Advances in Property-Based Testing for αProlog

James Cheney[1], Alberto Momigliano[2], Matteo Pessina[2]

[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), witnessed by the success of *CompCert* and subsequent projects [23, 36]. 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, 8], verification of more realistic optimizations seems even harder, e.g. *call arity* 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, [8]).

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* [25] has been suggested in combination with automatic generation of (expressive) test programs, most notably for C compilers with the *Csmith* tool [37] 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 [14], 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 formal 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 some traction in the field of semantics engineering, see in particular the QuickCheck/Nitpick combination offered in Isabelle/HOL [5] and random testing in PLT-Redex [20]. However, a particular dimension to validating for example optimizations in a compiler such as GHC, whose intermediate language *Core* 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 [13] 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 [11]. 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 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, [20]).

Searching for counterexamples means finding values 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)

[3] Xavier Leroy, personal communication. In fact, the encoding of the λ-calculus in [24] does not respect α-equivalence, nor does it implement substitutions in a capture avoiding way.

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 lacking 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 success set is included in the complement of the success set of the original predicates that occurred negatively. 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 [11], *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) that improves significantly on the performance of *NE* as described in [11], so much as to make it competitive *w.r.t.* the *NF* version;
2. and by evaluating our checker in comparison with some of its competitors in the logical framework niche, namely QuickCheck/Nitpick [5] and PLT-Redex [20]. 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 Appendix A contains additional experiments and some formal notions used in Section 3.

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 are available at [12].

2 A Brief Tour of αCheck

We specify the formal systems and the properties we wish to check as Horn logic programs in αProlog [13], a logic programming language based on *nominal logic*, a first-order theory axiomatizing names and name-binding introduced by Pitts [32].

In αProlog, there are several built-in types, functions, and relations with special behavior. There are distinguished *name types* that are 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 `a\M` in programs, considered equal up to α-renaming of the bound name. Thus, where one writes $\lambda x.M$, $\forall x.M$, etc. in a paper exposition, in αProlog one writes `lam(x\M)`, `forall(x\M)`, 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 αProlog as `x # t`; in particular, if t is also a name then freshness is name-inequality. For convenience, αProlog 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 core 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 λ-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:

$$
\begin{array}{lll}
\text{Types} & A,B & ::= int \,|\, ilist \,|\, A \rightarrow B \\
\text{Terms} & M & ::= x \,|\, \lambda x{:}A.\ M \,|\, M_1\ M_2 \,|\, c \,|\, err \\
\text{Constants} & c & ::= n \,|\, nil \,|\, cons \,|\, hd \,|\, tl \\
\text{Values} & V & ::= c \,|\, \lambda x{:}A.\ M \,|\, cons\ V \,|\, cons\ V_1\ V_2
\end{array}
$$

We start (see the top of Figure 1) by declaring the syntax of terms, constant and types, while values will be carved out *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, following standard practice.

We follow this up (see the remainder of Figure 1) with the static (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 [11]. 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*

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

```
ty: type.
intTy: ty.          funTy: (ty,ty) -> ty.     listTy: ty.
cst: type.
toInt: int -> cst.  nil: cst.  cons: cst.  hd: cst.  tl: cst.
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).
tc([(Y,_)|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)),VS)   :- 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

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:

```
tc(G,app(M,N),U) :- tc(G,M,funTy(T,U)), tc(G,N,U). % was funTy(U,T)
```

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

```
E = app(c(hd),c(toInt(N)))
T = intTy
```

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

