Abstract
Comparing provers on a formalization of the same problem is always a valuable exercise. In this paper, we present the formal proof of correctness of a non-trivial algorithm from graph theory that was carried out in three proof assistants: Why3, Coq, and Isabelle.

1 Introduction
Graph algorithms are notoriously obscure in the sense that it is hard to grasp why exactly they work. Therefore proof of correctness are more than welcome in this domain. In this paper, we consider Tarjan’s algorithm [28] for discovering the strongly connected components in a directed graph and present a formal proof of its correctness in three different systems: Why3, Coq and Isabelle. The algorithm is treated at an abstract level with a functional programming style manipulating finite sets, stacks and mappings, but it respects the linear time behaviour of the original presentation.

To our knowledge this is the first time that the formal correctness proof of a non-trivial program is carried out in three very different proof assistants: Why3 is based on a first-order logic with inductive predicates and automatic provers, Coq on an expressive theory of higher-order logic and dependent types, and Isabelle/HOL combines higher-order logic with automatic provers. We claim that our proof is direct, readable, elegant, and follows Tarjan’s presentation. Crucially for our comparison, the algorithm is defined at the same level of abstraction in all three systems, and the proof relies on the same arguments in the three formal systems. Note that a similar exercise but for a much more elementary proof (the irrationality of square root of 2) and using many more proof assistants (17) was presented in [32].

Formal and informal proofs of algorithms about graphs were already performed in [24, 30, 25, 13, 17, 29, 19, 27, 26, 15, 8]. Some of them are part of a larger library, others focus on the treatment of pointers or on concurrent algorithms. In particular, only Lammich and...
Neumann [17] gave an alternative formal proof of Tarjan’s algorithm within their framework for verifying graph algorithms in Isabelle/HOL. We expose here the key parts of the proofs. The interested reader can access the details of the proofs and run them on the web [7, 9, 20]. In this paper, we recall the principles of the algorithm in section 2; we describe the proofs in the three systems in sections 3, 4, and 5 by emphasizing the differences induced by the logics which are used; we conclude in sections 6 and 7 by commenting the developments and advantages of each proof system.

2 The algorithm

In a directed graph, two vertices $x$ and $y$ are strongly connected if there exists a path from $x$ to $y$ and a path from $y$ to $x$. A strongly connected component (scc) is a maximal set of vertices where all pairs of vertices are strongly connected. A fundamental property relates sccs and depth-first search (DFS) traversal in a directed graph: each scc is a prefix of a single subtree in the corresponding spanning forest (see figure 1c). Its root is named the base of the scc. Tarjan’s algorithm [28] relies on the detection of these bases and collects the sccs in a pushdown stack. It performs a single DFS traversal of the graph assigning a serial number $\text{num}[x]$ to any vertex $x$ in the order of the visit. It computes the following function for every vertex $x$:

$$\text{LOWLINK}(x) = \min\{\text{num}[y] \mid x \xrightarrow{*} z \xleftarrow{\rightarrow} y \land x \text{ and } y \text{ are in the same scc}\}$$

The relation $x \xrightarrow{*}$ means that $z$ is a son of $x$ in the spanning forest, the relation $\xrightarrow{\rightarrow}$ is its transitive and reflexive closure, and $z \xleftarrow{\rightarrow}$ means that there is a cross-edge from $z$ to $y$ in the spanning forest (a cross-edge is an edge of the graph which is not an edge in the spanning forest). In figure 1c, $\xrightarrow{*}$ is drawn in thick lines and $\xleftarrow{\rightarrow}$ in dotted lines; in figure 1b the table of the values of the $\text{LOWLINK}$ function is shown. The minimum of the empty set is assumed to be $+\infty$ (this is a slight simplification w.r.t. the original algorithm).

The base $x$ of an scc is found when $\text{LOWLINK}(x) \geq \text{num}[x]$, and the component is formed by the nodes of the subtree rooted at $x$ and pruned of the sccs already discovered in that subtree. Notice that $\text{LOWLINK}(x)$ need neither be the lowest serial number of a vertex accessible from $x$, nor of an ancestor of $x$ in the spanning forest. Take for instance, vertices 8 or 9 in figure 1c. Moreover, the DFS traversal sets to $+\infty$ the serial numbers of vertices in already discovered sccs. The definition of $\text{LOWLINK}$ can therefore be written as:

$$\text{LOWLINK}(x) = \min\{\text{num}[y] \mid x \xrightarrow{\rightarrow} z \xrightarrow{\rightarrow} y\}$$

Our implementation of graphs uses an abstract type $\text{vertex}$ for vertices, a constant $\text{vertices}$ for the finite set of all vertices in the graph, and a $\text{successors}$ function from vertices to their adjacency set. The algorithm maintains an environment $e$ implemented as a record of type $\text{env}$ with four fields: a stack $e.\text{stack}$, a set $e.\text{sccs}$ of strongly connected components, a fresh serial number $e.\text{sn}$, and a function $e.\text{num}$ from vertices to serial numbers.

The DFS traversal is organized by two mutually recursive functions $\text{dfs1}$ and $\text{dfs}$. The function $\text{dfs1}$ visits a new vertex $x$ and computes $\text{LOWLINK}(x)$. Furthermore it adds a new scc when $x$ is the base of a new scc. The function $\text{dfs}$ takes as argument a set $r$ of roots and
Figure 1 The vertices are numbered and pushed onto the stack in the order of their visit by the recursive function dfs1. When the first component \{0\} is discovered, vertex 0 is popped; similarly when the second component \{5, 6, 7\} is found, its vertices are popped; finally all vertices are popped when the third component \{1, 2, 3, 4, 8, 9\} is found. Notice that there is no cross-edge to a vertex with a number less than 5 when the second component is discovered. Similarly in the first component, there is no edge to a vertex with a number less than 0. In the third component, there is no edge to a vertex less than 1 since we have set the serial number of vertex 0 to \(+\infty\) when 0 was popped.

