Advances in Property-Based Testing for Prolog

Citation for published version:

Digital Object Identifier (DOI):
10.1007/978-3-319-41135-4_3

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published in:
Tests and Proofs

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Advances in Property-Based Testing for αProlog

James Cheney¹, Alberto Momigliano², Matteo Pessina²

¹ University of Edinburgh jcheney@inf.ed.ac.uk
² Università degli Studi di Milano momigliano@di.unimi.it, matteo.pessina3@studenti.unimi.it

Abstract. αCheck is a light-weight property-based testing tool built on top of αProlog, a logic programming language based on nominal logic. αProlog is particularly suited to the validation of the meta-theory of formal systems, for example correctness of compiler translations involving name-binding, alpha-equivalence and capture-avoiding substitution. In this paper we describe an alternative to the negation elimination algorithm underlying αCheck that substantially improves its effectiveness. To substantiate this claim we compare the checker performances w.r.t. two of its main competitors in the logical framework niche, namely the QuickCheck/Nitpick combination offered by Isabelle/HOL and the random testing facility in PLT-Redex.

1 Introduction

Formal compiler verification has come a long way from McCarthy and Painter’s “Correctness of a Compiler for Arithmetic Expression” (1967), as witnessed by the success of CompCert and subsequent projects [21,35]. However outstanding these achievements are, they are not a magic wand for every-day compiler writers: not only CompCert was designed with verification in mind, whereby the implementation and the verification were a single process, but there are only a few dozen people in the world able and willing to carry out such an endeavour. By verification, CompCert means the preservation of certain simulation relations between source, intermediate and target code; however, the translations involved are relatively simple compared to those employed by modern optimizing compilers. Despite some initial work [1,7], handling more realistic optimizations seems even harder, e.g. the verification of the call arity analysis and transformation in the Glasgow Haskell Compiler (GHC):

“The [Nominal] Isabelle development corresponding to this paper, including the definition of the syntax and the semantics, contains roughly 12,000 lines of code with 1,200 lemmas (many small, some large) in 75 theories, created over the course of 9 months” (page 11, [7]).

For the rest of us, hence, it is back to compiler testing, which is basically synonymous with passing a hand-written fixed validation suite. This is not completely satisfactory, as the coverage of those tests is difficult to assess and because, being fixed, these suites will not uncover new bugs. In the last few years, randomized differential testing [24] has been suggested in combination with automatic generation of (expressive) test programs, most notably for C compilers with the Csmith
tool [36] and to a lesser extent for GHC [30]. The oracle is comparison checking: Csmith feeds randomly generated programs to several compilers and flags the minority one(s), that is, those reporting different outputs from the majority of the other compilers under test, as incorrect. Similarly, the outcome of GHC on a random program with or without an optimization enabled is compared.

Property-based testing, as pioneered by QuickCheck [12], seems to leverage the automatic generation of test cases with the use of logical specifications (the properties), making validation possible not only in a differential way, but internally, w.r.t. (an abstraction of) the behavior of the source and intermediate code. In fact, compiler verification/validation is a prominent example of the more general field of verification of the meta-theory of formal systems. For many classes of (typically) shallow bugs, a tool that automatically finds counterexamples can be surprisingly effective and can complement formal proof attempts by warning when the property we wish to prove has easily-found counterexamples. The beauty of such meta-theory model checking is that, compared to other general forms of system validation, the properties that should hold are already given by means of the theorems that the calculus under study is supposed to satisfy. Of course, those need to be fine tuned for testing to be effective, but we are mostly free of the thorny issue of specification/invariant generation.

In fact, such tools are now gaining traction in the field of semantics engineering, see in particular the QuickCheck/Nitpick combination offered in Isabelle/HOL [4] and random testing in PLT-Redex [18]. However, a particular dimension to validating for example optimizations in a compiler such as GHC, whose intermediate language is a variant of the polymorphically typed λ-calculus, is a correct, simple and effective handling of binding signatures and associated notions such as α-equivalence and capture avoiding substitutions. A small but not insignificant part of the success of the CompCert project is due to not having to deal with any notion of binder [3]. The ability to encode possibly non-algorithmic relations (such as typing) in a declarative way would also be a plus.

The nominal logic programming language αProlog [11] offers all those facilities. Additionally, it was among the first to propose a form of property based testing for language specifications with the αCheck tool [9]. In contrast to QuickCheck/Nitpick and PLT Redex, our approach supports binding syntax directly and uses logic programming to perform exhaustive symbolic search for counterexamples. Systems lacking this kind of support may end up with ineffective testing capabilities or requiring an additional amount of coding, which needs to be duplicated in every case study:

Redex offers little support for handling binding constructs in object languages. It provides a generic function for obtaining a fresh variable, but no help in defining capture-avoiding substitution or α-equivalence [. . .] In one case [. . .] managing binders constitutes a significant portion of the overall time spent [. . .] Generators derived from grammars [. . .] require substantial massaging to

Xavier Leroy, personal communication. In fact, the encoding of the λ-calculus in [22] does not respect α-equivalence, nor does it implement substitutions in a capture avoiding way.
achieve high test coverage. This deficiency is particularly pressing in the case of typed object languages, where the massaging code almost duplicates the specification of the type system" (page 5, [18]).

αCheck extends αProlog with tools for searching for counterexamples, that is, substitutions that makes the antecedent of a specification true and the conclusion false. In logic programming terms this means fixing a notion of negation. To begin with, αCheck adopted the infamous negation-as-failure (NF) operation, “which put pains thousandfold upon the” logic programmers. As many good things in life, its conceptual simplicity and efficiency is marred by significant problems:

– the lack of an agreed intended semantics against which to carry a soundness proof: this concern is significant because the semantics of negation as failure has not yet been investigated for nominal logic programming;
– even assuming such a semantics, we know that NF is unsound for non-ground goals; hence all free variables must be instantiated before solving the negated conclusion. This is obviously exponentially expensive in an exhaustive search setting and may prevent optimizations by goal reordering.

