

Synthesizing Random Generators via Higher-Order Representations

Agustín Mista

Chalmers University of Technology
Gothenburg, Sweden
mista@chalmers.se

Alejandro Russo

Chalmers University of Technology
Gothenburg, Sweden
russo@chalmers.se

Abstract—*Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.*

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Integer placerat tristique nisl. Praesent augue. Fusce commodo. Vestibulum convallis, lorem a tempus semper, dui dui euismod elit, vitae placerat urna tortor vitae lacus. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec hendrerit tempor tellus. Donec pretium posuere tellus. Proin quam nisl, tincidunt et, mattis eget, convallis nec, purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla posuere. Donec vitae dolor. Nullam tristique diam non turpis. Cras placerat accumsan nulla. Nullam rutrum. Nam vestibulum accumsan nisl.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat [1].

The main contributions of this paper are:

- We identify two pathological scenarios for which standard type-driven automatic derivation tools fail to synthesize practical random generators, due to a lack of either type structure or domain knowledge (Section II).

- We present a generation technique able to encode stronger properties of the target data by reifying the static information present on the program codebase (Section III).
- We apply and extend the theory of branching processes to analytically predict the average distribution of generated values. Furthermore, we use the predictions to perform simulation-based optimization of the random generation parameters (Section IV).
- We provide an implementation of our ideas in the form of a Haskell library to perform automatic derivation of random generators capable to extract useful structure information from the user source code.

II. TYPE-DRIVEN RANDOM GENERATION

In this section we briefly describe the common technique used for automatically deriving random data generators in Haskell using a type-driven approach. Then, we introduce its main drawbacks by considering two scenarios where this technique gives poor results in practice. We remark that, although this work makes focus on Haskell data types, this technique is general enough to be applied to most programming languages with some level of support for composite types.

Haskell is a strongly typed programming language with a powerful type system. It lets programmers encode a considerable amount of information about the properties of their systems using data types that can be checked at compilation time. One of its key aspects is the support for Algebraic Data Types (ADTs). Essentially, an ADTs is a composite type defined by combining other types. In the most basic conception, types can be combined by **sums** (also known as *variant types* or *tagged disjoint unions*) and **products** (or tuples) of other data types. To exemplify this, consider the following definition of the data type *Exp* encoding integer expressions:

```
data Exp = Val Int
         | Add Exp Exp
         | Mul Exp Exp
```

In the previous definition, we declare *Exp* as the sum of three possible classes of values: *Val* represents literal values, whereas *Add* and *Mul* represent the addition and multiplication of two integer expressions, respectively. In Haskell, *Val*, *Add* and *Mul* are called *data constructors* and are used to

distiguish which variant of the data type we are reffering to. Each data constructor is then defined as a product of zero or more types known as *fields*. In particular, *Val* contains a field of type *Int*, while *Add* and *Mul* contain two fields of type *Exp* representing the operands of each operation. Note that *Exp* is used as a field of at least one of its data constructors (case *Add* and *Mul*), making it a recursively defined ADT. In general, we will say that a data constructor with no recursive fields is *terminal*, and *non-terminal* or *recursive* if it has at least one field of such nature. With this reprentation, the expression “2 + (5 * 6)” can be encoded simply by using *Exp* data constructors as (*Add* (*Val* 2) (*Mul* (*Val* 5) (*Val* 6))). Furthermore, an evaluation function from *Exps* to integer values can be defined very idiomatically:

```
eval :: Exp → Int
eval (Val n)    = n
eval (Add x y)  = eval x + eval y
eval (Mul x y)  = eval x * eval y
```

In the previous definition, *eval* is described using *pattern matching* over each possible kind of value. For the input case of a literal values, we simply return the value contained in the *Val* constructor. On the other hand, if the input value matches either an *Add* or a *Mul* data constructor, where *x* and *y* are *pattern variables* that match any subexpression, we recursively evaluate these subexpressions, combining them with the apropiate operation on each case.