In the body of dfs1, the auxiliary function add_stack_incr updates the environment by pushing \(x\) on the stack, assigning it the current fresh serial number, and incrementing that number in view of future calls. The function dfs1 performs a recursive call to dfs for the adjacent vertices of \(x\) as roots and the updated environment. If the returned integer value \(n1\) is less than the number \(n0\) assigned to \(x\), the function simply returns \(n1\) and the current environment. Otherwise, the function declares that a new SCC has been found, consisting of all vertices that are contained on top of \(x\) in the current stack. Therefore the stack is popped until \(x\); the popped vertices are stored as a new set in e.sccs; and their serial numbers are all set to \(+\infty\), ensuring that they do not interfere with future calculations of min values. The auxiliary functions split and set_infty are used to carry out these updates.

let add_stack_incr x e =
{stack = Cons x e.stack; sccs = e.sccs; sn = n+1; num = e.num[x ← n]}
4 Formal proofs of Tarjan’s SCC algorithm

```plaintext
let rec set_infty s f = match s with Nil → f
| Cons x s' → (set_infty s' f)[x ← +∞] end
let rec split x s = match s with Nil → (Nil, Nil)
| Cons y s' → if x = y then (Cons x Nil, s')
else let (s1', s2) = split x s' in (Cons y s1', s2) end
```

Figure 1 illustrates the behavior of the algorithm by an example. We presented the algorithm as a functional program, using data structures available in the Why3 standard library [3]. For lists we have the constructors Nil and Cons; the function elements returns the set of elements of a list. For finite sets, we have the empty set empty, and the functions add to add an element to a set, remove to remove an element from a set, choose to pick an arbitrary element in a (non-empty) set, and is_empty to test for emptiness. We also use maps with functions const denoting the constant function, _[_] to access the value of an element, and _[←_] for creating a map obtained from an existing map by setting an element to a given value.

For a correspondence between our presentation and the imperative programs used in standard textbooks, the reader is referred to [8]. The present version can be directly translated into Coq or Isabelle functions, and it respects the linear running time behaviour of the algorithm, since vertices could be easily implemented by integers, +∞ by the cardinal of vertices, finite sets by lists of integers and mappings by mutable arrays (see for instance [7]).

Thus for each environment e in the algorithm, the working stack e.stack corresponds to a cut of the spanning forest where strongly connected components to its left are pruned and stored in e.sccs. In this stack, any vertex can reach any vertex higher in the stack. And if a vertex is a base of an scc, no cross-edge can reach some vertex lower than this base in the stack, otherwise that last vertex would be in the same scc with a strictly lower serial number.

We therefore have to organize the proofs of the algorithm around these arguments. To maintain these invariants we will distinguish, as is common for DFS algorithms, three sets of vertices: white vertices are the non-visited ones, black vertices are those that are already fully visited, and gray vertices are those that are still being visited. Clearly, these sets are disjoint and white vertices can be considered as forming the complement in vertices of the union of the gray and black ones.

The previously mentioned invariant properties can now be expressed for vertices in the stack: no such vertex is white, any vertex can reach all vertices higher in the stack, any vertex can reach some gray vertex lower in the stack. Moreover, vertices in the stack respect the numbering order, i.e. a vertex x is lower than y in the stack if and only if the number assigned to x is strictly less than the number assigned to y.

3 The proof in Why3

The Why3 system comprises the programming language WhyML used in previous section and a many sorted first-order logic with inductive data types and inductive predicates to express the logical assertions. The system generates proof obligations w.r.t. the assertions, pre- and post-conditions and lemmas inserted in the WhyML program. The system is interfaced with off-the-shelf automatic provers and interactive proof assistants.

From the Why3 library, we use pre-defined theories for integer arithmetic, polymorphic lists, finite sets and mappings. There is also a small theory for paths in graphs. Here we define graphs, paths and sccs as follows.

```plaintext
axiom successors_vertices: ∀x. mem x vertices → subset (successors x) vertices
predicate edge (x y: vertex) = mem x vertices ∧ mem y (successors x)
inductive path vertex (list vertex) vertex =
```
where \textit{mem} and \textit{subset} denote membership and the subset relation for finite sets.

We add two ghost fields in environments for the black and gray sets of vertices. These fields are used in the proofs but not used in the calculation of the SCCs, which is checked by the type-checker of the language.¹

The functions now become:

```ocaml
let rec dfs1 x e =
  let n0 = e.sn in
  let (n1, e1) = dfs (successors x) (add_stack_incr x e) in
  if n1 < n0 then (n1, add_black x e1) else
  let (s2, s3) = split x e1.stack in
  (+∞, {black = add x e1.black; gray = e.gray; stack = s3; sccs = add (elements s2) e1.sccs; sn = e1.sn; num = set_infty s2 e1.num})
with dfs r e = ...

let tarjan () =
  let e = {black = empty; gray = empty; stack = Nil; sccs = empty; sn = 0; num = const (-1)} in
  let (_, e') = dfs vertices e in e'.sccs
```

with a new function \textit{add\_black} turning a vertex from gray to black and the modified \textit{add\_stack\_incr} adding a new gray vertex with a fresh serial number to the current stack.

```ocaml
let add_black x e =
  {black = add x e.black; gray = remove x e.gray; stack = e.stack; sccs = e.sccs; sn = e.sn; num = e.num}
let add_stack_incr x e =
  let n = e.sn in
  {black = e.black; gray = add x e.gray; stack = Cons x e.stack; sccs = e.sccs; sn = n+1; num = e.num[x ← n]}
```

The main invariant \((I)\) of our program states that the environment is well-formed:

```ocaml
predicate wf_env (e: env) =
  let {stack = s; black = b; gray = g} = e in
  wf_color e ∧ wf_num e ∧ simpleset s ∧ no_black_to_white b g ∧
  (∀x y. lmem x s → lmem y s → e.num[x] ≤ e.num[y] → reachable x y) ∧
  (∀y. lmem y s → ∃x. mem x g ∧ e.num[x] ≤ e.num[y] ∧ reachable y x) ∧
  (∀cc. mem cc e.sccs ↔ subset cc b ∧ is_scc cc)
```

where \textit{lmem} stands for membership in a list. The well-formedness property is the conjunction of seven clauses. The two first clauses express elementary conditions about the colored sets of vertices and the numbering function (see [7, 8] for a detailed description). The third clause states that there are no repetitions in the stack, and the fourth that there is no edge from a

¹ In Why3-1.2.0, this check is performed differently
black vertex to a white vertex. The next two clauses formally express the property already
stated above: any vertex in the stack reaches all higher vertices and any vertex in the stack
can reach a lower gray vertex. The last clause states that the \texttt{sccs} field is the set of all sccs
all of whose vertices are black.

