxisName, ylab = yAxisName,
  ...
}
```

The following pattern explains why many promises are both evaluated and meta-programmed. The call to `substitute` extracts the AST for `deg` and the next line evaluates `deg`.

```
function(deg) {
  degname <- deparse(substitute(deg))
  deg <- as.integer(deg)
  if (deg < 0 || deg > 1) stop(paste0("Error_", degname))
  deg
}
```

Another common pattern, that is a syntactic convenience, is to allow the use of symbols instead of strings. The `::` operator is used to prefix function names with their packages. It is implemented as a reflective function that expects two strings. But programmers would rather write `base::log` than `"base"::"log"`. To this end, the arguments are left uninterpreted, instead the function `deparse`s them to strings.

```
`::` <- function(pkg, name) {
  pkg <- as.character(substitute(pkg))
  name <- as.character(substitute(name))
  get(name, envir=asNamespace(pkg), inherits=FALSE)
}
```

Meta-programming is also used for better error reporting and logging. This example shows code that only retrieves the source text of the argument.

```
function(arg)
  if (!is.numeric(arg)) stop(paste(deparse(substitute(arg)), "is_not_numeric"))
```

We also found functions that leverage non-standard evaluation. The following definition is for `base::local` which provides limited form of sandboxing by evaluating code in a new environment. The argument is extracted and evaluated using `eval` in an empty or user-supplied environment.

```
function(arg, envir=new.env()) eval.parent(substitute(eval(quote(arg),envir)))
```

A combination of meta-programming, dynamic evaluation and first-class environments opens up the door for domain specific languages. The pipe operator heavily used in the tidyverse group of packages performs non-standard evaluation on its arguments. While the user writes code like this `df %>% mean`, what is actually executed is `mean(df)`. While the actual definition relies on more intricate non-standard evaluation techniques, a simple definition that achieves a similar effect turns both arguments into abstract syntax trees, and captures the calling environment.