To remedy this αCheck also offered negation elimination (NE) [3,26], a source-to-source transformation that replaces negated subgoals to calls to equivalent positively defined predicates. NE by-passes the previous issues arising for NF since, in the absence of local (existential) variables, it yields an ordinary (α)Prolog program, whose intended model is included in the complement of the model of the source program. In particular, it avoids the expensive term generation step needed for NF, it has been proved correct, and it may open up other opportunities for optimization. Unfortunately, in the experiments reported in our initial implementation of αCheck [9], NE turned out to be slower than NF.

Perhaps to the reader’s chagrin, this paper does not tackle the validation of compiler optimizations (yet). Rather, it lays the foundations by:

1. describing an alternative implementation of negation elimination, dubbed NEs—“s” for simplified: this improves significantly over the performance of NE as described in [9] by producing negative programs that are equivalent, but much more succinct, so much as to make the method competitive w.r.t. NF;
2. and by evaluating our checker in comparison with some of its competitors in the logical framework niche, namely QuickCheck/Nitpick [4] and PLT-Redex [18]. To the best of our knowledge, this is the first time any of these three tools have been compared experimentally.

In the next section we give a tutorial presentation of the tool and move then to the formal description of the logical engine (Section 3). In Section 4 we detail the NEs algorithm and its implementation, whereas Section 5 carries out the promised comparison on two case studies, a prototypical λ-calculus with lists and a basic type system for secure information flow. The sources for αProlog and αCheck can be found at https://github.com/aprolog-lang/aprolog. Supplementary material, including the full listing of the case studies presented here and an online appendix containing additional experiments and some formal notions
used in Section 3 but omitted here for the sake of space, are available at [10]. We assume some familiarity with logic programming.

2 A Brief Tour of $\alpha$Check

We specify the formal systems and the properties we wish to check as Horn logic programs in $\alpha$Prolog [11], a logic programming language based on nominal logic, a first-order theory axiomatizing names and name-binding introduced by Pitts [32].

In $\alpha$Prolog, there are several built-in types, functions, and relations with special behavior. There are distinguished name types that are populated with infinitely many name constants. In program text, a lower-case identifier is considered to be a name constant by default if it has not already been declared as something else. Names can be used in abstractions, written $\lambda x . M$ in programs, considered equal up to $\alpha$-renaming of the bound name. Thus, where one writes $\lambda x . M$, $\forall x . M$, etc. in a paper exposition, in $\alpha$Prolog one writes $\text{lam}(x \langle M \rangle)$, $\text{forall}(x \langle M \rangle)$, etc. In addition, the freshness relation $a \# t$ holds between a name $a$ and a term $t$ that does not contain a free occurrence of $a$. Thus, $x \notin FV(t)$ is written in $\alpha$Prolog as $x \# t$; in particular, if $t$ is also a name then freshness is name-inequality. For convenience, $\alpha$Prolog provides a function-definition syntax, but this is just translated to an equivalent (but more verbose) relational implementation via flattening.

Horn logic programs over these operations suffice to define a wide variety of object languages, type systems, and operational semantics in a convenient way. To give a feel of the interaction with the checker, here we encode a simply-typed $\lambda$-calculus augmented with constructors for integers and lists, following the PLT-Redex benchmark sltk.lists.rkt from http://docs.racket-lang.org/redex/benchmark.html, which we will examine more deeply in Section 5.1. The language is formally declared as follows:

```
Types  A, B ::= int | ilist | A -> B
Terms   M ::= x | $\lambda x : A. M$ | M1 M2 | c | err
Constants c ::= n | nil | cons | hd | tl
Values  V ::= c | $\lambda x : A. M$ | cons V | cons V1 V2
```

We start (see the top of Figure 1) by declaring the syntax of terms, constants and types, while we carve out values via an appropriate predicate. A similar predicate is_err characterizes the threading in the operational semantics of the err expression, used to model run time errors such as taking the head of an empty list.

We follow this up (see the remainder of Figure 1) with the static semantics (predicate tc) and dynamic semantics (one-step reduction predicate step), where we omit the judgments for the value predicate and subst function, which are analogous to the ones in [9]. Note that err has any type and constants are typed via a table tcf, also omitted.

Horn clauses can also be used as specifications of desired program properties of such an encoding, including basic lemmas concerning substitution as well as main theorems such as preservation, progress, and type soundness. This is realized via checking directives.
ty: type.
intTy: ty. funTy: (ty,ty) -> ty. listTy: ty.
cst: type.
id: name_type.
exp: type.
var: id -> exp. lam: (id\exp,ty) -> exp. app: (exp,exp) -> exp.
c: cst -> exp. err: exp.

type ctx = [(id,ty)].
pred tc (ctx,exp,ty).
tc(_,err,T). tc(_,c(C),T) :- tcf(C) = T.
tc([(X,T)|G],var(X),T) :- X # Y, tc(G,var(X),T).
tc(G,app(M,N),U) :- tc(G,M,funTy(T,U)), tc(G,N,T).
tc(G,lam(x\M,T),funTy(T,U)) :- x # G, tc([(x,T)|G],M,U).
pred step(exp,exp).
step(app(c(hd),app(app(c(cons),V),VS)),V) :- value(V), value(VS).
step(app(c(tl),app(app(c(cons),V),VS)),V) :- value(V), value(VS).
step(app(lam(x\M,T),V), subst(M,x,V)) :- value(V).
step(app(M1,M2),app(M1',M2)) :- step(M1,M1').
step(app(V1,M2),app(M1,M2')) :- value(V1), step(M2,M2').
pred is_err(exp).
is_err(err).
is_err(app(c(hd),c(nil)))).
is_err(app(c(tl),c(nil)))).
is_err(app(E1,E2)) :- is_err(E1).
is_err(app(V1,E2)) :- value(V1), is_err(E2).

Fig. 1. Encoding of the example calculus in αProlog

#check "spec" n : H1, ..., Hn => A.

where spec is a label naming the property, n is a parameter that bounds the search space, and H1 through Hn and A are atomic formulas describing the preconditions and conclusion of the property. As with program clauses, the specification formula is implicitly universally quantified. Following the PLT-Redex development, we concentrate here only on checking that that preservation and progress hold.