```
E = app(lam(x\app(var(x),err),funTy(T,intTy)),c(toInt(N)))
E' = app(c(toInt(N)),err)
T = intTy
```

3 The Core Language

In this section we give the essential notions concerning the core syntax and semantics of αProlog programs. The surface syntax used in the previous section desugars to this core language.

An αProlog *signature* is composed by sets Σ_D and Σ_N of, respectively, base data types δ, which includes a type o of *propositions*, and name types ν; a collection Σ_P of *predicate symbols* $p{:}\tau \to o$ and one Σ_F of *function symbol* declarations $f{:}\tau \to \delta$. Types τ are formed as specified by the following grammar:

$$\tau ::= \mathbf{1} \,|\, \delta \,|\, \tau \times \tau' \,|\, \nu \,|\, \langle \nu \rangle \tau$$

where $\delta \in \Sigma_D$ and $\nu \in \Sigma_N$. Given a signature, the language of *terms* is defined over sets $V = \{X,Y,Z,...\}$ of logical variables and sets $A = \{\mathsf{a},\mathsf{b},...\}$ of names:

$$\begin{aligned} t,u &::= \mathsf{a} \,|\, \pi \cdot X \,|\, \langle\rangle \,|\, \langle t,u \rangle \,|\, \langle \mathsf{a} \rangle t \,|\, f(t) \\ \pi &::= \mathsf{id} \,|\, (\mathsf{a}\ \mathsf{b}) \circ \pi \end{aligned}$$

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

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

$$\frac{\mathsf{a} \neq \mathsf{b}}{\mathsf{a} \,\#_\nu\, \mathsf{b}} \quad \frac{}{\mathsf{a} \,\#_{\mathbf{1}}\, \langle\rangle} \quad \frac{\mathsf{a} \,\#_\tau\, t}{\mathsf{a} \,\#_\delta\, f(t)} \quad \frac{\mathsf{a} \,\#_{\tau_1}\, t_1 \quad \mathsf{a} \,\#_{\tau_2}\, t_2}{\mathsf{a} \,\#_{\tau_1 \times \tau_2}\, \langle t_1,t_2 \rangle} \quad \frac{\mathsf{a} \,\#_{\nu'}\, \mathsf{b} \quad \mathsf{a} \,\#_\tau\, t}{\mathsf{a} \,\#_{\langle \nu' \rangle \tau}\, \langle \mathsf{b} \rangle t} \quad \frac{}{\mathsf{a} \,\#_{\langle \nu' \rangle \tau}\, \langle \mathsf{a} \rangle t}$$

In the same way we define the equality relation, which identifies terms modulo α-equivalence:

$$\frac{}{\mathsf{a} \approx_\nu \mathsf{a}} \quad \frac{}{\langle\rangle \approx_{\mathbf{1}} \langle\rangle} \quad \frac{t_1 \approx_{\tau_1} u_1 \quad t_2 \approx_{\tau_2} u_2}{\langle t_1,t_2 \rangle \approx_{\tau_1 \times \tau_2} \langle u_1,u_2 \rangle} \quad \frac{t \approx_\tau u}{f(t) \approx_\delta f(u)}$$

$$\frac{\mathsf{a} \approx_\nu \mathsf{b} \quad t \approx_\tau u}{\langle \mathsf{a} \rangle t \approx_{\langle \nu \rangle \tau} \langle \mathsf{b} \rangle u} \quad \frac{\mathsf{a} \,\#_\nu\, (\mathsf{b},u) \quad t \approx_\tau (\mathsf{a}\ \mathsf{b}) \cdot u}{\langle \mathsf{a} \rangle t \approx_{\langle \nu \rangle \tau} \langle \mathsf{b} \rangle u}$$

We omit the type subscript when it is clear from the context.

Given a signature, *goals* G and *program clauses* D have the following form:

$$\begin{aligned} A &::= t \approx u \,|\, t \# u \\ G &::= \bot \,|\, \top \,|\, A \,|\, p(t) \,|\, G \wedge G' \,|\, G \vee G' \,|\, \exists X{:}\tau.\ G \,|\, \text{И}\mathsf{a}{:}\nu.\ G \,|\, \forall^* X{:}\tau.\ G \\ D &::= \top \,|\, p(t) \,|\, D \wedge D' \,|\, G \supset D \,|\, \forall X{:}\tau.\ D \,|\, \bot \,|\, D \vee D' \end{aligned}$$

The productions shown in black yield a fragment of nominal logic called И-goal clauses [13], for which resolution based on nominal unification is sound and complete. We rely on the fact that D formulas in a program Δ can always be normalized to sets of clauses of the form $\forall \boldsymbol{X}{:}\boldsymbol{\tau}.\ G \supset p(t)$, denoted def$(p,\Delta)$. The *fresh-name* quantifier И, firstly introduced in [32], quantifies over names not occurring in a formula (or in the values of its variables). It can be defined in terms of freshness #, that is the formula И$\mathsf{a}{:}\nu.\ \phi$ is logically equivalent to $\exists A{:}\nu.\ A \# \boldsymbol{X} \wedge \phi$, where $\boldsymbol{X}$ are the free variables appearing in ϕ. Here we take И-quantified formulas as primitive.

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. An unusual feature is $\forall^*$, the *extensional* universal quantifier [17]. 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 τ, $\forall^* X{:}\tau.\ G$ succeeds if and only if $G[t/X]$ does for every ground term of type τ.

Constraints are G-formulas of the following form:

$$C ::= \top \mid t \approx u \mid t \# u \mid C \wedge C' \mid \exists X{:}\tau.\ C \mid \text{И}\mathsf{a}{:}\nu.\ C$$

We write $\mathcal{K}$ for a set of constraints and Γ for a context keeping track of the types of variables and names. Constraint-solving is modeled by the judgment $\Gamma;\mathcal{K} \models C$, which holds if for all valutations θ from variables to ground terms $\theta : \Gamma$ such that $\theta \models \mathcal{K}$, we have $\theta \models C$. The latter notion of satisfiability is standard, modulo handling of names: for example $\theta \models$ И$\mathsf{a}{:}\nu.\ C$ iff for some $\mathsf{b} \# (\theta, C), \theta \models C[\mathsf{b}/\mathsf{a}]$.

We can describe an idealized interpreter for αProlog with the "amalgamated" proof-theoretic semantics introduced in [13] and inspired by similar techniques stemming from CLP [22] — see Figure 2, sporting two kind of judgments, goal-directed proof search $\Gamma;\Delta;\mathcal{K} \Rightarrow G$ and focused proof search $\Gamma;\Delta;\mathcal{K} \xrightarrow{D} Q$. This semantics allows us to concentrate on the high-level proof search issues, without requiring us to introduce or manage low-level operational details concerning constraint solving. We refer the reader to [13] for more explanation and ways to make those judgments operational. Note that the rule $\forall^*\omega$ says that goals of the form $\forall^* X{:}\tau.G$ can be proved if $\Gamma, X{:}\tau;\Delta;\mathcal{K}, C \Rightarrow G$ is provable for every constraint C such that $\Gamma;\mathcal{K} \models \exists X{:}\tau.\ C$ 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 eigen-variable.

4 Specification Checking

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

$$\text{И}\mathbf{a}.\forall \boldsymbol{X}.\ G \supset A \tag{1}$$

where G is a goal and A an atomic formula (including equality and freshness constraints). Since the И-quantifier is self-dual, the negation of (1) is of the form И$\mathbf{a}.\exists \boldsymbol{X}.G \wedge \neg A$.

$$\frac{\Gamma;\mathcal{K}\models A}{\Gamma;\Delta;\mathcal{K}\Rightarrow A}\ con \qquad \frac{\Gamma;\Delta;\mathcal{K}\Rightarrow G_1 \quad \Gamma;\Delta;\mathcal{K}\Rightarrow G_2}{\Gamma;\Delta;\mathcal{K}\Rightarrow G_1\wedge G_2}\ \wedge R$$

$$\frac{\Gamma;\Delta;\mathcal{K}\Rightarrow G_i}{\Gamma;\Delta;\mathcal{K}\Rightarrow G_1\vee G_2}\ \vee R_i \qquad \frac{\Gamma;\mathcal{K}\models \exists X{:}\tau.\ C \quad \Gamma,X{:}\tau;\Delta;\mathcal{K},C\Rightarrow G}{\Gamma;\Delta;\mathcal{K}\Rightarrow \exists X{:}\tau.\ G}\ \exists R$$

$$\frac{\Gamma;\mathcal{K}\models И\mathsf{a}{:}\nu.\ C \quad \Gamma\#\mathsf{a}{:}\nu;\Delta;\mathcal{K},C\Rightarrow G}{\Gamma;\Delta;\mathcal{K}\Rightarrow И\mathsf{a}{:}\nu.\ G}\ ИR$$

$$\frac{}{\Gamma;\Delta;\mathcal{K}\Rightarrow \top}\ \top R \qquad \frac{\Gamma;\Delta;\mathcal{K}\xrightarrow{D} Q \quad D\in\Delta}{\Gamma;\Delta;\mathcal{K}\Rightarrow Q}\ sel$$

$$\frac{\bigwedge\{\Gamma,X{:}\tau;\Delta;\mathcal{K},C\Rightarrow G \mid \Gamma;\mathcal{K}\models \exists X{:}\tau.\ C\}}{\Gamma;\Delta;\mathcal{K}\Rightarrow \forall^{*}X{:}\tau.\ G}\ \forall^{*}\omega$$

....

$$\frac{\Gamma;\mathcal{K}\models t\approx u}{\Gamma;\Delta;\mathcal{K}\xrightarrow{p(t)} p(u)}\ hyp \qquad \frac{\Gamma;\Delta;\mathcal{K}\xrightarrow{D_i} Q}{\Gamma;\Delta;\mathcal{K}\xrightarrow{D_1\wedge D_2} Q}\ \wedge L_i$$

$$\frac{\Gamma;\Delta;\mathcal{K}\xrightarrow{D} Q \quad \Gamma;\Delta;\mathcal{K}\Rightarrow G}{\Gamma;\Delta;\mathcal{K}\xrightarrow{G\supset D} Q}\ \supset L$$

$$\frac{\Gamma;\mathcal{K}\models \exists X{:}\tau.\ C \quad \Gamma,X{:}\tau;\Delta;\mathcal{K},C\xrightarrow{D} Q}{\Gamma;\Delta;\mathcal{K}\xrightarrow{\forall X:\tau.\ D} Q}\ \forall L$$

$$\frac{}{\Gamma;\Delta;\mathcal{K}\xrightarrow{\perp} Q}\ \perp L \qquad \frac{\Gamma;\Delta;\mathcal{K}\xrightarrow{D_1} Q \quad \Gamma;\Delta;\mathcal{K}\xrightarrow{D_2} Q}{\Gamma;\Delta;\mathcal{K}\xrightarrow{D_1\vee D_2} Q}\ \vee L$$

Fig. 2. Proof search semantics of αProlog programs

A *(finite) counterexample* is a closed substitution θ providing values for $\boldsymbol{X}$ 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 *negation*.

In αCheck the reference implementation reads negation as *finite failure* (*not*):

$$И\mathbf{a}.\exists \boldsymbol{X}{:}\boldsymbol{\tau}.\ G\wedge gen[[\boldsymbol{\tau}]](\boldsymbol{X})\wedge not(A)$$

where $gen[[\boldsymbol{\tau}]]$ are type-indexed predicates that *exhaustively* enumerate the (ground) inhabitants of $\boldsymbol{\tau}$. As we have mentioned in the Introduction, 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 (α)Prolog program, whose success set is included in the complement of the success set of the original predicates that occurred negatively. In other terms, a predicate and its complement are mutually *exclusive*. *Exhaustivity* may or may not hold, depending on the decidability of the predicate in question, but 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 *extensional* universal quantifier presented in the previous section.

The generation of complementary predicates can be split into two phases: *term complementation* and *clause complementation*.

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 *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 [21]. This forces terms in source clause heads to be linear and free of names (including swapping and abstractions), by replacing them with logical variables, and, in case they occurred in abstractions, by constraining them in the clause body by a *concretion* to a fresh variable. A concretion, written $t@a$, is the elimination form for abstractions and does not require to be taken as a primitive since it can be implemented by translating the goal $G[t@a]$ to $\exists X.t \approx \langle a \rangle X \wedge G[X]$. For example, the clause for typing lambdas is normalized as follows:

```
tc(G,lam(M,T),funTy(T,U)):- new x. tc([(x,T) |G],M@x,U).
```

Hence, we can use a type-directed version of first-order term complementation, $not[[\tau]] : \tau \rightarrow \tau\ set$ and prove its correctness in term of *exclusivity* following [3, 27]: the intersection of the set of ground instances of a term and its complement is empty. *Exhaustivity* also holds, but will not be needed. The definition of $not[\![\tau]\!]$ is in the appendix A, but we offer the following example:

$$\begin{array}{l} not[[\mathtt{exp}]](\mathtt{app}(\mathtt{c}(\mathtt{hd}),\_)) = \\ \;\{\mathtt{lam}(\_,\_),\mathtt{err},\mathtt{c}(\_),\mathtt{var}(\_),\mathtt{app}(\mathtt{c}(\mathtt{tl}),\_),\mathtt{app}(\mathtt{c}(\mathtt{nil}),\_),\mathtt{app}(\mathtt{c}(\mathtt{toInt}(\_)),\_), \\ \;\mathtt{app}(\mathtt{var}(\_),\_),\mathtt{app}(\mathtt{err},\_),\mathtt{app}(\mathtt{lam}(\_,\_),\_),\mathtt{app}(\mathtt{app}(\_,\_),\_)\} \end{array}$$

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.

$$
\begin{array}{rlrl}
not^G(\top) &= \bot & not^D(\top) &= \bot \\
not^G(\bot) &= \top & not^D(\bot) &= \top \\
not^G(p(t)) &= p^{\neg}(t) & not^D(G \supset p(t)) &= \bigwedge\{\forall(p^{\neg}(u)) \mid u \in not[[\tau]](t)\} \wedge \\
& & & \quad (not^G(G) \supset p^{\neg}(t)) \\
not^G(t \approx_\tau u) &= neq[[\tau]](t,u) & & \\
not^G(a \#_\tau u) &= nfr[[\nu,\tau]](a,u) & & \\
not^G(G \wedge G') &= not^G(G) \vee not^G(G') & not^D(D \wedge D') &= not^D(D) \vee not^D(D') \\
not^G(G \vee G') &= not^G(G) \wedge not^G(G') & not^D(D \vee D') &= not^D(D) \wedge not^D(D') \\
not^G(\forall^* X{:}\tau.\ G) &= \exists X{:}\tau.\ not^G(G) & not^D(\forall X{:}\tau.\ D) &= \forall X{:}\tau.\ not^D(D) \\
not^G(\exists X{:}\tau.\ G) &= \forall^* X{:}\tau.\ not^G(G) & & \\
not^G(\text{И}\mathsf{a}{:}\nu.\ G) &= \text{И}\mathsf{a}{:}\nu.\ not^G(G) & not^D(\Delta) &= not^D(\mathrm{def}(p,\Delta))
\end{array}
$$

Fig. 3. Negation of a goal and of clause

The first ingredient is complementing the equality and freshness constraints, yielding (α-)inequality $neq[[\tau]]$ and non-freshness $nfr[[\nu,\delta]]$: we implement these using type-directed code generation within the αProlog interpreter and refer again to the appendix A for their generic definition.

Figure 3 shows goal and clause complementation: most cases of the former, *via* the not^G function, are intuitive, being classical tautologies. Note that the self-duality of the И-quantifier allows goal negation to be applied recursively. Complementing existential goals is where we introduce *extensional* quantification and invoke its proof-theory.

Clause complementation is where things get interesting and differ from the previous algorithm [11]. The complement of a clause $p(t) \leftarrow G$ 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 $not^G(G)$. We take clause complementation *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 produces the following disjunction:

```
(not_tc(_,err,_) /\ not_tc(_,var(_),_) /\ not_tc(_,app(_,_),_) /\
 not_tc(_,lam(_,_),_) /\ not_tc(_,c(C),T):- neq(tcf(C), T))
 \/
(not_tc(_,err,_) /\ not_tc(_,var(_),_) /\ not_tc(_,c(_),_) /\ not_tc(_,lam(_,_),_) /\
 not_tc(G,app(M,N),U):- forall* T. not_tc(G,M,arr(T,U));not_tc(G,N,T)).
```

Notwithstanding the top-level disjunction, we are *not* committing to any form of disjunctive logic programming: the key observation is that '$\vee$' can be restricted to a program constructor *inside* a predicate definition; therefore it can be eliminated by simulating unification in the definition:

$$(Q_1 \leftarrow G_1) \vee (Q_2 \leftarrow G_2) \equiv \theta(Q_1 \leftarrow G_1 \wedge G_2)$$

where $\theta = \mathrm{mgu}(Q_1, Q_2)$. Because $\vee$ 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 *backward* and *forward* subsumption, by using an extension of the αProlog interpreter itself to check entailment between newly generated clauses and the current database (and vice-versa). Despite the fact that this subsumption check is *partial*, because the current unification algorithm does not handle equivariant unification with mixed prefixes and extensional quantification [10], it makes the difference: the `not_is_err` predicate definition decreases from an unacceptable 128 clauses to a much more reasonable 18. The final definition of `not_tc`:

```
not_tc(_,c(C),T)              :- neq_ty(tccf(C),T).
not_tc([],var(_),_).
not_tc([(X,T)|G],var(X'),T') :- (neq_ty(T,T'); fresh_id(X,X')), not_tc(G,var(X'),T').
not_tc(G,app(M,N),U)          :- forall T:ty. not_tc(G,M,funTy(T,U)); not_tc(G,N,T).
not_tc(G,app(M,N),listTy)     :- forall T:ty. not_tc(G,M,funTy(T,listTy)); not_tc(G,N,T).
not_tc(G,app(M,N),intTy)      :- forall T:ty. not_tc(G,M,funTy(T,intTy)); not_tc(G,N,T)
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 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 [11]. And in exhaustive search, every clause counts.

The soundness of clause complementation is a crucial property for the purpose of model checking; we again express it in terms of exclusivity:

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

- $\Gamma;\Delta;\mathcal{K} \Rightarrow G$ *and* $\Gamma;not^D(\Delta);\mathcal{K} \Rightarrow not^G(G)$*;*
- $\Gamma;\Delta;\mathcal{K} \xrightarrow{D} Q$ *and* $\Gamma;not^D(\Delta);\mathcal{K} \xrightarrow{not^D(D)} not^G(Q)$*.*

The proof follows the lines of [26].

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 [35], as suggested in [6]. 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 [15] 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 validation of the meta-theory of object languages, with significant success [20]. 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 [5]. In the appendix A we report some comparison with Isabelle/HOL's QuickCheck, but here we concentrate on *Nitpick* [6], a higher-order model finder in the *Alloy* lineage supporting (co)inductive definitions. It 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++ [7]. It would be natural to couple Isabelle/HOL's QuickCheck and/or Nitpick's capabilities with *Nominal* Isabelle [34], 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 on how much of the search space the three techniques explore, keeping in mind that what is traversed is very different in shape and 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.

However, our mission is finding counterexamples and so we compare the *time to find counterexamples* (TFCS) 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 a 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.

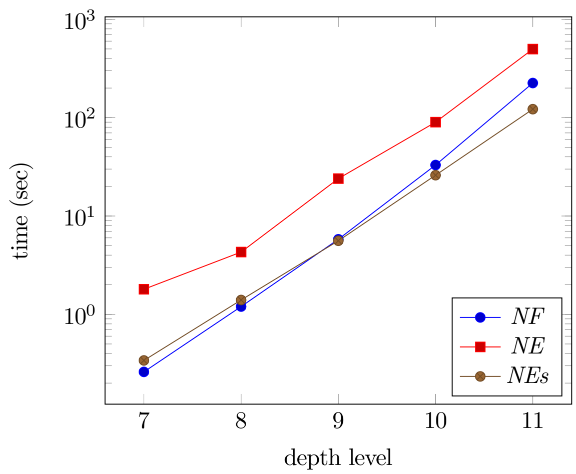

Fig. 4. Loglinear-plot of TESS on the progress theorem

The above results 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 of an order of magnitude (bugs 3 (ii) and 7). Further, *NEs* never under-performs *NE*, probably because it locates counterexample at at lower depth. In rare occasions (bug