Since at the end of the \texttt{tarjan} function, all vertices are black, the \texttt{sccs} field will contain
exactly the set of all strongly connected components.

\begin{verbatim}
let tarjan () = return \{r \rightarrow \forall cc. mem cc r \iff subset cc vertices \land is_scc cc\}
  let e = \{black = empty; gray = empty; 
  stack = Nil; sccs = empty; sn = 0; num = const (-1)\} in
  let (_, e') = dfs vertices e in
  assert \{subset vertices e'.black\};
  e'.sccs
\end{verbatim}

Our functions \texttt{dfs1} and \texttt{dfs} modify the environment in a monotonic way. Namely
they augment the set of visited vertices (the black ones); they keep invariant the set of the ones
currently under visit (the gray set); they increase the stack with new black vertices; they
also discover new sccs and they keep invariant the serial numbers of vertices in the stack,

\begin{verbatim}
predicate subenv (e e’: env) = 
  subset e.black e’.black \land e.gray == e’.gray
\land (\exists s. e’.stack = s ++ e.stack \land subset (elements s) e’.black)
\land subset e.sccs e’.sccs \land (\forall x. lmem x e.stack \rightarrow e.num[x] = e’.num[x])
\end{verbatim}

Once these invariants are expressed, it remains to locate them in the program text and
to add assertions which help to prove them. The pre-conditions of \texttt{dfs1} are quite natural:
the vertex \(x\) must be a white vertex of the graph, and it must be reachable from all gray
vertices. Moreover invariant (I) must hold. The post-conditions of \texttt{dfs1} are of three kinds.
 Firstly (I) and the monotony property \texttt{subenv} hold in the resulting environment. Vertex
\(x\) is black at the end of \texttt{dfs1}. Finally we express properties of the integer value \(n\) returned
by this function which should be \texttt{LOWLINK}(\(x\)) as noted previously. In this proof, we give
three implicit properties for characterizing \(n\). First, the returned value is never higher than
the number of \(x\) in the final environment. Secondly, the returned value is either \(+\infty\) or the
number of a vertex in the stack reachable from \(x\). Finally, if there is an edge from a vertex \(y’\)
in the new part of the stack to a vertex \(y\) in its old part, the resulting value \(n\) must be lower
or equal to the serial number of \(y\).

\begin{verbatim}
let rec dfs1 x e =
  (* pre-condition *)
  requires \{mem x vertices \land not mem x (union e.black e.gray)\}
  requires \{\forall y. mem y e.gray \rightarrow reachable y x\}
  requires \{wf_env e\}  (* I *)
  (* post-condition *)
  return \{(\_, e’) \rightarrow wf_env e’ \land subenv e e’\}
  return \{(\_, x) \rightarrow mem x e’.black\}
  return \{(\_, e’) \rightarrow n = e’.num[x]\}
  return \{(n, e’) \rightarrow n = +\infty \lor \text{num_of_reachable_in_stack} n x e’\}
  return \{(n, e’) \rightarrow \forall y. xedge_to e’.stack e.stack y \rightarrow n \leq e’.num[y]\}
\end{verbatim}

The auxiliary predicates used above are formally defined in the following way.

\begin{verbatim}
predicate num_of_reachable_in_stack (n: int) (x: vertex)(e: env) = 
  \exists y. lmem y e.stack \land n = e.num[y] \land reachable x y
predicate xedge_to (s1 s3: list vertex) (y: vertex) = 
  (\exists s2. s1 = s2 ++ s3 \land \exists y’. lmem y’ s2 \land edge y’ y) \land lmem y s3
\end{verbatim}

Notice that the definition of \texttt{xedge_to} fits the definition of \texttt{LOWLINK} when the cross edge
ends at a vertex residing in the stack before the call of \texttt{dfs1}. The pre- and post-conditions
for the function \(dfs\) are quite similar up to a generalization to sets of vertices considered as
the roots of the algorithm (see [7]).

We now add seven assertions in the body of the \(dfs1\) function to help the automatic
provers. In contrast, the function \(dfs\) needs no extra assertions in its body. In \(dfs1\), when the
number \(n0\) of \(x\) is strictly greater than the number \(n1\) resulting from the call to its successors,
the first assertion states that \(n1\) cannot be \(+\infty\); it helps proving the next assertion. The
second assertion states that a lower gray vertex is reachable from \(x\) and that thus the scc of
\(x\) is not fully black at end of \(dfs1\). In that assertion the inequality \(y \neq x\) is redundant, but
helps showing the \(sccs\) constraint at the end of \(dfs1\). When \(n1 \geq n0\), the next four assertions
show that the strongly connected component \(s2\) of \(x\) is on top of \(x\) in the current
stack and that then \(x\) is the base of that scc. The seventh assertion helps proving that the
coloring constraint is preserved at the end of \(dfs1\).

```
let n0 = e.sn in
let (n1, e1) = dfs (successors x) (add_stack_incr x e) in
if n1 < n0 then begin
  assert(n1 \neq +\infty);
  assert\(\{y. y \neq x \land \text{mem } e1\gray \land e1\num[y] < e1\num[x] \land \text{in_same_scc } x y\};
  (n1, add_black x e1)
else begin
  let (s2, s3) = split x e1.stack in
  assert\(\{\text{is_last } x s2 \land s3 = e1\stack \land \text{subset } (\text{elements } s2) (\text{add } x e1\black)\};
  assert\(\{\text{is_subsc } (\text{elements } s2)\};
  assert\(\{\text{in_same_scc } y \rightarrow \text{mem } y s2\};
  assert\(\{\text{is_scc } (\text{elements } s2)\};
  assert\(\{\text{inter } e\gray (\text{elements } s2) = \text{empty}\};
  \(+\infty, \{\text{black } = \text{add } x e1\black; \text{gray } = e1\gray; \text{stack } = s3;\)
  \{sccs = \text{add } (\text{elements } s2) e1\sccs; sn = e1.sn; num = set_infty s2 e1.num\}
end

where \(\text{inter}\) is set intersection, and \(\text{is_last}\) is defined below.
```

```
predicate is_last (x: α) (s: list α) = \exists s'. s = s' ++ Cons x Nil
```