#check "pres" 7 : tc([],E,T), step(E,E') => tc([],E',T).
#check "prog" 7 : tc([],E,T) => progress(E).

Here, progress is a predicate encoding the property of “being either a value, an error, or able to make a step”. The tool will not find any counterexample, because, well, those properties are (hopefully) true of the given setup. Now, let us insert
a typo that swaps the range and domain types of the function in the application rule, which now reads:

\[
\text{tc}(G, \text{app}(M, N), U) \leftarrow \text{tc}(G, M, \text{funTy}(T, U)), \text{tc}(G, N, U). \% \text{ was funTy}(U, T)
\]

Does any property become false? The checker returns immediately with this counterexample to progress:

\[
E = \text{app}(c(\text{hd}), c(\text{toInt}(N)))
\]
\[
T = \text{intTy}
\]

This is abstract syntax for \(\text{hd n}\), an expression erroneously well-typed and obviously stuck. Preservation meets a similar fate: \((\lambda x: T \to \text{int. } x \text{ err})\ n\) steps to an ill-typed term.

\[
E = \text{app}(\text{lam}(x\\text{app}(\text{var}(x), \text{err}), \text{funTy}(T, \text{intTy})), c(\text{toInt}(N)))
\]
\[
E' = \text{app}(c(\text{toInt}(N)), \text{err})
\]
\[
T = \text{intTy}
\]

3 The Core Language

In this section we give the essential notions concerning the core syntax, to which the surface syntax used in the previous section desugars, and semantics of \(\alpha\)-Prolog programs.

An \(\alpha\)-Prolog signature is composed by sets \(\Sigma_D\) and \(\Sigma_N\) of, respectively, base types \(\delta\), which includes a type \(o\) of propositions, and name types \(\nu\); a collection \(\Sigma_P\) of predicate symbols \(p:\tau \to o\) and one \(\Sigma_F\) of function symbol declarations \(f:\tau \to \delta\). Types \(\tau\) are formed as specified by the following grammar:

\[
\tau ::= \delta \mid \tau \times \tau' \mid 1 \mid \nu \mid \langle \nu \rangle \tau
\]

where \(\delta \in \Sigma_D\) and \(\nu \in \Sigma_N\) and 1 is the unit type. Given a signature, the language of terms is defined over sets \(V = \{X, Y, Z, \ldots\}\) of logical variables and sets \(A = \{a, b, \ldots\}\) of names:

\[
t, u ::= a \mid \pi \cdot X \mid \langle \rangle \mid \langle t, u \rangle \mid \langle a \rangle t \mid f(t)
\]
\[
\pi ::= \text{id} \mid (a \ b) \circ \pi
\]

where \(\pi\) are permutations, which we omit in case \(\text{id} \cdot X\), \(\langle \rangle\) is unit, \(\langle t, u \rangle\) is a pair and \(\langle a \rangle t\) is the abstract syntax for name-abstraction. The result of applying the permutation \(\pi\) (considered as a function) to \(a\) is written \(\pi(a)\). Typing for these terms is standard, with the main novelty being that name-abstractions \(\langle a \rangle t\) have abstraction types \(\langle \nu \rangle \tau\) provided \(a : \nu\) and \(t : \tau\).

The freshness \((s \# \tau, u)\) and equality \((t \approx \tau, u)\) constraints, where \(s\) is a term of some name type \(\nu\), are the new features provided by nominal logic. The former relation is defined on ground terms by the following inference rules, where \(f : \tau \to \delta \in \Sigma_F\):

\[
\begin{align*}
\frac{a \neq b}{a \#_{\nu} b} & \quad \frac{a \#_{\tau} t}{a \#_{\nu} f(t)} & \quad \frac{a \#_{\tau_1} t_1 \ a \#_{\tau_2} t_2}{a \#_{\tau_1 \times \tau_2} \langle t_1, t_2 \rangle} & \quad \frac{a \#_{\langle \nu \rangle \tau} b \ a \#_{\tau} t}{a \#_{\langle \nu \rangle \tau} \langle b \rangle t} & \quad \frac{a \#_{\langle \nu \rangle \tau} \langle a \rangle t}{a \#_{\langle \nu \rangle \tau} \langle a \rangle t}
\end{align*}
\]
In the same way we define the equality relation, which identifies terms modulo α-equivalence, where \((a \ b) \cdot u\) denotes *swapping* two names in a term:

\[
\begin{align*}
\text{a} \approx_r \text{a} & \quad \langle \rangle \approx_1 \langle \rangle \\
\text{t}_1 \approx_r \text{u}_1 & \quad \text{t}_2 \approx_r \text{u}_2 \\
\langle \text{t}_1, \text{t}_2 \rangle \approx_{r_1 \times r_2} \langle \text{u}_1, \text{u}_2 \rangle & \quad f(\text{t}) \approx_s f(\text{u})
\end{align*}
\]

Given a signature, goals \(G\) and program clauses \(D\) have the following form:

\[
\begin{align*}
A & ::= t \approx u \mid t \not\approx u \\
G & ::= \bot \mid \top \mid A \mid p(t) \mid G \land G' \mid G \lor G' \mid \exists X: \tau. G \mid \forall \alpha: \nu. G \mid \forall^* X: \tau. G \\
D & ::= \top \mid p(t) \mid D \land D' \mid G \supset D \mid \forall X: \tau. D \mid \bot \mid D \lor D'
\end{align*}
\]

The productions shown in black yield a fragment of nominal logic called \(\forall\)-goal clauses \([11]\), for which resolution based on nominal unification is sound and complete. This is in contrast to the general case where the more complicated *equivariant unification* problem must be solved \([8]\). We rely on the fact that \(D\) formulas in a program \(\Delta\) can always be normalized to sets of clauses of the form \(\forall X: \tau. G \supset p(t)\), denoted \(\text{def}(p, \Delta)\). The *fresh-name* quantifier \(\forall\), firstly introduced in \([32]\), quantifies over names not occurring in a formula (or in the values of its variables). The extensions shown in red here in the language BNF (and in its proof-theoretic semantics in Figure 2) instead are constructs brought in from the negation elimination procedure (Section 4.1) and which will not appear in any source programs. In particular, an unusual feature is the *extensional* universal quantifier \(\forall^*\) \([15]\). Differently from the *intensional* universal quantifier \(\forall\), for which \(\forall X: \tau. G\) holds if and only if \(G[x/X]\) holds, where \(x\) is an eigenvariable representing any terms of type \(\tau\), \(\forall^* X: \tau. G\) succeeds if and only if \(G[t/X]\) does for every ground term of type \(\tau\).