A. Type-Driven Generation of Random Values

In order to perform random testing of a Haskell codebase involving user defined data types, most approaches require the user to provide a random data generator for each one of them. This tends to be a cumbersome and error prone task that closely follows the data type structure. For instance, consider the following definition of a *QuickCheck* random generator for the type *Exp*:

```
genExp :: Gen Exp
genExp = sized gen
  where
    gen 0 = Val ($) arbitrary
    gen n = oneof
      [ Val ($) arbitrary
      , Add ($) gen (n-1) (*) gen (n-1)
      , Mul ($) gen (n-1) (*) gen (n-1) ]
```

This random generator is defined using *QuickCheck*’s combinator *sized* :: (*Int* → *Gen a*) → *Gen a* to paremetrize the generation process up to an external integer number known as the *generation size*. This parameter is used to limit the maximum amount of recursive call that this random generator can perform. When called with a positive generation size (case *gen n*), the generator is free to randomly pick with uniform probability among any data constructor of *Exp*. In the case it picks to generate a *Val* data constructor, it also generates a random *Int* value using the standard overloaded generator *arbitrary* (*QuickCheck* provides standard random generators

for most base types like *Int*, *Bool*, etc.). On the other hand, when it randomly picks to generate either an *Add* or a *Mul* data constructor, it also generates independently two random subexpressions corresponding to the data constructor fields by calling itself recursively (*gen (n-1)*), ensuring to decrease the generation size by a unit on each recursive invocation.

This procedure keeps calling itself recursively until the generation size reaches zero (case *gen 0*), where the generator is constrained to pick only among terminal data constructors, being *Val* the only possible choice in our particular case. Strictly decreasing the the generation size by one on each recursive call results in a useful property over the generated data: every generated value has at most *n* levels, where *n* is the generation size set by the user. This property enables us to model the generation process using the theory of branching processes introduced by **CITATION NEEDED**, and extended in this work as described in Section IV.

Having discussed the previous random generator definition, it easy to understand how to extend this generation procedure to any data type defined in an algebraic fashion. Fortunately, there exists different meta-programming tools to avoid the user from having to mechanically write random generators over and over again for each user-defined ADT. The simplest tool for such purpose is *MegaDeTH*. Given the name of the target data type, it synthesizes a random generator for it that behaves pretty much like the previously defined one. However, picking among different data constructors with uniform probability can lead to a generation process biased to generate (in average) very small values, regardless of the generation sized set by the user **CITATION NEEDED**.

DRAGEN is meta-programming tool conceived to mitigate this problem. Instead of setting a uniform generation probability of data constructors, this tool considers the scenario where each one can be generated following a different (although fixed over time) generation probability. Then, this tool uses the theory of branching processes to analitically predict the average distribution of data constructors generated on each random value. This prediction mechanism is used to feedback a simulation-based optimization process that adjusts the generation probability of each data constructor in order to obtain a particular distribution of values that can be specified by the user, offering a more flexible testing environment while still being mostly automated.

Both *MegaDeTH* and **DRAGEN** synthesize random generators that are theoretically capable to generate the whole space of values of the target data type, provided that we set a generation size suficiently big. However, the limitations arise quickly when we consider that the underlying generation model is essentially the same: they pick a single data constructor and recursively generate each required subexpression independently. In practice, this procedure is too generic to generate complex data that is useful enough to be used for random testing.

In this work we identify two sources of additional structure information which are not considered by the aforementioned automatic derivation tools to obtain random data generators:

- 1) The target code behaves differently over inputs matching very specific patterns.
- 2) The target code yields part of the responsibility of preserving its invariants to the functions of its abstract interface.

This information can be reified in compile time and used to synthesize richer random generators automatically. We proceed to exemplify the previous points in detail.

B. Presence of Complex Pattern Matchings

To exemplify the first problematic scenario that may arise, suppose we want to use randomly generated *Exps* to test a property comprising the following function:

```
foo :: Exp → Exp
foo (Add (Add x (Val 50)) (Add (Val 25) y))
  = error "pattern #1"
foo (Mul (Val 50) (Mul (Val x) y))
  = error "pattern #2"
foo x = x
```

This function behaves very much like an identity function, with the exception that it fails for two very specific patterns of input values defined using nested pattern matching. This is, *foo* does not only match against the root data constructor, but also against the data constructors of its subexpressions and sub-subexpressions.