bug	check	NF	NE	NEs	cex	Description/Class
1	pres	0.3 (7)	1 (7)	0.37 (7)	$(\lambda x.\ x\ err)\ n$	app rule: range of function
	prog	0 (5)	3.31 (9)	0.27 (5)	*hd n*	matched to the arg (S)
2	prog	0.27 (8)	t.o. (11)	85.3 (12)	*(cons n) nil*	value *(cons v) v* omitted (M)
3	pres	0.04 (6)	0.04 (6)	0.3 (6)	$(\lambda x.\ n)\ m$	order of types swapped
	prog	0 (5)	3.71 (9)	0.27 (8)	*hd n*	in function pos of app (S)
4	prog	t.o.	t.o.	t.o.	?	the type of cons is incorrect (S)
5	pres	t.o. (9)	t.o. (10)	41.5 (10)	*tl ((cons n) err)*	tail reduction returns the head (S)
6	prog	29.8 (11)	t.o. (11)	t.o. (12)	*hd ((cons n) nil)*	hd reduction on partially applied cons (M)
7	prog	1.04 (9)	18.5 (10)	1.1 (9)	$hd\ ((\lambda x.\ err)\ n)$	no evaluation for argument of app (M)
8	pres	0.02 (5)	0.03 (5)	0.1 (5)	$(\lambda x.\ x)\ nil$	lookup always returns int (U)
9	pres	0 (5)	0.02 (5)	0.1 (5)	$(\lambda x.\ y)\ n$	vars may not match in lookup (S)

Table 1. TFCE on the *Stlc* benchmark, Redex-style encoding

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 (2,6), 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 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 [20], Redex allows the user to write judgments such as typing in a declarative style, provided they can be given a functional mode, but more complex systems, such as typing rule for a polymorphic version of a similar calculus, require very indirect encoding, e.g. in CPS-style. We simulate addition on integers with natural numbers (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, i.e. no appeal to built-in arithmetic, 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, where constructors such as `hd` are *not* treated as constants, but are first class, e.g.:

```
tc(G,hd(E),intTy)        :- tc(G,E,listTy).
step(hd(cons(H,Tl)), H) :- value(H),value(Tl).
```

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

bug#	check	NF	NE	NEs	cex
1	pres	0.05 (5)	2.79 (5)	0.04 (5)	$(\lambda x.\ hd\ x)\ N$
2	prog	0 (4)	7.76 (9)	0.8 (7)	*(cons N) nil*
3	pres	0 (4)	0.05 (4)	0 (4)	$(\lambda x.\ nil)\ nil$
4	prog	0.15 (7)	t.o. (10)	199.1 (12)	*N + (cons N nil)*
5	pres	0(4)	0.04 (4)	0(4)	*tl (cons N) nil*
7	prog	5.82 (9)	151.2 (11)	19.54. (10)	$(\lambda x.\ nil)\ (N{+}M)$
8	pres	0.01 (4)	0.04 (4)	0.1 (4)	$(\lambda x.\ x)\ nil$
9	pres	0 (4)	0.04 (4)	0.1 (4)	$(\lambda x.\ y)\ N$

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

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 [35], whereby the basic imperative language *IMP* is endowed with a type system that prevents information flow from private to public variables[4]. 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 $\forall x.\ sec\ x < l \rightarrow \sigma_1(x) = \sigma_2(x)$. Assume a standard big-step evaluation semantics for IMP, relating an initial state σ and a command c to a final state τ:

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 $l \vdash_d c$ 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 *termination-sensitive* (syntax-directed) version $l \vdash_{\Downarrow} c$, where non-terminating programs do not leak information and its declarative cousin $l \vdash_{\Downarrow d} c$. 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:

st↔std $l \vdash c$ iff $l \vdash_d c$ and **sT↔sTd** $l \vdash_{\Downarrow} c$ iff $l \vdash_{\Downarrow d} c$.

[4] For an interesting case study regarding instead *dynamic* information flow and carried out in Haskell, see [19]. A large part of the paper is dedicated to the fine tuning of custom generators and shrinkers.

bug	check	Nitpick	NF	NE	NEs	Description
1	conf	(sp)	0.03 (5)	4.4 (8)	2.1 (7)	second premise of the seq rule omitted
	non-inter	t.o.	9.13 (8)	6.71 (8)	6.1 (8)	ditto
2	non-inter	(sp)	3.3 (8)	2.1 (8)	1.9 (8)	var swapping in $\leq$ premise of assn rule
3	st→std	0.95	t.o.	t.o	t.o.	inversion of $\leq$ in antimono rule
	std→st	0.75	0.8 (7)	0.3 (7)	0.3 (7)	ditto
4	st→std					$\leq$ assumption omitted in IF: **theorem holds**
	std→st	1.3	0.9 (7)	t.o.	t.o.	ditto
5	st→std	5.1(sp)	24.5 (11)	t.o.	t.o.	same as 2 but on decl version of the rule
	std→st	1.1	0.2 (7)	t.o.	24.6 (11)	ditto
6	sT→sTd	5.1(sp)	t.o.	t.o.	t.o.	same as 2 but on term version of the rule
	sTd→sT	1.0	0.01 (5)	0.32 (7)	0.05 (6)	ditto
7	sT→sTd					same as 4 but on term-decl rule: **theorem holds**
	sTd→sT	1.6	1.7 (8)	12.5 (9)	1.2(8)	ditto

Table 3. αCheck vs. Nitpick on the Volpano benchmark suite. (sp) indicates that Nitpick produced a spurious counterexample.

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`),*sec* `x=0,l=1` and σ maps 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,l=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 levels so as to find the relevant bugs.

The comparison with Nitpick[5] 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 $\vdash_{\Downarrow d}$, 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 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

[5] Settings: `[sat_solver=MiniSat_JNI,max_threads=1,check_potential,timeout = 200]`

settings. Yet, we could not select more binding intensive examples due to the current difficulties on running Nitpick with *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 similar 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" [18] 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*, for example using *isabelle mutabelle*.

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 [16]. In more pragmatical terms, providing generators and reflection mechanism for built-in datatypes and associated operators is a priority.

Finally, we'd like to implement improvements in nominal equational unification algorithms, which would make subsumption complete, *via equivariant* unification [10], 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.