*Constraints* are \(G\)-formulas of the following form:

\[
C ::= \top \mid t \approx u \mid t \not\approx u \mid C \land C' \mid \exists X: \tau. C \mid \forall \alpha: \nu. C
\]

We write \(K\) for a set of constraints and \(\Gamma\) for a context keeping track of the types of variables and names. Constraint-solving is modeled by the judgment \(\Gamma; K \models C\), which holds if for all maps \(\theta\) from variables in \(\Gamma\) to ground terms if \(\theta \models K\) then \(\theta \models C\). The latter notion of satisfiability is standard, modulo handling of names: for example \(\theta \models \forall \alpha: \nu. C\) iff for some \(b\) fresh for \(\theta\) and \(C, \theta \models C[b/a]\).

We can describe an idealized interpreter for \(\alpha\)Prolog with the “amalgamated” proof-theoretic semantics introduced in \([11]\) and inspired by similar techniques stemming from CLP \([21]\) — see Figure 2 sporting two kind of judgments, goal-directed proof search \(\Gamma; \Delta; K \Rightarrow G\) and focused proof search \(\Gamma; \Delta; K \xrightarrow{D} Q\). This semantics allows us to concentrate on the high-level proof search issues, without requiring to introduce or manage low-level operational details concerning constraint solving. We refer the reader to \([11]\) for more explanation and ways to make those judgments operational. Note that the rule \(\forall^* \omega\) says that goals of
Fig. 2. Proof search semantics of αProlog programs

the form $\forall^* X : \tau . G$ can be proved if $\Gamma ; X : \tau ; \Delta ; K , C \Rightarrow G$ is provable for every constraint $C$ such that $\Gamma ; K \models \exists X : \tau . C$ holds. Since this is hardly practical, the number of candidate constraints $C$ being infinite, we approximate it by modifying the interpreter so as to perform a form of case analysis: at every stage, as dictated by the type of the quantified variable, we can either instantiate $X$ by performing a one-layer type-driven case distinction and further recur to expose the next layer by introducing new $\forall^*$ quantifiers, or we can break the recursion by instantiation with an eigenvariable.

4 Specification Checking

Informally, #check specifications correspond to specification formulas of the form

$$\forall \alpha . \forall X . G \supset A$$ (1)
where $G$ is a goal and $A$ an atomic formula (including equality and freshness constraints). Since the $\forall$-quantifier is self-dual, the negation of (1) is of the form $\forall a. \exists X. G \land -A$. A \textit{(finite) counterexample} is a closed substitution $\theta$ providing values for $X$ such that $\theta(G)$ is derivable, but the conclusion $\theta(A)$ is not. Since we live in a logic programming world, the choice of what we mean by “not holding” is crucial, as we must choose an appropriate notion of \textit{negation}.

In $\alpha$Check the reference implementation reads negation as \textit{finite failure} ($not$):

$$\forall a. \exists X: \tau. G \land gen[\tau](X) \land not(A)$$

(2)

where $gen[\tau]$ are type-indexed predicates that \textit{exhaustively} enumerate the (ground) inhabitants of $\tau$. For example, $gen[ty]$ yields the predicate:

\begin{verbatim}
gen_ty(intTy).
gen_ty(listTy).
gen_ty(funTy(T1,T2)) :- gen_ty(T1), gen_ty(T2).
\end{verbatim}

A check such as (2) can simply be executed as a goal in the $\alpha$Prolog interpreter, using the number of resolution steps permitted to solve each subgoal as a bound on the search space. This method, combined with a complete search strategy such as iterative deepening, will find a counterexample, if one exists. This realization of specification checking is simple and effective, while not escaping the traditional problems associated with such an operational notion of negation.

4.1 Negation Elimination

Negation Elimination [3,26] is a source-to-source transformation that replaces negated subgoals with calls to a combination of equivalent positively defined predicates. In the absence of local (existential) variables, NE yields an ordinary \((\alpha)\)Prolog program, whose intended model is included in the complement of the model of the source program. In other terms, a predicate and its complement are mutually \textit{exclusive}. Exhaustivity, that is whether a program and its complement coincide with the Herbrand base of the program’s signature may or may not hold, depending on the decidability of the predicate in question; nevertheless, this property, though desirable, is neither frequent nor necessary in a model checking context. When local variables are present, the derived positivized program features the \textit{extensional} universal quantifier presented in the previous section.

The generation of complementary predicates can be split into two phases: \textit{term complementation} and \textit{clause complementation}.

\textit{Term complementation} A cause of atomic goal failure is when its arguments do not unify with any of the program clause heads in its definition. The idea is then to generate the complement of the term structure in each clause head by constructing a set of terms that differ in at least one position. However, and similarly to the higher-order logic case, the complement of a nominal term containing free or bound names cannot be represented by a \textit{finite} set of nominal terms. For our application nonetheless, we can pre-process clauses so that the standard complementation algorithm for (linear) first order terms applies [19]. This forces terms
Clause complementation The idea of the clause complementation algorithm is to compute the complement of each head of a predicate definition using term complementation, while clause bodies are negated pushing negation inwards until atoms are reached and replaced by their complement and the negation of constraints is computed. The contributions (in fact a disjunction) of each of the original clauses are finally merged. The whole procedure can be seen as a negation normal form procedure, which is consistent with the operational semantics of the language.