```
function(lhs, rhs) {  
  lhs <- substitute(lhs)  
  rhs <- substitute(rhs)  
  eval(call(pipe, rhs, lhs), parent.frame(), parent.frame())  
}
```

Overall, the use of meta-programming is widespread and falls in two rough categories: access to the source text of an argument in the direct caller and non-standard evaluation of an argument. The latter is the source of much of the expressive power of the language and is critical to some of the most widely used packages such as `ggplot` and `dplyr`.

7.4 Revisiting the Traditional Benefits of Laziness

The next research question asks whether the benefits of lazy evaluation that were advocated by Hudak [1989] are realized in the R ecosystem.

RQ4: *Are the traditional benefits of laziness realized in R?*

These benefits are that programmers need not worry about the cost of unused arguments and they are able to define and use unbounded data structures. We posit the following hypothesis, if programmers understand how call-by-need works, they will feel free to pass complex computations in non-strict positions. If true, one could hope to observe a difference in running time for arguments known to be strict (promises passed to *Always* parameters) and those that are not (promises passed to *Sometimes* parameters). Figure 19 shows the probability density of promise running times with times smaller than a millisecond discarded. The promises that are passed to *Sometimes* arguments

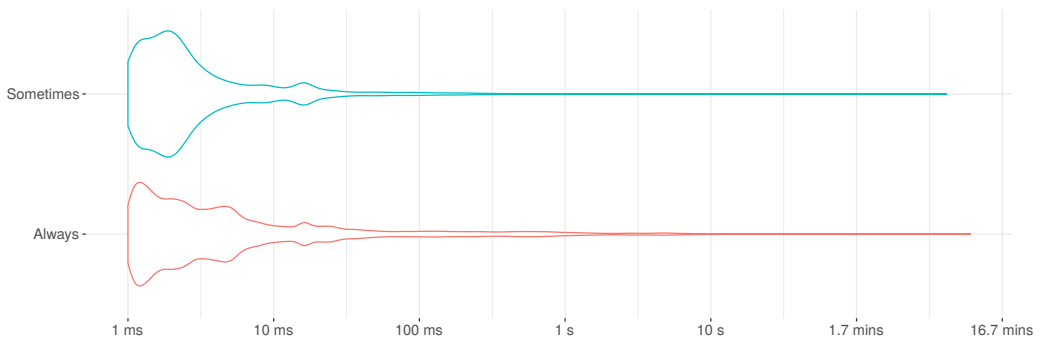


Fig. 19. Promise evaluation

tend to be less expensive to evaluate. While there is a difference in the profiles, the data does not allow us to confirm that programmers are taking advantage of laziness.

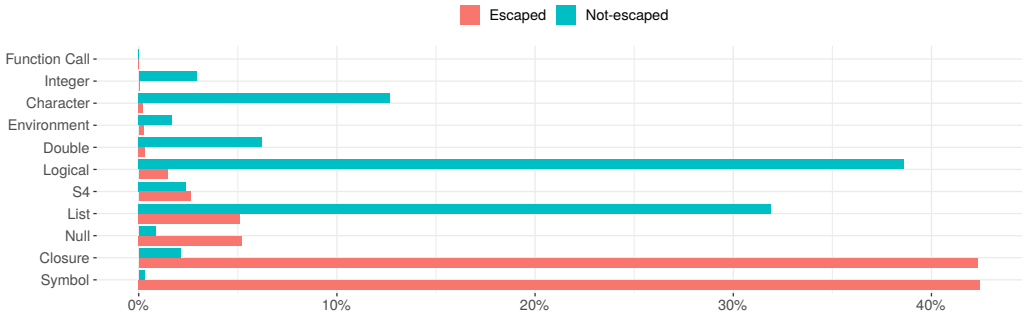


Fig. 20. Function results

Laziness makes it possible to compute over infinite data structures. While R does not provide such data structures, it is conceivable that programmers created some in their code. As we have shown, one can use promises together with environments to create unbounded data structures. While it is difficult to measure this directly, we can measure *escaped promises*. These promises outlive the function they are passed into, and these would be a superset of lazy data structures. Of the 261 B promises we have observed, only 11.6 M escape. This is a rather small number. We need to establish the reason why they escape. Figure 20 compares the return types of functions which have at least one of their promises escaping, and functions that do not have escaping promises. The main difference between them is that functions with escaping promises have a large number of symbols and closures as their return values. The next section performs a qualitative analysis to understand how those closures are used.

Qualitative analysis. It is not easy to assess, from the quantitative results alone, whether programmers benefit from laziness. Consider a call, $f(a+b, c)$, and imagine that depending on the value of c , the first argument may or may not be evaluated. If a and b are large matrices and c is infrequently true, a programmer aware of laziness would not worry about the performance of this code. Without laziness, the API of the function would likely have to change so that instead of passing the result of the computation one would pass the individual arguments and let the function perform the addition if needed. In our manual inspection, we have not found any code suggesting that programmers are concerned about the cost of evaluating expressions, but this is most likely due to the fact that many users are not performance sensitive.

One promising use of laziness is related to `delayedAssign`. We observed 82 M calls to this function. We inspected manually 36 packages that use it. We found several recurring patterns that aim to avoid unnecessary computation. Examples are: the `AzureML` package uses delayed assignment to avoid loading unneeded parts of its workspace; `crunch` uses it to delay fetching data from a server; and (slightly surprisingly) `callCC` uses it in conjunction with non-local return to implement the `call/cc` function. Overall we found little evidence of programmers taking advantage of call-by-need, other than in cases where they explicitly called `delayedAssign`, in the sample of functions we inspected. We did find cases where the authors of the code seemed to want to enforce a consistent evaluation order and prevent argument-induced side effects from happening in the midst of evaluation of the function.

Closure:

```
function(arg, e=10^-5) { function(x) (arg(x+e)-arg(x-e))/(2*e) }
```

S4 object:

```
function(arg) new("FLXcomponent",df=arg$df)
```

Environment:

```
function(arg) { env<-new.env(); env$fn<-function(x){...out<-arg(x)...}; env }
```

Finalizer:

```
function(arg) reg.finalizer(environment(), function(...) dbDisconnect(arg))
```

Delayed assignment:

```
function(arg, e) delayedAssign(x, get(from, arg), assign.env=as.environment(e))
```

Formula:

```
function(arg, i) as.formula(arg[, 1] ~ arg[, i])
```

Fig. 21. Escaping promises

To detect an infinite data structure we looked at occurrences of promises that outlive the function in which they were passed. We inspected 100 functions with escaping arguments and observed the following patterns: (1) arguments captured in closures, (2) arguments captured in S4 objects, (3) arguments stored in environments, (4) arguments passed into finalizers, (7) argument passed into delayed assignments, and (8) arguments passed into formulas. Figure 21 gives examples of each of these categories. In terms of linguistic mechanisms, all but the last two end up as variants of closure-captured promises. Formula is interesting, because it is really a domain specific language that is interpreted with different semantics. In our time spent working with R, we found a single package, *Rstackdeque* [O’Neil 2015] that advertised the use of lazy data structures, specifically fully persistent queues based on [Okasaki 1995]. This package, which depends on lazy lists, is the only use of lazy data structures we are aware of in the R ecosystem.

8 RELATED WORK

Lazy functional programming languages have a rich history. The earliest lazy programming language was Algol 60 [Backus and al. 1963] which had a call-by-name evaluation strategy. This was followed by a series of purely functional lazy languages [Augustsson 1993; Turner 1979, 1985]. The motivations for the pursuit of laziness were modularity, referential transparency and the ability to work with infinite data structures [Hughes 1989]. These languages inspired the design of Haskell [Hudak et al. 2007].

The meta-programming support of R is reminiscent of fexprs [Wand 1998] in Lisp. Fexprs are first class functions with unevaluated arguments. In R, functions always have access to their unevaluated and evaluated arguments. Pitman [Pitman 1980] argued in favor of macros over fexprs. Macros are transparent, their definition can be understood by expanding them to primitive language forms before the evaluation phase. fexprs on the other hand perform code manipulation during evaluation. This makes it harder for compilers to statically optimize fexprs. Furthermore, expression manipulation such as substitution of an expression for all evaluable occurrences of some other expression can be performed correctly by macros because they expand before evaluation to primitive forms.

Purdue FastR [Kalibera et al. 2014] is an AST interpreter for R written in Java to explore the applicability of simple compiler optimization techniques, within the reach of scientific community

lacking expertise in language run-times. The authors implement an optimization technique that defers element wise operations on vectors by constructing expression trees called Views, which are evaluated on demand. This prevents the materialization of temporary vectors in a chain of vectorized mathematical operations. Like promises, views cache the result of evaluating the expression. However, unlike promises which are exposed to the user through meta-programming, views are completely transparent to the user. Promises are built by packaging arbitrary argument expressions but views are built incrementally by piling referentially transparent vector operations such as `+`, `-`, `log`, `ceil`, etc. Promises are evaluated very quickly due to the eager nature of most functions, but the expression trees of views are evaluated only when the entire result vector or its subset is demanded or a selected aggregate operation such as `sum` is applied.

Building upon the implicit argument quoting of promises is a data structure called quosure, short for quoted closure, that bundles an expression and its evaluation environment for explicit manipulation at the language level. A quosure is thus an explicit promise object exposed to the user, with APIs to access the underlying expression and environment. Quosures are a central component of a collection of R packages for data manipulation, Tidyverse [Wickham 2017], that have a common design language and underlying data structures. Dplyr [Wickham et al. 2018], a package of Tidyverse, implements a DSL for performing SQL like data transformations on tabular data and ggplot2 [Wickham 2016] implements a declarative language for graphing data, inspired by the Grammar of Graphics. These packages quote, unquote and quasiquote user supplied expressions and evaluate them in appropriate environments. To facilitate this, these packages also provide an evaluation function, `eval_tidy` that extends the base `eval` to deal with quosures. This suggests that reifying promises can be useful.