A Some formal definitions

The effect of a permutation π on a name:

$$\mathsf{id}(\mathsf{a}) = \mathsf{a}$$
$$((\mathsf{a}\ \mathsf{b})\circ\pi)(\mathsf{c}) = \begin{cases} \mathsf{b} & \pi(\mathsf{c})=\mathsf{a} \\ \mathsf{a} & \pi(\mathsf{c})=\mathsf{b} \\ \pi(\mathsf{c}) & \pi(\mathsf{c})\notin\{\mathsf{a},\mathsf{b}\} \end{cases}$$

The swapping operation *ground* terms:

$$\begin{aligned} \pi\cdot\langle\rangle &= \langle\rangle & \pi\cdot f(t) &= f(\pi\cdot t) \\ \pi\cdot\langle t,u\rangle &= \langle\pi\cdot t,\pi\cdot u\rangle & \pi\cdot\mathsf{a} &= \pi(\mathsf{a}) \\ \pi\cdot\langle\mathsf{a}\rangle t &= \langle\pi\cdot\mathsf{a}\rangle\pi\cdot t \end{aligned}$$

Constraint satisfaction:

$$\begin{aligned} \theta &\models \top \\ \theta \models t\approx u &\iff \theta(t)\approx\theta(u) \\ \theta \models t\,\#\,u &\iff \theta(t)\,\#\,\theta(u) \\ \theta \models C\wedge C' &\iff \theta\models C \text{ and } \theta\models C' \\ \theta \models \exists X{:}\tau.\ C &\iff \text{for some } t:\tau,\ \theta[X:=t]^6\models C \\ \theta \models \text{И}\mathsf{a}{:}\nu.\ C &\iff \text{for some } \mathsf{b}\,\#\,(\theta,C),\ \theta\models C[\mathsf{b}/\mathsf{a}] \end{aligned}$$

A context Γ is a sequence of bindings between variables (or names) and types.

$$\Gamma ::= \cdot \mid \Gamma,X{:}\tau \mid \Gamma\#\mathsf{a}{:}\nu$$

where we write name-bindings as $\Gamma\#\mathsf{a}{:}\nu$, to remind us that a must be fresh for other names and variables in Γ.
Term complementation:

$$\begin{aligned} \mathit{not}[[\tau]] &: \tau\to\tau\ \mathit{set} \\ \mathit{not}[[\tau]](t) &= \emptyset \qquad\qquad \text{when } \tau\in\{\mathbf{1},\nu,\langle\nu\rangle\tau\} \text{ or } t \text{ is a variable} \\ \mathit{not}[[\tau_1\times\tau_2]](t_1,t_2) &= \{(s_1,\_)\mid s_1\in\mathit{not}[[\tau_1]](t_1)\}\cup\{(\_,s_2)\mid s_s\in\mathit{not}[[\tau_2]](t_2)\} \\ \mathit{not}[[\delta]](f(t)) &= \{g(\_)\mid g\in\Sigma, g:\sigma\to\delta, f\ \neq g\}\cup\{f(s)\mid s\in\mathit{not}[[\tau]](t)\} \end{aligned}$$

The correctness of the algorithm for term complementation can be stated in the following constraint-conscious way, as required by the proof of the main soundness theorem:

Lemma 1 (Term Exclusivity).
Let $\mathcal{K}$ be consistent, $s\in\mathit{not}[[\tau]](t)$, $FV(u)\subseteq\Gamma$ and $FV(s,t)\subseteq\boldsymbol{X}$. It is not the case that both $\Gamma;\mathcal{K}\models\exists\boldsymbol{X}{:}\boldsymbol{\tau}.\ u\approx t$ and $\Gamma;\mathcal{K}\models\exists\boldsymbol{X}{:}\boldsymbol{\tau}.\ u\approx s$.

Inequality and non-freshness:

$$
\begin{array}{rcl}
neq[[\tau]] & : & \tau \times \tau \to o \\
neq[[\mathbf{1}]](t,u) & = & \bot \\
neq[[\tau_1 \times \tau_2]](t,u) & = & neq[[\tau_1]](\pi_1(t),\pi_1(u)) \vee neq[[\tau_2]](\pi_2(t),\pi_2(u)) \\
neq[[\delta]](t,u) & = & neq_\delta(t,u) \\
neq[[\langle\nu\rangle\tau]](t,u) & = & \text{И}\mathsf{a}{:}\nu.\ neq[[\tau]](t@\mathsf{a},u@\mathsf{a}) \\
neq[[\nu]](t,u) & = & t \# u \\
neq_\delta(t,u) & :- & \bigvee\{\exists X,Y{:}\tau.\ t \approx f(X) \wedge u \approx f(Y) \wedge neq[[\tau]](X,Y) \\
& & \qquad \mid f : \tau \to \delta \in \Sigma\} \\
& & \vee \bigvee\{\exists X{:}\tau, Y{:}\tau'.\ t \approx f(X) \wedge u \approx g(Y) \\
& & \qquad \mid f : \tau \to \delta, g : \tau' \to \delta \in \Sigma, f \neq g\}
\end{array}
$$

$$
\begin{array}{rcl}
nfr[[\nu,\tau]] & : & \nu \times \tau \to o \\
nfr[[\nu,\mathbf{1}]](a,t) & = & \bot \\
nfr[[\nu,\tau_1 \times \tau_2]](a,t) & = & nfr[[\nu,\tau_1]](a,\pi_1(t)) \vee nfr[[\nu,\tau_2]](a,\pi_2(t)) \\
nfr[[\nu,\delta]](a,t) & = & nfr_{\nu,\delta}(a,t) \\
nfr[[\nu,\langle\nu'\rangle\tau]](a,t) & = & \text{И}\mathsf{b}{:}\nu'.\ nfr[[\tau]](a,t@\mathsf{b}) \\
nfr[[\nu,\nu]](a,b) & = & a \approx b \\
nfr[[\nu,\nu']](a,b) & = & \bot \quad (\nu \neq \nu') \\
nfr_{\nu,\delta}(a,t) & :- & \bigvee\{\exists X{:}\tau.\ t \approx f(X) \wedge nfr[[\nu,\tau]](a,X) \mid f : \tau \to \delta \in \Sigma\}
\end{array}
$$

B Other experiments

Random testing has been present in Isabelle/HOL's since [4] and has been recently enriched with a notion of *smart* test generators to improve its success rate w.r.t. conditional properties. Exhaustive and symbolic testing follow the SmallCheck approach [33]. Notwithstanding all these improvements, QuickCheck requires all code and specs to be *executable* in the underlying functional language, while many of the specifications that we are interested in are best seen as *partial* and *not terminating*.