The first ingredient is complementing the equality and freshness constraints, yielding $(\alpha)$-inequality $\text{not}[\tau]$ and non-freshness $\text{not}[\tau]$; we implement these using type-directed code generation within the $\alpha$Prolog interpreter and refer again to the appendix [10] for their generic definition.
Figure 3 shows goal and clause complementation: most cases of the former, via the \( \text{not}^G \) function, are intuitive, being classical tautologies. Note that the self-duality of the \( \mathcal{N} \)-quantifier allows goal negation to be applied recursively. Complementing existential goals is where we introduce \textit{extensional} quantification and invoke its proof-theory.

Clause complementation is where things get interesting and differ from the previous algorithm [9]. The complement of a clause \( G \supset p(t) \) must contain a “factual” part, built via term complementation, motivating failure due to clash with (some term in) the head. We obtain the rest by negating the body with \( \text{not}^{\neg G}(G) \). We take clause complementation \textit{definition-wise}, that is the negation of a program is the conjunction of the negation of all its predicate definitions. An example may help: negating the typing clauses for constants and application (\( \text{tc} \) from Fig 2) produces the following disjunction:

\[
\begin{align*}
& (\text{not}_{tc}(\_,\text{err},\_)) \lor \text{not}_{tc}(\_,\text{var}(\_),\_) \lor \text{not}_{tc}(\_,\text{app}(\_),\_)) \lor \\
& \text{not}_{tc}(\_,\text{lam}(\_),\_) \lor \text{not}_{tc}(\_,\text{c}(\C),\text{T}) := \text{neq}_{ty}(\text{tcf}(\C), \text{T}) \\
& \lor \\
& (\text{not}_{tc}(\_,\text{err},\_)) \lor \text{not}_{tc}(\_,\text{var}(\_),\_) \lor \text{not}_{tc}(\_,\text{c}(\_),\_) \lor \\
& \text{not}_{tc}(\_,\text{lam}(\_),\_) \lor \\
& \text{not}_{tc}(\_\text{app}(\text{M},\text{N}),\text{U}) := \text{forall}^* \text{T}. \text{not}_{tc}(\_\text{G},\text{fun}_{ty}(\text{T},\text{U})) \lor \\
& \text{not}_{tc}(\_\text{app}(\text{M},\text{N}),\text{U}) := \text{forall}^* \text{T}. \text{not}_{tc}(\_\text{G},\text{N},\text{T})
\end{align*}
\]

Notwithstanding the top-level disjunction, we are \textit{not} committing to any form of disjunctive logic programming: the key observation is that ‘\( \lor \)’ can be restricted to a program constructor \textit{inside} a predicate definition; therefore it can be eliminated by simulating unification in the definition:

\[
(G_1 \supset Q_1) \lor (G_2 \supset Q_2) \equiv \theta(G_1 \land G_2 \supset Q_1)
\]

where \( \theta = \text{mgu}(Q_1, Q_2) \). Because \( \lor \) is commutative and associative we can perform this merging operation in any order. However, as with many bottom-up operations, merging tends to produce a lot of redundancies in terms of clauses that are instances of each other. We have implemented \textit{backward} and \textit{forward} subsumption [23], by using an extension of the \( \alpha \)-Prolog interpreter itself to check entailment between newly generated clauses and the current database (and vice-versa). Despite the fact that this subsumption check is \textit{partial}, because the current unification algorithm does not handle equivariant unification with mixed prefixes [25] and extensional quantification [8], it makes all the difference: the \( \text{not}_{\text{is} \_ \text{err}} \) predicate definition decreases from an unacceptable 128 clauses to a much more reasonable 18. The final definition of \( \text{not}_{tc} \) follows, where we (as in Prolog) use the semicolon as concrete syntax for disjunction in the body:

\[
\begin{align*}
& \text{not}_{tc}(\_,\text{c}(\C),\text{T}) :- \text{neq}_{ty}(\text{tcf}(\C),\text{T}). \\
& \text{not}_{tc}(\_,\text{var}(\_),\_) \\
& \text{not}_{tc}([X,T]|G),\text{var}(X'),T') :- (\text{neq}_{ty}(T,T') ; \text{fresh}_{id}(X,X')), \\
& \text{not}_{tc}(G,\text{var}(X'),T'). \\
& \text{not}_{tc}(\_\text{app}(\text{M},\text{N}),\text{U}) :- \text{forall}^* \text{T}. \text{not}_{tc}(\_\text{G},\text{fun}_{ty}(\text{T},\text{U})); \\
& \text{not}_{tc}(\_\text{G},\text{N},\text{T}). \\
& \text{not}_{tc}(\_\text{app}(\text{M},\text{N}),\text{list}_{ty}) :- \text{forall}^* \text{T}. \text{not}_{tc}(\_\text{G},\text{M},\text{fun}_{ty}(\text{T},\text{list}_{ty})); \\
& \text{not}_{tc}(\_\text{G},\text{N},\text{T}). \\
& \text{not}_{tc}(\_\text{app}(\text{M},\text{N}),\text{int}_{ty}) :- \text{forall}^* \text{T}. \text{not}_{tc}(\_\text{G},\text{M},\text{fun}_{ty}(\text{T},\text{int}_{ty}));
\end{align*}
\]
not_tc(_,lam(_),listTy).
not_tc(_,lam(_),intTy).
not_tc(G,lam(M,T),funTy(T,U)):- new x:id. not_tc([x,T]|G],M@x,U).

Regardless of the presence of two subsumed clauses in the app case that our approach failed to detect, it is a big improvement in comparison to the 38 clauses generated by the previous algorithm [9]. And in exhaustive search, every clause counts.

Having synthesized the negation of the tc predicate, αCheck will use it internally while searching, for instance in the preservation check, for

$$\exists E. \exists T. \text{tc}([], E, T), \text{step}(E, E'), \text{not_tc}([], E', T)$$

Soundness of clause complementation is crucial for the purpose of model checking; we again express it in terms of exclusivity. The proof follows the lines of [26].