All proofs are discovered by the automatic provers except for two proofs carried out
interactively in Coq. One is the proof of the black extension of the stack in case \(n1 < n0\).
The provers could not work with the existential quantifier, although the Coq proof is quite
short. The second Coq proof is the fifth assertion in the body of \(dfs1\), which asserts that any
\(y\) in the scc of \(x\) belongs to \(s2\). It is a maximality assertion which states that the set \(\text{elements}\)
\(s2\) is a complete scc. The proof of that assertion is by contradiction. If \(y\) is not in \(s2\), there
must be an edge from \(x'\) in \(s2\) to some \(y'\) not in \(s2\) such that \(x'\) reaches \(y'\) and \(y'\) reaches \(y\).
There are three cases, depending on the position of \(y'\). Case 1 is when \(y'\) is in \(sccs\): this is
not possible since \(x\) would then be in \(sccs\) which contradicts \(x\) being gray. Case 2 is when \(y'\)
is an element of \(s'\): the serial number of \(y'\) is strictly less than the one of \(x\) which is \(n0\). If
\(x' \neq x\), the cross-edge from \(x'\) to \(y'\) contradicts \(n1 \geq n0\) (post-condition 5); if \(x' = x\), then \(y'\)
is a successor of \(x\) and again it contradicts \(n1 \geq n0\) (post-condition 3). Case 3 is when \(y'\)
is white, then \(x' \neq x\) is impossible since \(x'\) is then black in \(s2\) and would be the origin of a
black-to-white edge to \(y'\); if \(x' = x\), then \(y'\) is not white by post-condition 2 of \(dfs\).

Some quantitative information about the Why3 proof is listed in table 1. Alt-Ergo 2.3
and CVC4 1.5 proved the bulk of the proof obligations.\(^2\) The proof uses 49 lemmas that were
all proved automatically, but with an interactive interface providing hints to apply inlining,
splitting, or induction strategies. This includes 13 lemmas on sets, 16 on lists, 5 on lists

\(^2\) In addition to the results reported in the table, Spass was used to discharge one proof obligation.
Table 1 Performance results with provers in Why3-0.88.3 (in seconds, on a 3.3 GHz Intel Core i5 processor). Total time is 341.15 seconds. The two last columns contain the numbers of verification conditions and proof obligations. Notice that there may be several VCs per proof obligation.

without repetitions, 3 on paths, 5 on sccs and 7 very specialized lemmas directly involved in the proof obligations of the algorithm. Among the lemmas, a critical one is the lemma xpath_xedge on paths which reduces a predicate on paths to a predicate on edges. In fact, most of the Why3 proof works on edges which are handled more robustly by the automatic provers than paths. Another important lemma is subscc_after_last_gray which shows that the stack elements on top of the last gray vertex form a subset of an scc. This means that another program with the split call before the if-statement would make a simpler proof, but it would be a non-linear-time program. The two Coq proofs are only 9 and 81 lines long (the Coq files of 677 and 680 lines include preambles that are automatically generated during the translation from Why3 to CoQ). The interested reader is refered to [7] where the full proof is available.

The proof explained so far only showed the partial correctness of the algorithm. But after adding two lemmas about union and difference for finite sets, termination is automatically proved by the following lexicographic ordering on the number of white vertices and roots.

```coq
let rec dfs1 x e = variant {cardinal (diff vertices (union e.black e.gray)), 0}
with dfs r e = variant {cardinal (diff vertices (union e.black e.gray)), 1, cardinal r}
```

4 The proof in Coq

Coq is based on type theory and the calculus of constructions, a higher order lambda-calculus, for expressing formulae and proofs. Some basic notions of graph theory are provided by the Mathematical Components Library [18]. Our formalization is parameterized by a finite type $V$ for the vertices and the function successors such that successors $x$ is the adjacency set of any vertex $x$. The boolean $\text{gconnect } x \ y$ indicates that a path connects the vertex $x$ to the vertex $y$. It is straightforward to define the set gscs of the sccs using gconnect. Components are represented as sets of sets ($\{\text{set } \{V\}\}$). We use library operations for creating singletons ($\text{set } x$), taking unions ($S_1 \cup S_2$), differences ($S_1 \setminus S_2$), complements ($\sim S$), and unions of all sets of a set of sets (cover $S$).

Coq proposes several mechanisms to put together properties (boolean conjunction, propositional conjunction, record, inductive family) that have their own specificities. In order to make the presentation more readable for a non-CoQ expert, we write them all with the propositional conjunction $[\land P_1, \ldots \land P_n]$. We refer to [9] for the actual code.

The CoQ proof differs from the one in Why3: it uses natural numbers only and does not mention colors (white, gray and black). In particular, the number $\infty$ is defined as the cardinality of $V$, vertices with $\infty+1$ as serial number correspond to the white vertices of the previous section and the environment is defined as a record with only two fields, a set of sccs and the mapping assigning serial numbers to vertices:

```coq
let rec dfs1 x e = variant {cardinal (diff vertices (union e.black e.gray)), 0}
with dfs r e = variant {cardinal (diff vertices (union e.black e.gray)), 1, cardinal r}
```
Given an environment $e$, the set of visited vertices is 
visited $e$ (the vertices with serial number less or equal to $\infty$), the current fresh serial number is $sn\ e$ (the cardinal of visited vertices), and the stack is stack $e$ (the list of elements $x$ which satisfy $num\ e\ x < sn\ e$, sorted by increasing serial number).