While not terribly exciting, these benchmarks, proposed and measured in [9] and taken from Isabelle *List.thy* theory are useful to set up a rough comparison with Isabelle's QuickCheck. We show the checks in our logic programming formulation, leaving to the reader the obvious meaning, noting only that we use numerals as datatype.

```
D1: distinct([X|XS]) => distinct(XS).
D2: distinct(XS),remove1(X,XS,YS) => distinct(YS).
D3: distinct(XS),distinct(YS),zip(XS,YS,ZS) => distinct(ZS).
S1: sorted(XS),remove_dupls(XS,YS) => sorted(YS).
S2: sorted(XS),insert(X,XS,YS) => sorted(YS).
```

		9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
`D1`	S	0	0	0	0.2	0.7	3.8	22	135	862								
	NF	0	0	0	0	0	0	0	0	0	0.07	0.12	0.2	0.32	0.52	0.83	1.36	2.22
	NE	0	0	0	0	0	0	0	0	0	0.06	0.11	0.18	0.3	0.49	1.8	1.3	2.1
	NEs	0	0	0	0	0	0	0	0	0	0.06	0.11	0.18	0.3	0.4	0.6	1.0	1.7
`D2`	S	0	0	0.1	0.4	2.5	16	98	671									
	NF	0	0	0	0	0	0	0	0	0	0	0.07	0.19	0.32	0.51	0.83	1.36	2.23
	NE	0	0	0	0	0	0	0	0	0	0.6	0.11	0.18	0.3	0.49	0.8	1.32	2.17
	NEs	0	0	0	0	0	0	0	0	0	0.6	0.11	0.18	0.2	0.39	0.6	1.1	1.7
`D3`	S	4.3	157															
	NF	0	0	0	0.08	0.14	0.35	0.76	1	3	6	12	24	45	82	155	286	580
	NE	0	0	0	0.08	0.13	0.32	0.68	1.3	3	6	11	22	42	79	150	280	586
	NEs	0	0	0	0.08	0.13	0.22	0.5	0.9	2.1	4.5	8	17	3	63	121	225	448
`S1`	S	0	0	0	0	0	0	0	0	0.10	0.2	0.3	0.8	1.7	3.6	7.8	17	36
	NF	0	0	0	0	0	0	0	0	0	0	0.6	0.08	0.11	0.15	0.21	0.27	0.35
	NE	0	0	0	0	0	0	0	0	0	0	0.06	0.08	0.11	0.15	0.2	0.27	0.36
	NEs	0	0	0	0	0	0	0	0	0	0	0	0.04	0.06	0.08	0.11	0.16	0.2
`S2`	S	0	0	0	0	0	0.1	0.1	0.2	0.5	1.1	2.5	5.5	12	28	61	135	292
	NF	0	0	0	0	0	0	0	0	0	0	0	0.05	0.07	0.1	0.13	0.18	0.23
	NE	0	0	0	0	0	0	0	0	0	0.06	0.08	0.11	0.15	0.19	0.25	0.33	0.44
	NEs	0	0	0	0	0	0	0	0	0	0.02	0.04	0.04	0.06	0.08	0.11	0.16	0.2
`S3`	S	0	0	0	0	0.1	0.1	0.2	0.4	0.9	2.2	5.1	12	26	59	136	311	708
	NF	0	0	0.05	0.08	0.13	0.2	0.32	0.48	0.73	1	1.5	2.2	3.2	4.5	6.4	8.9	12
	NE	0	0	0	0.05	0.08	0.12	0.18	0.27	0.4	0.57	0.83	1.1	1.6	2.2	3.2	4.3	5.7
	NEs	0	0	0	0	0	0	0.04	0.09	0.1	0.28	0.4	0.5	0.8	1.1	1.5	2.1	2.9

Table 4. TESS for list benchmark.

```
S3: sorted(XS),length(XS,N),less_equal(I,J),less(J,N),
                              nth(I,XS,X),nth(J,XS,Y) => less_equal(X,Y).
```

Table B shows the TESS run time up to a given size (25), that in our case we interpret as depth-bound. We extrapolated from Table 2 in [9] the *S* (for *smart generator*) rows. We omit the results for *exhaustive* and *narrowing-based* testing; the point of their inclusion was to show how smart generation outperforms the latter two over checks with hard-to-satisfy premises. Again, these measurements are only suggestive, since QuickCheck's result are taken with another hardware (empty cells denote timeout after 1h as in [9]'s setup). Still, we are largely superior, possibly due to smart generation trying to replicate in a functional setting what logic programming naturally offers. Note however that tests in Isabelle/QuickCheck are efficiently run by code generation at the ML level, while our bounded solver is just a non-optimized logic programming interpreter – to name one, it does not have yet first-argument indexing.

As usual in TESS, negation elimination tends to outperform *NF*, especially when, as here, it does not require extensional quantification. *NEs* only marginally improves on *NE*, because the negated predicates (`distinct,sorted` etc.) are already quite simple.