**Theorem 1 (Exclusivity).** Let $K$ be consistent. It is not the case that:

- $\Gamma; \Delta; K \vdash G$ and $\Gamma; \text{not}^D(\Delta); K \vdash \text{not}^G(G)$;
- $\Gamma; \Delta; K \xrightarrow{D} Q$ and $\Gamma; \text{not}^D(\Delta); K \xrightarrow{\text{not}^G(\Delta)} \text{not}^G(Q)$.

5 Case Studies

We have chosen as case studies here the Stlc benchmark suite, introduced in Section 2, and an encoding of the Volpano et al. security type system [34], as suggested in [5]. For the sake of space, we report at the same time our comparison between the various forms of negation, in particular NEs vs. NE, and the other systems of reference, accordingly, PLT-Redex and Nitpick.

**PLT-Redex** [13] is an executable DSL for mechanizing semantic models built on top of DrRacket. Redex has been the first environment to adopt the idea of random testing a la QuickCheck for validating the meta-theory of object languages, with significant success [18]. As we have mentioned, the main drawbacks are the lack of support for binders and low coverage of test generators stemming from grammar definitions. The user is therefore required to write her own generators, a task which tends to be demanding.

The system where proofs and disproofs are best integrated is arguably Isabelle/HOL [4]. In the appendix [10] we report some comparison with its version of QuickCheck, but here we concentrate on Nitpick [5], a higher-order model finder in the Alloy lineage supporting (co)inductive definitions. Nitpick works translating a significant fragment of HOL into first-order relational logic and then invoking Alloy’s SAT-based model enumerator. The tool has been used effectively in several case studies, most notably weak memory models for C++ [6]. It would be natural to couple Isabelle/HOL’s QuickCheck and/or Nitpick’s capabilities with Nominal Isabelle [33], but this would require strengthening the latter’s support for computation with names, permutations and abstract syntax modulo α-conversion. So,
at the time of writing, αCheck is unique as a model checker for binding signatures and specifications.

All test have been performed under Ubuntu 15.4 on a Intel Core i7 CPU 870, 2.93GHz with 8GB RAM. We time-out the computation when it exceeds 200 seconds. We report 0 when the time is <0.01. These tests must be taken with a lot of salt: not only is our tool under active development but the comparison with the other systems is only roughly indicative, having to factor differences between logic and functional programming (PLT-Redex), as well as the sheer scale and scope of counter-examples search in a system such as Isabelle/HOL.

5.1 Head-to-Head with PLT-Redex

We first measure the amount of time to exhaust the search space (TESS) using the three versions of negations supported in αCheck, over a bug-free version of the Stlc benchmark for \( n = 1, 2, \ldots \) up to the point where we time-out. This gives some indication of how much of the search space the three techniques explore, keeping in mind that what is traversed is very different in shape; hence the more reliable comparison is between NE and NEs. As the results depicted in Figure 4 suggests, NEs shows a clear improvement over NE, while NF holds its ground, however hindered by the explosive exhaustive generation of terms.

![Loglinear-plot of TESS on prog theorem](image)

Fig. 4. Loglinear-plot of TESS on prog theorem

However, our mission is finding counterexamples and so we compare the time to find counterexamples (TFCE) using NF, NE, NEs on the said benchmarks. We list in Table 1 the 9 mutations from the cited site. Every row describes the mutation inserted with an informal classification inherited from ibidem — (S)imple, (M)edium or (U)nusual, better read as artificial. We also list the counterexamples found by αCheck under NF (NE(s) being analogous but less instantiated) and the depths at which those are found or a time-out occurred.

The results in Table 1 show a remarkable improvement of NEs over NE, in terms of counter-examples that were timed-out (bug 2 and 5), as well as major speedups of more than an order of magnitude (bugs 3 (ii) and 7). Further, NEs never under-performs NE, probably because it locates counterexample at a lower depth. In rare occasions (bug 5 again) NEs even outperforms NF and in several cases it is comparable (bug 1, 3, 7, 8 and 9). Of course there are occasions (2 and
Table 1. TFCE on the Stlc benchmark, Redex-style encoding

<table>
<thead>
<tr>
<th>bug check</th>
<th>NF</th>
<th>NE</th>
<th>NEs</th>
<th>cex</th>
<th>Description/Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 (7)</td>
<td></td>
<td>1 (7)</td>
<td>0.37 (7)</td>
<td>(λx. x err) n</td>
<td>range of function in app rule</td>
</tr>
<tr>
<td>3.31 (9)</td>
<td>0.27 (5)</td>
<td>hd n</td>
<td>matched to the arg. (S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.27 (8)</td>
<td>t.o. (11)</td>
<td>85.3 (12)</td>
<td>(cons n) nil</td>
<td>value (cons v) v omitted (M)</td>
<td></td>
</tr>
<tr>
<td>0.04 (6)</td>
<td>0.04 (6)</td>
<td>0.3 (6)</td>
<td>(λx. n) m</td>
<td>order of types swapped</td>
<td></td>
</tr>
<tr>
<td>0.04 (6)</td>
<td>0.37 (9)</td>
<td>0.27 (8)</td>
<td>hd n</td>
<td>in function pos of app (S)</td>
<td></td>
</tr>
<tr>
<td>t.o.</td>
<td>t.o.</td>
<td>t.o.</td>
<td>tail red. returns the head (S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.8 (11)</td>
<td>t.o.</td>
<td>t.o. (12)</td>
<td>hd ((cons n) nil)</td>
<td>hd red. on part. applied cons (M)</td>
<td></td>
</tr>
<tr>
<td>1.04 (9)</td>
<td>18.5 (10)</td>
<td>1.1 (9)</td>
<td>hd ((λx. err) n)</td>
<td>no eval for argument of app (M)</td>
<td></td>
</tr>
<tr>
<td>0.02 (5)</td>
<td>0.03 (5)</td>
<td>0.1 (5)</td>
<td>(λx. x) nil</td>
<td>lookup always returns int (U)</td>
<td></td>
</tr>
<tr>
<td>0.02 (5)</td>
<td>0.1 (5)</td>
<td>(λx. y) n</td>
<td>vars do not match in lookup (S)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6), where NF is still dominant, as NEs counter-examples live at steeper depths (12 and 16, respectively) that cannot yet be achieved within the time-out.