QuickCheck's default implementation for random *Ints* pick a number in the interval $[-n, n]$ with uniform distribution. Hence, in principle we need to recognize what is a suitable generation size, which should be big enough for our generator to be able to generate the numbers we pattern match against to, otherwise we will not be able to generate any value to match against *foo* clauses, leaving fragments of code completely untested. Suppose then that we pick the a generation size 50, i.e. the minimum size that is big enough to produce an *Int* number equal to 50. Under this consideration, the probability of generating a value matching the first clause of *foo* (and hence triggering the first error) results as follows:

$$\begin{aligned}
P(\text{match}(\text{foo}\#1)) &= P(\text{Add}) \\
&\quad * P(\text{Add}) * 1 * (P(\text{Val}) * (1/100)) \\
&\quad * P(\text{Add}) * (P(\text{Val}) * (1/100)) * 1
\end{aligned}$$

If we use *MegaDeTH* to automatically derive a random generator for *Exp*, we obtain a uniform generation probability distribution over constructors, i.e., $P(\text{Val}) = P(\text{Add}) = P(\text{Mul}) = 1/3$. In this setting, $P(\text{match}(\text{foo}\#1))$ results $1/2430000$, meaning that, in average, we will need to generate over than two million test cases in order to be able to test the first clause of *foo* only once. This situation can be somewhat improved if we decide to use **DRAGEN** to obtain a random generator. By optimizing the generation probabilities in order to benefit the generation of *Add* constructor in a ratio of 20 : 1 with respect to the rest of them, we obtained a distribution of values where $P(\text{match}(\text{foo}\#1)) \approx 1/300000$.

Although this certainly improves the probability of generating a matching value, this probability is not substantial enough to become practical. Additionally, by favoring the generation probabilities towards the *Add* constructor, we found that the probability of generating a value matching the second clause of *foo* (which does not match against it) also gets diminished into an impractical value.

Despite the fact that the previous example might look artificial, branching against specific patterns of the input data is a common task. For instance, the balancing function of a Red-Black tree need to consider a quite specific combination of color, left and right subtrees and sub-subtrees in order to preserve the height invariant. Moreover, Common Subexpression Elimination (CSE) is a compiler optimization that needs to consider very specific sequences of instructions that may be regrouped in a computationally more efficient way, to cite a few.

C. Data Invariants Encoded on Abstract Interfaces

A common choice when implementing a data structure is to transfer the responsibility of preserving its invariants to the functions that manipulates its values. Consider for instance the following possible implementation of a basic *HTML* manipulation library:

```
data Html = Html String
head :: Html → Html
head (Html inner)
  = Html ("<head>" ++ inner ++ "</head>")
body :: Html → Html
body (Html inner)
  = Html ("<body>" ++ inner ++ "</body>")
br :: Html
br = Html "<br />"
concat :: Html → Html → Html
concat (Html x) (Html y) = Html (x ++ y)
```

In the previous definition, the *Html* data type is defined using a single data constructor that contains the textual representation of the *Html* code it represents in plain text. Later, the combinator functions defined over *HTML* are the ones in charge of transforming this plain text representation with the invariant that, given valid *Html* parameters, they always return a valid *Html* value. This is a very common programming pattern that can be found in a variety of Haskell libraries **CITATION NEEDED**.

If we use a standard type-driven approach to derive a random generator for *Html*, the only “structural” information that we can use in the derivation process is that *Html* values are composed of strings. As a consequence, we essentially end up generating random strings, which rarely represents a valid (or at least well structured) *Html* value. Additionally, if our test suite contains properties constrained by certain preconditions, this lack of domain knowledge may lead in an impractically high discard ratio of randomly generated test. For instance,

suposse we write a property to verify that *head* preserves the invariant previously mentioned:

```
prop_head_inv :: HtmL → Bool
prop_head_inv x = valid x ⇒ valid (head x)
```

In the previous definition we state that *head* always returns a valid *HtmL* value (*valid (head x)*), provided that we are given a valid *HtmL* as input (*valid x*). Then, while testing this property we rarely satisfy its precondition, in which case *QuickCheck* discards the whole test, retrying with a diferent random input, degradating this way the testing performance.

We believe that, if certain patterns of values are relevant enough to appear in the codebase being tested, a practical random testing methodology should be able to produce values satisfying such patterns in a substantial proportion. The next section introduces a representation model that let us encode the structure information presented here into our automatically derived random generators in a modular and flexible way.