References

1. D. Aspinall, L. Beringer, and A. Momigliano. Optimisation validation. *Electron. Notes Theor. Comput. Sci.*, 176(3):37–59, July 2007.
2. D. Baelde, A. Gacek, D. Miller, G. Nadathur, and A. Tiu. The Bedwyr system for model checking over syntactic expressions. In F. Pfenning, editor, *CADE*, volume 4603 of *Lecture Notes in Computer Science*, pages 391–397. Springer, 2007.
3. R. Barbuti, P. Mancarella, D. Pedreschi, and F. Turini. A transformational approach to negation in logic programming. *J. of Log. Program.*, 8:201–228, 1990.
4. S. Berghofer and T. Nipkow. Random testing in Isabelle/HOL. In *SEFM*, pages 230–239. IEEE Computer Society, 2004.
5. J. C. Blanchette, L. Bulwahn, and T. Nipkow. Automatic proof and disproof in Isabelle/HOL. In C. Tinelli and V. Sofronie-Stokkermans, editors, *FroCoS*, volume 6989 of *Lecture Notes in Computer Science*, pages 12–27. Springer, 2011.
6. J. C. Blanchette and T. Nipkow. Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In M. Kaufmann and L. Paulson, editors, *ITP 2010*, volume 6172 of *LNCS*, pages 131–146. Springer, 2010.
7. J. C. Blanchette, T. Weber, M. Batty, S. Owens, and S. Sarkar. Nitpicking C++ concurrency. In P. Schneider-Kamp and M. Hanus, editors, *Proceedings of the 13th International ACM SIGPLAN Conference on Principles and Practice of Declarative Programming*, pages 113–124. ACM, 2011.
8. J. Breitner. Formally proving a compiler transformation safe. In *Proceedings of the 2015 ACM SIGPLAN Symposium on Haskell*, Haskell '15, pages 35–46, New York, NY, USA, 2015. ACM.
9. L. Bulwahn. Smart testing of functional programs in Isabelle. In N. Bjørner and A. Voronkov, editors, *LPAR*, volume 7180 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2012.
10. J. Cheney. Equivariant unification. *Journal of Automated Reasoning*, 45(3):267–300, 2010.
11. J. Cheney and A. Momigliano. Mechanized metatheory model-checking. In M. Leuschel and A. Podelski, editors, *PPDP*, pages 75–86. ACM, 2007.
12. J. Cheney, A. Momigliano, and M. Pessina. Appendix to Advances in property-based testing for αProlog. `http://momigliano.di.unimi.it/alphaCheck.html`, 2016.
13. J. Cheney and C. Urban. Nominal logic programming. *ACM Transactions on Programming Languages and Systems*, 30(5):26, August 2008.
14. K. Claessen and J. Hughes. QuickCheck: a lightweight tool for random testing of Haskell programs. In *Proceedings of the 2000 ACM SIGPLAN International Conference on Functional Programming (ICFP 2000)*, pages 268–279. ACM, 2000.
15. M. Felleisen, R. B. Findler, and M. Flatt. *Semantics Engineering with PLT Redex*. The MIT Press, 2009.
16. B. Fetscher, K. Claessen, M. H. Palka, J. Hughes, and R. B. Findler. Making random judgments: Automatically generating well-typed terms from the definition of a type-system. In J. Vitek, editor, *ESOP 2015, ETAPS 2015. Proceedings*, volume 9032 of *Lecture Notes in Computer Science*, pages 383–405. Springer, 2015.
17. J. Harland. Success and failure for hereditary Harrop formulae. *J. Log. Program.*, 17(1):1–29, 1993.
18. Q. Heath and D. Miller. A framework for proof certificates in finite state exploration. In C. Kaliszyk and A. Paskevich, editors, *Proceedings Fourth Workshop on Proof eXchange for Theorem Proving, PxTP 2015, Berlin, Germany, August 2-3, 2015.*, volume 186 of *EPTCS*, pages 11–26, 2015.

19. C. Hritcu and co authors. Testing noninterference, quickly. In *ICFP*, pages 455–468. ACM, 2013.
20. C. Klein and co authors. Run your research: on the effectiveness of lightweight mechanization. In *POPL '12*, pages 285–296. ACM, 2012.
21. J.-L. Lassez and K. Marriott. Explicit representation of terms defined by counter examples. *Journal of Automated Reasoning*, 3(3):301–318, Sept. 1987.
22. J. Leach, S. Nieva, and M. Rodríguez-Artalejo. Constraint logic programming with hereditary Harrop formulas. *TPLP*, 1(4):409–445, July 2001.
23. X. Leroy. Formal verification of a realistic compiler. *Commun. ACM*, 52(7):107–115, 2009.
24. X. Leroy and H. Grall. Coinductive big-step operational semantics. *Inf. Comput.*, 207(2):284–304, 2009.
25. W. M. McKeeman. Differential testing for software. *Digital Technical Journal*, 10(1):100–107, 1998.
26. A. Momigliano. Elimination of negation in a logical framework. In P. Clote and H. Schwichtenberg, editors, *CSL*, volume 1862 of *Lecture Notes in Computer Science*, pages 411–426. Springer, 2000.
27. A. Momigliano and F. Pfenning. Higher-order pattern complement and the strict lambda-calculus. *ACM Trans. Comput. Log.*, 4(4):493–529, 2003.
28. T. Nipkow and G. Klein. *Concrete Semantics - With Isabelle/HOL*. Springer, 2014.
29. S. Owre. Random testing in PVS. In *Workshop on Automated Formal Methods (AFM)*, 2006.
30. M. H. Palka, K. Claessen, A. Russo, and J. Hughes. Testing an optimising compiler by generating random lambda terms. In *Proceedings of the 6th International Workshop on Automation of Software Test*, AST '11, pages 91–97, New York, NY, USA, 2011. ACM.
31. Z. Paraskevopoulou, C. Hritcu, M. Dénès, L. Lampropoulos, and B. C. Pierce. Foundational property-based testing. In C. Urban and X. Zhang, editors, *Interactive Theorem Proving - 6th International Conference, ITP 2015, Proceedings*, volume 9236 of *Lecture Notes in Computer Science*, pages 325–343. Springer, 2015.
32. A. M. Pitts. Nominal logic, a first order theory of names and binding. *Information and Computation*, 183:165–193, 2003.
33. C. Runciman, M. Naylor, and F. Lindblad. Smallcheck and lazy SmallCheck: automatic exhaustive testing for small values. In A. Gill, editor, *Haskell Workshop*, pages 37–48. ACM, 2008.
34. C. Urban and C. Kaliszyk. General bindings and alpha-equivalence in Nominal Isabelle. *Logical Methods in Computer Science*, 8(2), 2012.
35. D. Volpano, C. Irvine, and G. Smith. A sound type system for secure flow analysis. *J. Comput. Secur.*, 4(2-3):167–187, Jan. 1996.
36. J. Ševčík, V. Vafeiadis, F. Zappa Nardelli, S. Jagannathan, and P. Sewell. CompCertTSO: A verified compiler for relaxed-memory concurrency. *J. ACM*, 60(3):22:1–22:50, June 2013.
37. X. Yang, Y. Chen, E. Eide, and J. Regehr. Finding and understanding bugs in c compilers. In *PLDI '11*, pages 283–294, New York, NY, USA, 2011. ACM.