We do not report TFCE of PLT-Redex, because, being based on randomized testing, what we really should measure is time spent on average to find a bug. The two encodings are quite different: Redex has very good support for evaluation contexts, while we use congruence rules. Being untyped, the Redex encoding treats err as a string, which is then procedurally handled in the statement of preservation and progress, whereas for us it is part of the language. Since [18], Redex allows the user to write certain judgments in a declarative style, provided they can be given a functional mode, but more complex systems, such as typing for a polymorphic version of a similar calculus, require very indirect encoding, e.g. CPS-style. We simulate addition on integers with numerals (omitted from the code snippets presented in Section 2 for the sake of space), as we currently require our code to be pure in the logical sense, as opposed to Redex that maps integers to Racket’s ones. W.r.t. lines of code, the size of our encoding is roughly 1/4 of the Redex version, not counting Redex’s built-in generators and substitution function. The adopted checking philosophy is also somewhat different: they choose to test preservation and progress together, using a cascade of three built-in generators and collect all the counterexamples found within a timeout.

The performance of the negation elimination variants in this benchmark is not too impressive. However, if we adopt a different style of encoding (let’s call it PCF, akin to what we used in [9]), where constructors such as hd are not treated as constants, but are first class, e.g.:

\[
tc(G,hd(E),\text{intTy}) \\ step(hd(cons(H,Tl)), H) \\ step(hd(cons(H,Tl)), H) \\ t(\text{intTy})
\]

then all counter-examples are found very quickly, as reported in Table 2. In bug 4, NEs struggles to get at depth 13; on the other hand PLT-Redex fails to find that very bug. Bug 6 as well as several counterexamples disappear as not well-typed. This improved efficiency may be due to the reduced amount of nesting of
terms, which means lower depth of exhaustive exploration. This is not a concern for random generation and (compiled) functional execution as in PLT-Redex.

### 5.2 Nitpicking Security Type Systems

To compare Nitpick with our approach, we use the security type system due to Volpano, Irvine and Smith [34], whereby the basic imperative language IMP is endowed with a type system that prevents information flow from private to public variables. For our test, we actually selected the more general version of the type system formalized in [28], where the security levels are generalized from high and low to natural numbers. Given a fixed assignment sec of such security levels to variables, then lifted to arithmetic and Boolean expressions, the typing judgment \( l \vdash c \) reads as “command \( c \) does not contain any information flow to variables \( < l \) and only safe flows to variables \( \geq l \).” Following [28], we call this system syntax-directed.

The main properties of interest relate states that agree on the value of each variable (strictly) below a certain security level, denoted as \( \sigma_1 \approx_{<l} \sigma_2 \) iff \( \forall x. \text{sec} x < l \to \sigma_1(x) = \sigma_2(x) \). Assume a standard big-step evaluation semantics for IMP, relating an initial state \( \sigma \) and a command \( c \) to a final state \( \tau \):

**Confinement** If \( \langle c, \sigma \rangle \downarrow \tau \) and \( l \vdash c \) then \( \sigma \approx_{<l} \tau \);

**Non-interference** If \( \langle c, \sigma \rangle \downarrow \sigma', \langle c, \tau \rangle \downarrow \tau', \sigma \approx_{\leq l} \tau \) and \( 0 \vdash c \) then \( \sigma' \approx_{\leq l} \tau' \);

We extend this exercise by considering also a declarative version (std) \( l \vdash_d c \) of the syntax directed system, where anti-monotonicity is taken as a primitive rule instead of an admissible one as in the previous system; finally we encode also a syntax-directed termination-sensitive (stT) version \( l \vdash_{stT} c \), where non-terminating programs do not leak information and its declarative cousin (stTd) \( l \vdash_{stTd} c \). We then insert some mutations in all those systems, as detailed in Table 3 and investigate whether the following equivalences among those systems still hold:

---

Table 2. TFCE on the Stlc benchmark, PCF-style encoding. NEs cex shown

<table>
<thead>
<tr>
<th>bug#</th>
<th>check</th>
<th>NF</th>
<th>NE</th>
<th>NEs</th>
<th>cex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pres</td>
<td>0.05 (5)</td>
<td>2.79 (5)</td>
<td>0.04 (5)</td>
<td>((\lambda x. \text{hd} x)) N</td>
</tr>
<tr>
<td>2</td>
<td>prog</td>
<td>0 (4)</td>
<td>7.76 (9)</td>
<td>0.8 (7)</td>
<td>((\text{cons} N)) nil</td>
</tr>
<tr>
<td>3</td>
<td>prog</td>
<td>0 (4)</td>
<td>0.05 (4)</td>
<td>0 (4)</td>
<td>((\lambda x. \text{nil})) nil</td>
</tr>
<tr>
<td>4</td>
<td>pres</td>
<td>0.15 (7)</td>
<td>199.1 (12)</td>
<td>N + ((\text{cons} N) nil)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>prog</td>
<td>0 (4)</td>
<td>0.04 (4)</td>
<td>0 (4)</td>
<td>t ((\lambda x. \text{nil})) nil</td>
</tr>
<tr>
<td>7</td>
<td>prog</td>
<td>5.82 (9)</td>
<td>151.2 (11)</td>
<td>19.54. (10)</td>
<td>((\lambda x. \text{nil})) (N + M)</td>
</tr>
<tr>
<td>8</td>
<td>prog</td>
<td>0 (4)</td>
<td>0.04 (4)</td>
<td>0.1 (4)</td>
<td>((\lambda x. x)) nil</td>
</tr>
<tr>
<td>9</td>
<td>pres</td>
<td>0 (4)</td>
<td>0.04 (4)</td>
<td>0.1 (4)</td>
<td>((\lambda x. y)) N</td>
</tr>
</tbody>
</table>

---

For an interesting case study regarding instead dynamic information flow and carried out in Haskell, see [17]. A large part of the paper is dedicated to the fine tuning of custom generators and shrinkers.
Table 3. αCheck vs. Nitpick on the Volpano benchmark suite. (sp) indicates that Nitpick produced a spurious counterexample.