Another difference with the Why3 algorithm is the disentanglement of the mutually recursive function tarjan into two separate functions. The first one dfs1 treats a vertex $x$ and the second one dfs a set of vertices roots in an environment $e$.


definition dfs1 dfs x e :=
  e1 e2 :=
  e :=
  k r e := if k is k

where visit x e produces the environment where $x$ gets the next serial number, store stores a new strongly connected component.

Then, the two functions are glued together in a recursive function rec where the parameter $k$ controls the maximal recursive height.

\[ \text{Fixpoint rec } k x e := \text{if } k \text{ is } k'.+1 \text{ then } \text{dfs} (\text{dfs} (\text{rec} k')) (\text{rec} k') r e \text{ else } (\infty, e). \]

If $k$ is not zero (i.e., it is a successor of some $k'$), rec calls dfs taking care that its parameters can only use recursive calls to rec with a smaller recursive height, here $k'$. This ensures termination. A dummy value is returned in the case where $k$ is zero. Finally, the top level tarjan calls rec with the proper initial arguments.

\[ \text{Definition tarjan := let: } (\text{rec } (\infty + \infty.+2) V (\text{Env } \emptyset \text{ ffun } \Rightarrow \infty.+1)) \text{ in esccs e.} \]

Initially, the roots are all the vertices ($V$) and the environment has no component and all vertices are not visited (their number is $\infty.+1$). As both dfs and dfs1 cannot be applied more than the number of vertices, the value $\infty \times \infty.+2$ encodes the lexicographic product of the two maximal heights. It gives rec enough fuel to never encounter the dummy value so tarjan correctly terminates the computation. This allows us to separate the proof of the termination from the algorithm itself, and this last statement is of course proved formally later and named rec_terminates.

The invariants of the Coq proof are usually shorter than in the Why3 proof since they do not mention colors. We first define well-formed environments and their valid extension:

\[ \text{Definition wf_env e := } \land\ esccs e \subseteq gscs, \]

\[ \forall x, \text{num } e x < \infty \rightarrow \text{num } e x < sn e, \]

\[ \forall x, (\text{num } e x = \infty) = (x \in \text{cover} (esccs e)) \land \]

\[ \forall x y, \text{num } e x \leq \text{num } e y < \text{sn } e \rightarrow \text{gconnect } x y]. \]

\[ \text{Definition subenv e1 e2 := } \land\ esccs e1 \subseteq esccs e2, \]

\[ \forall x, \text{num } e1 x < \infty \rightarrow \text{num } e2 x = \text{num } e1 x \land \forall x, \text{num } e2 x < \text{sn } e1 \rightarrow \text{num } e1 x < \text{sn } e1]. \]

Then we state that new visited vertices are the ones reachable by paths accessible from roots with non-visited vertices (i.e., white paths in the colored setting). The function nexts such that nexts $D$ $X$ returns the set of vertices reachable from the set $X$ by a path which only contains vertices in $D$ except maybe the last one.
The post-condition is the conjunction of these three properties and the characterization of the output rank:

\[
\text{dfs_spec}(\text{ne} : \text{nat} \times \text{env}) (\text{roots} : \{\text{set} \ V\}) e := \text{let} : (n, e') := \text{ne} \text{ in }
\]

\[
\land n = \min_{x \in \text{nxts}} (\neg \text{visited e} \land \text{roots e} \land \text{inord}(\text{num} e' x),
\]

\[
\text{wf_env e'}, \text{subenv e e'} \land \text{outenv roots e e'}.
\]

Here, the argument \(ne'\) is the result of a \(dfs\). The output rank \(n\) is the minimum of the serial numbers of the vertices which can be reached from the roots through a path where all the vertices except maybe the last one were not already visited. Note that this characterization differs from the notion of \(LOWLINK\) which requires that the last vertex was visited.

Finally, we express correctness as the implication between pre- and post-conditions:

\[
\text{dfs_correct} \text{dfs (roots : \{set \ V\}) e} := \text{wf_env e} \rightarrow
\]

\[
(\forall x y, x \in \text{stack e} \rightarrow y \in \text{roots e} \rightarrow \text{gconnect x y}) \rightarrow \text{dfs_spec (dfs roots e) roots e}.
\]

These invariants are expressed differently from the formulation in Why3, but they reflect essentially the same ideas. Rephrasing the invariants made it possible to reduce by approximately 50% the size of the Coq proofs. The two central theorems are:

\[
\text{Lemma dfsP} \text{dfs1 dfsrec (roots : \{set \ V\}) e} : (\forall x, x \in \text{roots e} \rightarrow \text{dfs_correct dfs1 x e}) \rightarrow
\]

\[
(\forall x, x \in \text{roots e} \rightarrow \forall e1, \text{subenv e e1} \rightarrow \text{dfs_correct dfsrec (roots \set x) e1}) \rightarrow
\]

\[
\text{dfs_correct (dfs dfs1 dfsrec) roots e}.
\]

\[
\text{Lemma dfs1P} \text{dfs x e : dfs_correct dfs (successors x) (visit x e) \rightarrow}
\]

\[
\text{dfs1_correct (dfs dfs1) x e}.
\]

They state that \(dfs\) and \(dfs1\) are correct if their respective recursive calls are correct. The proof of the first lemma is straightforward since \(dfs\) simply iterates on a list. It mostly requires book-keeping between what is known and what needs to be proved. This is done in about 54 lines. The second one is more intricate and requires 124 lines. Gluing these two theorems together and proving termination gives us an extra 12 lines to prove the theorem:

\[
\text{Theorem rec_terminates} k (\text{roots : \{set \ V\}}) e:
\]

\[
k \geq \#\neg\exists \text{visited e} = \infty \rightarrow \text{dfs_correct (rec k) roots e}.
\]