Renjin is an implementation of R built on the Java virtual machine designed to analyze large data sets and facilitate integration with enterprise systems. Renjin supports delaying evaluation of side effect free computation [Mühleisen et al. 2018]. Instead of returning the actual result of a computation, Renjin returns placeholders which look and behave exactly like the actual computation result, but will only calculate results if forced to. The difference between Renjin and FastR, both systems are more “lazy” than GNU R, lies in Renjin’s support for relational-style optimization.

Morandat et al. [2012] implemented a tool called TraceR for profiling R programs. The architecture of TraceR was similar to that of the pipeline presented here, but it did not target large scale data collection and has gone unmaintained for several years. Our current infrastructure is less invasive than the previous implementation and is being considered for inclusion in GNUR.

Finally, we compare our semantics to the work of Bodin et al. [2018]. Our semantics makes no claims of being correct (there is no specification of R) or of being faithful to the language. The semantics is useful in as much it provides a readable account of delayed evaluation in R. Bodin’s work is more ambitious, it aims to provide an executable semantics. The benefits of executable semantics is that they can be tested against an implementation, in this case the GNU R virtual machine. The semantics consists of 28,026 lines of Coq and 1,689 lines of ML. Validation is done through testing and visual comparison between the GNU R’s C code and Coq code. Unfortunately, in the current state Bodin’s specification is still far from complete. Out of 20,976 tests, only 6,370 pass. Inspection of the specification reveals that key functions for laziness such as `force`, `forceAndCall`, and `delayedAssign` are not implemented. Only a handful of the provided tests deal with lazy evaluation (they check that promises are evaluated only when forced). Furthermore, package loading and interaction with C code is not supported, thus packages from our corpus cannot be tested. We tried to match our semantics to theirs but the DLS’18 paper does not describe their treatment of laziness. Due to the size of the Coq codebase and lack of documentation, it was unclear how to align the two artifacts.

9 CONCLUSION

This paper offers a glimpse into the design, implementation and usage of call-by-need in the R programming language. Call-by-need is the default in R, but our data suggests that it is used less than one would expect. To deal with side effects and manage programmers' expectation, many functions are stricter than they need to be. We found little evidence of lazy data structures or that users leverage lazy evaluation to avoid unnecessary computation. We found only two broad categories of usages that benefited from it. The first is the creation of delayed bindings. These, in our experience, are always explicit. The second is for meta-programming. Within that category, uses are split between accessing the source text of an expression for debugging purposes and performing non-standard evaluation.

The costs of lazy evaluation in performance and memory use are substantial. Every argument to a function must be boxed in a promise, retaining a reference to the function's environment until evaluated. Every access to a variable must check if it is bound to a promise and either evaluate it or read the cached value. Lazy evaluation complicates the task of compilers and program analysis tools as they must deal with the possibility of any variable access causing side effects. Lastly, the majority of users do not expect arguments to be evaluated in a lazy fashion, thus leading to hard to understand bugs.

If laziness is mostly unused, could it be eliminated? Any change to the semantics of a widely used language has to be minimally invasive. We are considering the following combination of ideas. For the meta-programming use-cases that require only source code we propose to offer a function which returns the caller expression for any argument, this is possible as the information is present in the debug meta-data of the interpreter. For functions that do non-standard evaluation, we propose to add annotations on their parameters to request a promise to be generated. This can be implemented by adding a run-time check before function calls.

While we propose removing laziness, there is also an argument for strengthening it. In many ways, R is only *weakly* lazy, it forces promises in many places where other languages would not. The works of Wickham [2017], Mühleisen et al. [2018] and Kalibera et al. [2014] suggest that more laziness can bring interesting optimization opportunities, especially when performing operation on large data objects.

ACKNOWLEDGMENTS