<table>
<thead>
<tr>
<th>bug check</th>
<th>Nitpick NF</th>
<th>NE</th>
<th>NEs</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 conf</td>
<td>(sp) 0.03 (5) 4.4 (8) 2.1 (7)</td>
<td>second premise of seq rule omitted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-inter</td>
<td>t.o. 9.13 (8) 6.71 (8) 6.1 (8)</td>
<td>ditto</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 non-inter (sp) 3.3 (8) 2.1 (8) 1.9 (8)</td>
<td>var swapping in ≤ premise of assn rule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 st→std 0.95 t.o. t.o t.o.</td>
<td>inversion of ≤ in antimono rule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>std→st 0.75 0.8 (7) 0.3 (7) 0.3 (7)</td>
<td>ditto</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 st→std ≤ assumption omitted in IF: true</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>std→st 1.3 0.9 (7) t.o. t.o.</td>
<td>ditto</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 st→std 5.1(sp) 24.5 (11) t.o. t.o.</td>
<td>as 2 but on decl. version of the rule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>std→st 1.1 0.2 (7) t.o. 24.6 (11)</td>
<td>ditto</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 stT→stTd 5.1(sp) t.o. t.o. t.o.</td>
<td>as 2 but on term. version of the rule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stTd→stT 1.0 0.01 (5) 0.32 (7) 0.05 (6)</td>
<td>ditto</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 stT→stTd same as 4 but on term-decl. rule: true</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stTd→stT 1.6 1.7 (8) 12.5 (9) 1.2(8)</td>
<td>ditto</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\text{st↔std} \vdash c$ iff $\vdash d c$ and $\text{stT↔stTd} \vdash y c$ iff $\vdash y d c$.

Again the experimental evidence is quite pleasing as far as NE vs. NEs goes, where the latter is largely superior (5 (ii), 1 (i), 7 (ii)). In one case NEs improves on NF (1 (ii)) and in general competes with it save for 4 (ii) and 5 (i) and (ii). To have an idea of the counterexamples found by αCheck, the command (SKIP ; x := 1), see x = 0, 1 = 1 and state σ mapping x to 0 falsifies confinement 1 (i); in fact, this would not hold were the typing rule to check the second premise. A not too dissimilar counterexample falsifies non-interference 1 (ii): $c$ is (SKIP ; x := y), sec x, y = 0, 1, 1 = 0 and σ maps y to 0 and x undefined (i.e. to a logic variable), while τ maps y to 1 and keeps x undefined. We note in passing that here extensional quantification is indispensable, since ordinary generic quantification is unable to instantiate security levels so as to find the relevant bugs.

The comparison with Nitpick is more mixed. On one hand Nitpick fails to find 1 (ii) within the timeout and in other four cases it reports spurious counterexamples, which on manual analysis turn out to be good. On the other it nails down, quite quickly, two other cases where αCheck fails to converge at all (3 (i), 6 (i)). This despite the facts that relations such as evaluations, $\vdash d$ and $\vdash y d$, are reported not well founded requiring therefore a problematic unrolling.

The crux of the matter is that differently from Isabelle/HOL’s mostly functional setting (except for inductive definition of evaluation and typing), our encoding is fully relational: states and security assignments cannot be seen as partial functions but are reified in association lists. Moreover, we pay a significant price in not being able to rely on built-in types such as integers, but have to deploy our

---

5 Settings: [sat Solver=MiniSatJNI, max_threads=1, check potential, timeout = 200]
clearly inefficient versions. This means that to falsify simple computations such as \( n \leq m \), we need to provide a derivation for that failure. Finally, this case study does not do justice to the realm where αProlog excels, namely it does not exercise binders intensely: we are only using nominal techniques in representing program variables as names and freshness to guarantee well-formedness of states and of the table encoding the variable security settings. Yet, we could not select more binding intensive examples due to the current difficulties with running Nitpick under Nominal Isabelle.

6 Conclusions and Future Work

We have presented a new implementation of the NE algorithm underlying our model checker αCheck and experimental evidence showing satisfying improvements w.r.t. the previous incarnation, so as to make it competitive with the NF reference implementation. The comparison with PLT-Redex and Nitpick, systems of considerable additional maturity, is also, in our opinion, favourable: αCheck is able to find similar counterexamples in comparable amounts of time; it is able to find some counterexamples that Redex or Nitpick respectively do not; and in no case does it report spurious counterexamples. Having said that, our comparison is at most just suggestive and certainly partial, as many other proof assistants have incorporated some notion of PBT, e.g. [29,31]. A notable absence here is a comparison with what at first sight is a close relative, the Bedwyr system [2], a logic programming engine that allows a form of model checking directly on syntactic expressions possibly containing binding. Since Bedwyr uses depth-first search, checking properties for infinite domains should be approximated by writing logic programs encoding generators for a finite portion of that model. Our initial experiments in encoding the Stlc benchmark in Bedwyr have failed to find any counterexample, but this could be imputed simply to our lack of experience with the system. Recent work about “augmented focusing systems” [16] could overcome this problem.

All the mutations we have inserted so far have injected faults in the specifications, not in the checks. This make sense for our intended use, where the properties we validate are the main theorems that our calculi should satisfy. However, it would be interesting to see how our tool would fare w.r.t. mutation testing of theorems.

Exhaustive term generation has served us well so far, but it is natural to ask whether random generation could have a role in αCheck, either by simply randomizing term generation under NF or more generally the logic programming interpreter itself, in the vein of [14]. More practically, providing generators and reflection mechanism for built-in datatypes and associated operators is a priority.

Finally, we would like to implement improvements in nominal equational unification algorithms, which would make subsumption complete, via equivariant unification [8], and more ambitiously introduce narrowing, so that functions could be computed rather then simulated relationally. In the long run, this could open the door to use αCheck as a light-weight model checker for (a fragment) of Nominal Isabelle.
References