The correctness of \(tarjan\) follows directly in 19 lines of straightforward proof.

\[
\text{Theorem tarjan_correct} : \text{tarjan = gsccs}.
\]

We now provide some quantitative information. The Coq contribution is composed of two files. The \texttt{extra_nocolors} file defines the \texttt{bigmin} operator and some notions of graph theory that we intend to add to Mathematical Components. This file is 294 lines long. The main file is \texttt{tarjan_nocolors} and is 605 lines long. It is compiled in 12 seconds with a memory footprint of 800 Mb (3/4 of which is resident) on a Intel® i7 2.60GHz quad-core laptop running Linux. The proofs are performed in the SSReflect proof language [14] with very little automation. The proof script is mostly procedural, alternating book-keeping tactics \((move)\) with transformational ones (mostly \texttt{rewrite} and \texttt{apply}), but often intermediate steps
Table 2 Distribution of the numbers of lines of the 43 proofs in the file `tarjan_nocolors`.

<table>
<thead>
<tr>
<th>Number of lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>124</td>
</tr>
</tbody>
</table>

are explicitly declared with the `have` tactic. There are more than fifty of such intermediate steps in the 320 lines of proof of the file `tarjan_nocolors`. Table 2 gives the distribution of the numbers of lines of these proofs. Most of them are very short (26 are less than 2 lines) and the only complicated proof is the one corresponding to the lemma `dfs1P`.

5 The proof in Isabelle/HOL

Isabelle/HOL [21] is the encoding of simply typed higher-order logic in the logical framework Isabelle [23]. Unlike Why3, it is not primarily intended as an environment for program verification and does not contain specific syntax for stating pre- and post-conditions or intermediate assertions in function definitions. Logics and formalisms for program verification have been developed within Isabelle/HOL (e.g., [16]), but they target imperative rather than functional programming, so we simply formalize the algorithm as an Isabelle function. Isabelle/HOL provides an extensive library of data structures and proofs. In this development we mainly rely on the set and list libraries. We start by introducing a `locale`, fixing parameters and assumptions for the remainder of the proof. We explicitly assume that the set of vertices is finite.

```isabelle
goalie graph =
fixes vertices :: 'nu set and successors :: 'nu set
assumes finite vertices and
\forall v \in vertices. successors v \subseteq vertices
```

We introduce reachability in graphs using an inductive predicate definition, rather than via an explicit reference to paths as in the Why3 definition. Isabelle then generates appropriate induction theorems for use in proofs.

```isabelle
inductive reachable where
reachable x x
| \[ y \in successors x; reachable y z \] \implies reachable x z
```

The definition of strongly connected components mirrors that used in Why3. The following lemma states that SCCs are disjoint; its one-line proof is found automatically using Sledgehammer [2], which heuristically selects suitable lemmas from the set of available facts (including Isabelle’s library), invokes several automatic provers, and finally reconstructs a proof that is checked by the Isabelle kernel.

```isabelle
lemma scc-partition:
assumes is-scc S and is-scc S’ and x \in S \setminus S’
shows S = S’
```

Environments are represented by records, similar to the formalization in Why3, except that there is no distinction between regular and “ghost” fields. Also, the definition of the well-formedness predicate closely mirrors that used in Why3.\(^3\)

```isabelle
record 'nu env =
  black :: 'nu set  
  gray :: 'nu set
  stack :: 'nu list  
  sccs :: 'nu set set
  sn :: nat
  num :: 'nu \Rightarrow int
definition wf_env where wf_env e \equiv
  wf_color e \land wf_num e \land distinct (stack e) \land no_black_to_white e
```

\(^3\) We use the infix operator `\preceq` to denote precedence in lists.
The definition of the two mutually recursive functions \(dfs1\) and \(dfs\) again closely follows their representation in Why3.

\[
\text{function (domintros) dfs1 and dfs where} \\
\text{dfs1 x e =} \\
\text{let (n1,e1) = dfs (successors x) (add_stack_incr x e) in} \\
\text{if n1 < int (sn e) then (n1, add_black x e1) } \\
\text{else (let (l,r) = split_list x (stack e1) in} \\
\text{(+∞, (| black = insert x (black e1), gray = gray e,} \\
\text{ stack = r, sn = sn e1, sccs = insert (set l) (sccs e1),} \\
\text{ num = set_infty l (num e1) | )))) } \\
\text{and} \\
\text{dfs roots e =} \\
\text{(if roots = {} then (+∞, e) } \\
\text{else (let x = SOME x. x ∈ roots;} \\
\text{res1 = (if num e x \neq -1 then (num e x, e) else dfs1 x e);} \\
\text{res2 = dfs (roots - {x}) (snd res1) } \\
\text{in (min (fst res1) (fst res2), snd res2))} \\
\]

The \textbf{function} keyword introduces the definition of a recursive function. Isabelle checks that the definition is well-formed and generates appropriate simplification and induction theorems. Because HOL is a logic of total functions, it introduces two proof obligations: the first one requires the user to prove that the cases in the function definitions cover all type-correct arguments; this holds trivially for the above definitions. The second obligation requires exhibiting a well-founded ordering on the function parameters that ensures the termination of recursive function invocations, and Isabelle provides a number of heuristics that work in many cases. However, the functions defined above will in fact not terminate for arbitrary calls, in particular for environments that assign sequence number \(-1\) to non-white vertices.