We thank the reviewers for constructive comments that helped us improve the presentation. Early prototypes of our analysis were implemented by Konrad Siek and Jan Noha, we thank them for their contribution. We would also like to thank Ben Chung, Artem Pelenitsyn, Filip Křikava, Tomas Kalibera, Luke Tierney, Stepan Sindelar, Alex Bertram, and Hadley Wickham for their comments and encouragement. This work received funding from the Office of Naval Research (ONR) award 503353, the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement 695412), the NSF (awards 1518844, 1544542, and 1617892), and the Czech Ministry of Education, Youth and Sports (grant agreement CZ.02.1.010.00.015_0030000421).

REFERENCES

- Lennart Augustsson. 1993. The Interactive Lazy ML System. *J. Funct. Program.* 3, 1. <https://doi.org/10.1017/S0956796800000617>
- John W. Backus and *al.* 1963. Revised Report on the Algorithm Language ALGOL 60. *Commun. ACM* 6, 1. <https://doi.org/10.1145/366193.366201>
- Richard A. Becker, John M. Chambers, and Allan R. Wilks. 1988. *The New S Language*. Chapman & Hall.
- Martin Bodin, Tomás Diaz, and Éric Tanter. 2018. A trustworthy mechanized formalization of R. In *Symposium on Dynamic Languages (DLS)*. <https://doi.org/10.1145/3276945.3276946>
- Olivier Flückiger, Guido Chari, Jan Ječmen, Ming-Ho Yee, Jakob Hain, and Jan Vitek. 2019. R Melts Brains: An IR for First-Class Environments and Lazy Effectful Arguments. In *Dynamic Language Symposium (DLS)*. <https://doi.org/10.1145/3359619.3359744>
- Robert Gentleman, Vince J. Carey, and *al.* 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5. <https://doi.org/10.1186/gb-2004-5-10-r80>
- Paul Hudak. 1989. Conception, Evolution, and Application of Functional Programming Languages. *ACM Comput. Surv.* 21, 3. <https://doi.org/10.1145/72551.72554>
- Paul Hudak, John Hughes, Simon L. Peyton Jones, and Philip Wadler. 2007. A history of Haskell: being lazy with class. In *History of Programming Languages Conference (HOPL-III)*. <https://doi.org/10.1145/1238844.1238856>
- John Hughes. 1989. Why Functional Programming Matters. *Comput. J.* 32, 2. <https://doi.org/10.1093/comjnl/32.2.98>
- Ross Ihaka and Robert Gentleman. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5, 3. <http://www.amstat.org/publications/jcgs/>
- Tomas Kalibera, Petr Maj, Floreal Morandat, and Jan Vitek. 2014. A Fast Abstract Syntax Tree Interpreter for R. In *Conference on Virtual Execution Environments (VEE)*. <https://doi.org/10.1145/2576195.2576205>
- Filip Křikava and Jan Vitek. 2018. Tests from traces: automated unit test extraction for R. In *International Symposium on Software Testing and Analysis (ISSTA)*. <https://doi.org/10.1145/3213846.3213863>
- Uwe Ligges. 2017. 20 Years of CRAN (Video on Channel9). In *UseR! Conference*.
- Floreal Morandat, Brandon Hill, Leo Osvald, and Jan Vitek. 2012. Evaluating the Design of the R Language: Objects and Functions for Data Analysis. In *European Conference on Object-Oriented Programming (ECOOP)*. https://doi.org/10.1007/978-3-642-31057-7_6
- Hannes Mühleisen, Alexander Bertram, and Maarten-Jan Kallen. 2018. Database-Inspired Optimizations for Statistical Analysis. *Journal of Statistical Software* 87, 4. <https://doi.org/10.18637/jss.v087.i04>
- Chris Okasaki. 1995. Simple and efficient purely functional queues and dequeues. *Journal of Functional Programming* 5, 4. <https://doi.org/10.1017/S0956796800001489>
- Shawn T. O'Neil. 2015. Implementing Persistent O(1) Stacks and Queues in R. *The R Journal* 7, Issue 1. <https://doi.org/10.32614/RJ-2015-009>
- Kent M. Pitman. 1980. Special Forms in LISP. In *LISP Conference*. <https://doi.org/10.1145/800087.802804>
- David Smith. 2011. The R Ecosystem. In *The R User Conference 2011*.
- Luke Tierney. 2019. A Byte Code Compiler for R. www.stat.uiowa.edu/~luke/R/compiler/compiler.pdf
- David A. Turner. 1979. A New Implementation Technique for Applicative Languages. *Softw., Pract. Exper.* 9, 1. <https://doi.org/10.1002/spe.4380090105>
- David A. Turner. 1985. Miranda: A Non-Strict Functional language with Polymorphic Types. In *Functional Programming Languages and Computer Architecture (FPCA)*. https://doi.org/10.1007/3-540-15975-4_26
- Mitchell Wand. 1998. The Theory of Fexprs is Trivial. *Lisp and Symbolic Computation* 10, 3. <https://doi.org/10.1023/A:100772063>
- Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. <http://ggplot2.org>
- Hadley Wickham. 2017. *tidyverse: Easily Install and Load the 'Tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>
- Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Müller. 2018. *dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>
- Andrew K. Wright and Matthias Felleisen. 1992. A Syntactic Approach to Type Soundness. *Information and Computation* 115. <https://doi.org/10.1006/inco.1994.1093>