III. EXTRACTING STRUCTURE

In this work, we ...

```
data HRep_Val a = Mk_Val Int
data HRep_Add a = Mk_Add a a
data HRep_Mul a = Mk_Mul a a
```

For this purpose, we employ Haskell's type classes mechanism **CITATION NEEDED**. We define an evaluation type class (\Downarrow) which specifies how to perform a single transformation *step* from the higher-order representation back into a concrete value of the target:

```
class (rep ⇓ target) where
  step :: rep target → target
```

Then, we can define the following overloaded instances of the *step* operation for the canonical representations of data constructors simply by mapping each lifted constructor back into its corresponding one.

```
instance (HRep_Val ⇓ Exp) where
  step (Mk_Val n) = Val n
instance (HRep_Add ⇓ Exp) where
  step (Mk_Add x y) = Add x y
instance (HRep_Mul ⇓ Exp) where
  step (Mk_Mul x y) = Mul x y
```

With these individual representations, we can define a type combinator (\oplus) to compose two representations into a single one:

```
data (f ⊕ g) a = L (f a) | R (g a)
```

Then, a composite representation can be transformed back into the concrete target type by pattern matching on the data type variant and applying the *step* tranformation to the inner representation:

```
instance (f ⊕ g ⇓ Exp) where
  step (L f) = step f
  step (R g) = step g
```

Furthermore, we can define a type combinator (\otimes) to tag every representation with an explicit generation frequency:

```
data (f ⊗ n) a = Freq (f a)
```

This combinator is evaluated back to our target data type simply by piping the result from the inner representation. It does not change the evaluation semantics, as it is only considered at generation time:

```
instance (f ⊗ n ⇓ Exp) where
  step (Freq f) = step f
```

With the introduced combinators, we can easily create a type synonym HRep_{Exp} to refer to the canonical representation of our original data type *Exp*, tagging for instance the representation of the constructor *Add* to be generated in double the proportion of rest of the representations:

```
type HRep_Exp = HRep_Val
               ⊕ HRep_Add ⊗ 2
               ⊕ HRep_Mul
```

```
data HRep_foo#1 a = Mk_foo#1 a a
data HRep_foo#2 a = Mk_foo#2 Int a
```

```
instance (HRep_foo#1 ⇓ Exp) where
  step (Mk_foo#1 x y)
    = Add (Add x (Val 50)) (Add (Val 25) y)
instance (HRep_foo#2 ⇓ Exp) where
  step (Mk_foo#2 x y)
    = Mul (Val 50) (Mul (Val x) y)
```

```
type HRep_foo = HRep_foo#1
               ⊕ HRep_foo#2
```

```
data HRep_ten    a = Mk_ten
data HRep_square a = Mk_square a
data HRep_minus  a = Mk_minus a a
```

```
instance (HRep_ten ⇓ Exp) where
  step Mk_ten = ten
instance (HRep_square ⇓ Exp) where
  step (Mk_square x) = square x
instance (HRep_minus ⇓ Exp) where
  step (Mk_minus x y) = minus x y
```

```
type HRep_M = HRep_ten
              ⊕ HRep_square
              ⊕ HRep_minus
```

```
type Spec_Exp = HRep_Exp ⊗ 4
               ⊕ HRep_foo ⊗ 2
               ⊕ HRep_M
```

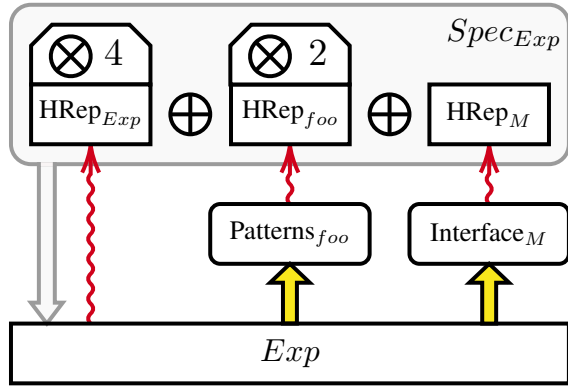


Fig. 1: Higher order representation of the data type Exp , using structural information from the function foo and the abstract interface of the module M .

This previous definition can be interpreted graphically as it is shown in the Figure 1. Curly arrows represent the structural information extracted using meta-programming.

IV. RANDOM GENERATORS SYNTHESIS

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec hendrerit tempor tellus. Donec pretium posuere tellus. Proin quam nisl, tincidunt et, mattis eget, convallis nec, purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla posuere. Donec vitae dolor. Nullam tristique diam non turpis. Cras placerat accumsan nulla. Nullam rutrum. Nam vestibulum accumsan nisl.

V. RELATED WORK

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim

bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

VI. FINAL REMARKS

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

REFERENCES

- [1] G. Grieco, M. Ceresa, A. Mista, and P. Buiras, “QuickFuzz testing for fun and profit,” *Journal of Systems and Software*, vol. 134, no. Supp. C, 2017.