The \texttt{domintros} attribute instructs Isabelle to consider these functions as “partial”. More precisely, it introduces an explicit predicate representing the domains for which the functions are defined. This “domain condition” appears as a hypothesis in the simplification rules that mirror the function definitions so that the user can assert the equality of the left- and right-hand sides of the definitions only if the domain predicate holds. Isabelle also proves (mutually inductive) rules for proving when the domain condition is guaranteed to hold. Our first objective is therefore to establish sufficient conditions that ensure the termination of the two functions. Assuming the domain condition, we prove that the functions never decrease the set of colored vertices and that vertices are never explicitly assigned the number \(-1\) by our functions. Denoting the union of gray and black vertices as \textbf{colored}, we introduce the predicate

\[
\text{definition colored_num where colored_num e ≡} \\
\forall v ∈ \text{colored e. } v ∈ \text{vertices} \land \text{num e v} \neq -1
\]

and show that this predicate is an invariant of the functions. We then prove that the triple defined as

\[
(\text{vertices - colored e, } \{x\}, 1) \\
(\text{vertices - colored e, roots, 2})
\]

for the arguments of \(dfs1\) and \(dfs\), respectively, decreases w.r.t. lexicographical ordering on finite subset inclusion and \(<\) on natural numbers across recursive function calls, provided that \textit{colored_num} holds when the function is called and \(x\) is a white vertex. These conditions
are therefore sufficient to ensure that the domain condition holds:\footnote{Observe that Isabelle introduces a single operator corresponding to the two mutually recursive functions whose domain is the disjoint sum of the domains of both functions.}

\begin{verbatim}
theorem dfs1_dfs_termination:
  \[(x : vertices - colored e; colored_num e) \implies dfs1_dfs_dom (Inl(x,e))\]
  \[(roots : vertices; colored_num e) \implies dfs1_dfs_dom (Inr(roots,e))\]
\end{verbatim}

The proof of partial correctness follows the same ideas as the proof presented for Why3. We define the pre- and post-conditions of the two functions as predicates in Isabelle. For example, the predicates for \texttt{dfs1} are defined as follows:

\begin{verbatim}
definition dfs1_pre where dfs1_pre e ≡
  wf_env e ∧ x ∈ vertices ∧ x /∈ colored e ∧ (∀g ∈ gray e. reachable g x)
definition dfs1_post where dfs1_post x e res ≡
  let n = fst res; e' = snd res
  in wf_env e' ∧ subenv e e' ∧ roots ⊆ colored e'
  ∧ (∀x ∈ roots. n ≤ num e' x)
  ∧ (n = +∞ ∨ (∃y in set (stack e'). num e' y = n ∧ reachable x y))
\end{verbatim}

We now show the following theorems:

- The pre-condition of each function establishes the pre-condition of every recursive call appearing in the body of that function. For the second recursive call in the body of \texttt{dfs} we also assume the post-condition of the first recursive call.
- The pre-condition of each function, plus the post-conditions of each recursive call in the body of that function, establishes the post-condition of the function.

Combining these results, we establish partial correctness:

\begin{verbatim}
theorem dfs_partial_correct:
  \[dfs1_dfs_dom (Inl(x,e)); dfs1_pre x e] =⇒ dfs1_post x e (dfs1 x e)
  \[dfs1_dfs_dom (Inr(roots,e)); dfs_pre roots e] =⇒ dfs_post roots e (dfs roots e)
\end{verbatim}

We define the initial environment and the overall function.

\begin{verbatim}
definition init_env where init_env ≡
  \{(black = {}, gray = {}, stack = [], sccs = {}, sn = 0, num = λ_. -1\}
definition tarjan where tarjan ≡
  sccs (snd (dfs vertices init_env))
\end{verbatim}

It is trivial to show that the arguments to the call of \texttt{dfs} in the definition of \texttt{tarjan} satisfy the pre-condition of \texttt{dfs}. Putting together the theorems establishing termination and partial correctness, we obtain the desired total correctness results.

\begin{verbatim}
theorem dfs_correct:
  dfs1_pre x e =⇒ dfs1_post x e (dfs1 x e)
  dfs_pre roots e =⇒ dfs_post roots e (dfs roots e)

theorem tarjan_correct:
  tarjan = \{ C . is_scc C ∧ C ⊆ vertices \}
\end{verbatim}

The intermediate assertions appearing in the Why3 code guided the overall proof: they are established either as separate lemmas or as intermediate steps within the proofs of the above theorems. Similarly to the Coq proof, the overall induction proof was explicitly decomposed into individual lemmas as laid out above. In particular, whereas Why3 identifies the predicates that can be used from the function code and its annotation with pre- and post-conditions, these assertions appear explicitly in the intermediate lemmas used in the
proof of theorem $dfs\_partial\_correct$. The induction rules that Isabelle generated from the function definitions were helpful for finding the appropriate decomposition of the overall correctness proof.

Despite the extensive use of Sledgehammer for invoking automatic back-end provers, including the SMT solvers CVC4 and Z3, from Isabelle, we found that in comparison to Why3, significantly more user interactions were necessary in order to guide the proof. Although many of those were straightforward, a few required thinking about how a given assertion could be derived from the facts available in the context. Table 3 indicates the distribution of the number of interactions used for the proofs of the 46 lemmas the theory contains. These numbers cannot be compared directly to those shown in Table 2 for the Coq proof because an Isabelle interaction is typically much coarser-grained than a line in a Coq proof. As in the case of Why3 and Coq, the proofs of partial correctness of $dfs1$ (split into two lemmas following the case distinction) and of $dfs$ required the most effort. It took about one person-month to carry out the case study, starting from an initial version of the Why3 proof. Processing the entire Isabelle theory on a laptop with a 2.7 GHz Intel® Core i5 (dual-core) processor and 8 GB of RAM takes 35 seconds of CPU time.

### 6 General comments about the proof

Our formal proofs refer to colors, finite sets, and the stack, although the informal correctness argument is about properties of strongly connected components in spanning trees. The algorithmic would explain the algorithm with spanning trees as in Tarjan’s article. It would be nice to extract a program from such a proof, but programmers like to understand the proof in terms of variables and data that their program is using.

A first version of the formal proof used ranks in the working stack and a flat representation of environments by adding extra arguments to functions for the black, gray, scc sets and the stack. That was perfect for the automatic provers of Why3. But after remodelling the proof in Coq and Isabelle/HOL, it was simpler to gather these extra arguments in records and have a single extra argument for environments. Also ranks disappeared in favor of the $num$ function and the precedence relation, which are easier to understand. The automatic provers have more difficulties with the inlining of environments, but with a few hints they could still succeed.

Our proof is mainly about the correctness of Tarjan’s algorithm. It relies on surprisingly few and elementary concepts of finite graphs. With the exception of the use of the Mathematical Components library for Coq, we therefore did not use existing libraries formalizing advanced concepts of graph theory [11, 22].

Finally, coloring of vertices is usual for graph algorithms. The stack used in our algorithm is also not necessary since it is just used to efficiently output new strongly connected components. The Coq formalization actually shows that proof can be done with just serial numbers and the store of connected components. The stack and current serial number could be added back using a program refinement, in order to recover a linear time computation.

There is always a tension between the concision of the proof, its clarity and its relation to the real program. In our presentation, we have allowed for a few redundancies.
The formal proof expressed in this article was initially designed and implemented in Why3 [8] as the result of a long process, nearly a 2-year half-time work with many attempts of proofs about various graph algorithms (depth first search, Kosaraju strong connectivity, bi-connectivity, articulation points, minimum spanning tree). Why3 has a clear separation between programs and the logic. It makes the correctness proof quite readable for a programmer. Also first-order logic is easy to understand. Moreover, one can prove partial correctness without caring about termination.

Another important feature of Why3 is its interface with various off-the-shelf theorem provers (mainly SMT provers). Thus the system benefits from the current technology in theorem provers and clerical sub-goals can be delegated to these provers, which makes the overall proof shorter and easier to understand. Although the proof must be split in more elementary pieces, this has the benefit of improving its readability. Several hints about inlining or induction reasoning are still needed and two Coq proofs were used. The system records sessions and facilitates incremental proofs. However, the automatic provers are sometimes no longer able to handle a proof obligation after seemingly minor modifications to the formulation of the algorithm or the predicates, making the proof somewhat unstable.

The Coq and Isabelle proofs were inspired by the Why3 proof. Their development therefore required much less time although their text is longer. The Coq proof uses SSReflect and the Mathematical Components library, which helps reduce the size of the proof compared to classical Coq. The proof also uses the bigops library and several other higher-order features which makes it more abstract and closer to Tarjan’s original proof.

In Coq, one could prove termination using well-foundedness [1, 4], but because of nested recursion the Function command fails, and both Equations and Program Fixpoint require the addition of an extra proof argument to the function. Instead, we define the functionals dfs1 and dfs and recombine them in rec and tarjan by recursion on a natural number used as fuel. We prove partial correctness on functionals and postpone termination on rec.

Our Coq proof does not use significant automation.\(^5\) All details are explicitly expressed, but many of them were already present in the Mathematical Components library. Moreover, a proof certificate is produced and a functional program could in principle be extracted. The absence of automation makes the system very stable to use since the proof script is explicit, but it requires a higher degree of expertise from the user.

The Isabelle/HOL proof can be seen as a mid-point between the Why3 and Coq proofs. It uses higher order logic and the level of abstraction is close to the one of the Coq proof.

\(^5\) Hammers exist for Coq [10, 12] but unfortunately they currently perform badly when used in conjunction with the Mathematical Components library.
although more readable in this case study. The proof makes use of Isabelle’s extensive support for automation. In particular, Sledgehammer [2] was very useful for finding individual proof steps. It heuristically selects lemmas and facts available in the context and then calls automatic provers (SMT solvers and superposition-based provers for first-order logic). When one of these provers finds a proof, Sledgehammer attempts to find a proof that can be certified by the Isabelle kernel, using various proof methods such as combinations of rewriting and first-order reasoning (blast, fastforce etc.), calls to the metis prover or reconstruction of SMT proofs through the smt proof method. Unlike in Why3, the automatic provers used to find the initial proof are not part of the trusted code base because ultimately the proof is checked by the kernel. The price to pay is that the degree of automation in Isabelle is still significantly lower compared to Why3. Adapting the proof to modified definitions was fast: the Isabelle/jEdit GUI eagerly processes the proof script and quickly indicates those steps that require attention.

The Isabelle proof also faces the termination problem to achieve general consistency. We chose to delay handling termination, using the domintros attribute. The proofs of termination and of partial correctness are independent; in particular, we obtain a weaker predicate ensuring termination than the one used for partial correctness. Although the basic principle of the termination proof is very similar to the Coq proof and relies on considering functionals of which the recursive functions are fixpoints, the technical formulation is more flexible because we rely on proving well-foundedness of an appropriate relation rather than computing an explicit upper bound on the number of recursive calls.

One strong point of Isabelle/HOL is its nice \LaTeX output and the flexibility of its parser, supporting mathematical symbols. Combined with the hierarchical Isar proof language [31], the proof is in principle understandable without actually running the system, although some familiarity with the system is still required.

In the end, the three systems Why3, Coq, and Isabelle/HOL are mature, and each one has its own advantages w.r.t. readability, expressivity, stability, ease of use, automation, partial-correctness, code extraction, trusted base and length of proof (see table 4). Coming up with invariants that are both strong enough and understandable was by far the hardest part in this work. This effort requires creativity and understanding, although proof assistants provide some help: missing predicates can be discovered by understanding which parts of the proof fail. We think that formalizing the proof in all three systems was very rewarding and helped us better understand the state of the art in computer-aided deductive program verification. It could be also interesting to implement this proof in other formal systems and establish comparisons based on this quite challenging example.6

Another interesting work would be to verify an implementation of this algorithm with imperative programs and concrete data structures. This will make the proof more complex, since mutable variables and mutable data structures have to be considered. There is support for verifying imperative programs in general-purpose proof assistants [5, 6, 16], and it would be interesting to also develop them simultaneously in various formal systems and to understand how these proofs can be derived from ours.

A final and totally different remark is about teaching of algorithms. Do we want students to formally prove algorithms, or to present algorithms with assertions, pre- and post-conditions, and make them prove these assertions informally as exercises? In both cases, we believe that our work could make a useful contribution.

---

6 We have set up a Web page http://www-sop.inria.fr/marelle/Tarjan/contributions.html in order to collect formalizations.
References


