Adapting Proof Automation to Adapt Proofs

Talia Ringer  
University of Washington, USA

John Leo  
Halfaya Research, USA

Nathaniel Yazdani  
University of Washington, USA

Dan Grossman  
University of Washington, USA

Abstract
We extend proof automation in an interactive theorem prover to analyze changes in specifications and proofs. Our approach leverages the history of changes to specifications and proofs to search for a patch that can be applied to other specifications and proofs that need to change in analogous ways.

We identify and implement five core components that are key to searching for a patch. We build a patch finding procedure from these components, which we configure for various classes of changes. We implement this procedure in a Coq plugin as a proof-of-concept and use it on real Coq code to change specifications, port definitions of a type, and update the Coq standard library. We show how our findings help drive a future that moves the burden of dealing with the brittleness of small changes in an interactive theorem prover away from the programmer and into automated tooling.

CCS Concepts • Software and its engineering → Software verification; Software evolution; Programming by example;

Keywords proof automation, proof repair, proof evolution

1 Introduction

Proof automation makes verification more accessible to programmers, but it is often intractable without programmer guidance. In interactive theorem proving (ITP), the programmer guides the proof search process. The guidance reduces the search space to make proof automation more tractable. This in turns helps the programmer, who does not have to manually verify the entire system.

Despite automation, programming in these proof assistants is brittle: Even a minor change to a definition or theorem can break many dependent proofs. This is a major source of development inefficiency in proof assistants based on dependent type theory [13, 29, 64].

Traditional proof automation does not consider how proofs, definitions, and theorems change over time. Instead, it is driven by the state of the current proof, sometimes with supplementary information from other proofs, definitions, and theorems (as in hint databases [3] and rippling [20]). This puts the burden of dealing with brittleness on the programmer.

We present a new approach to proof automation that accounts for breaking changes. In our approach, the programmer guides proof search by providing an example of how to adapt proofs to changes in definitions or theorems. A tool then generalizes the example adaptation into a reusable patch that the programmer can use to fix other broken proofs.

In doing so, we chart a path for a future that moves the burden of brittleness away from the programmer and into proof automation. Programmers typically address brittleness through design principles that make proofs resilient to change [13, 29, 64], or through program-specific proof automation [24]. These techniques, while useful, have limitations: Planning a verification effort around future change is challenging, and program-specific automation requires specialized knowledge. The programmer’s ability to anticipate likely changes determines the robustness of both techniques in the face of change. Even then, many breaking changes are outside of the programmer’s control. Updating proof assistant versions, for example, can break proofs regardless of planning or automation [55].

We identify a set of core components that are critical to searching an example for patches in Coq. We use the components to build a procedure for finding patches in a Coq plugin as a proof-of-concept. Case studies on real projects like CompCert [43] and the Coq standard library suggest that patches are useful for realistic scenarios, and that it is simple to compose the components to handle different classes of changes. We test the boundaries of the components on a suite of tests and show how our findings can help build the future we envision.
In summary, we contribute the following:

1. We identify a set of core components that are key to searching for patches.
2. We demonstrate that these core components are useful on real changes in Coq code.
3. We explore how to drive a future that moves the burden of change in ITP away from the programmer and into automated tooling.

2 Automation, Reimagined

Traditional proof automation considers only the current state of theorems, proofs, and definitions. This is a missed opportunity: Verification projects are rarely static. Like other software, these projects evolve over time.

Currently, the burden of change largely falls on programmers. This does not have to be true. Proof automation can view theorems, proofs, and definitions as fluid entities: When a proof or specification changes, a tool can search the difference between the old and new versions for a reusable patch that can fix broken proofs.

Status Quo Experienced Coq programmers use design principles and custom tactics to make proofs resilient to change. These techniques are useful for large proof developments, but they place the burden of change on the programmer. This can be problematic when change occurs outside of the programmer’s control.

Consider a commit from the upcoming Coq 8.7 release [49]. This commit redefined injection from integers to reals (Figure 1). This change broke 18 proofs in the standard library.

The Coq developer who committed the change fixed the broken proofs, then made an additional 12 commits to address the change in coq-contribs, a regression suite of projects that the Coq developers maintain as versions change. Many of these changes were simple. For example, the developer wrote a lemma that describes the change:

```coq
Lemma INR_IPR : ∀ p, INR (Pos.to_nat p) = IPR p.
```

The developer then used this lemma to fix broken proofs within the standard library. For example, one proof broke on this line:

```coq
rewrite Pos2Nat.inj_sub by trivial.
```

It succeeded with the lemma:

```coq
rewrite <- 3!INR_IPR, Pos2Nat.inj_sub by trivial.
```

These changes are outside-facing: Coq users have to make similar changes to their own proofs when they update to Coq 8.7. The Coq developer can update some tactics to account for this, but it is impossible to account for every tactic that users could use. Furthermore, while the developer responsible for the changes knows about the lemma that describes the change, the Coq user does not. The Coq user must determine how the definition has changed and how to address the change, perhaps by reading documentation or by talking to the developers.

Our Vision When a user updates Coq, a tool can determine that the definition has changed, then analyze changes in the standard library and in coq-contribs that resulted from the change in definition (in this case, rewriting by the lemma). It can extract a reusable patch from those changes, which it can automatically apply within broken user proofs. The user never has to consider how the definition has changed.

3 Generating Reusable Patches

We identify five components that are key to finding reusable patches. We implement these components in a prototype Coq plugin, which we call Pumpkin Patch (Proof Updater Mechanically Passing Knowledge Into New Proofs, Assisting The Coq Hacker), or Pumpkin.

We focus the Pumpkin prototype on the proof assistant Coq [2]. In Coq, each theorem is a type, and a proof of that theorem is a term that inhabits that type. Rather than write proof terms directly, users write proof scripts in a high-level tactic language; Coq then uses these scripts to guide search for a proof term.

The Pumpkin repository contains a detailed user guide. To use Pumpkin, the programmer modifies a single proof script to provide an example of how to adapt a proof to a change. Pumpkin generalizes the example adaptation into a reusable patch: a function that can be used to fix other broken proofs, which Pumpkin defines as a Coq term. We focus on the problem of finding these patches; we discuss how to extend Pumpkin to automatically apply patches it finds in Section 7.

We motivate the core components (Section 3.1) and describe how they compose to find patches (Section 3.2). We find patches for real code in Section 4, and we describe our implementation in Section 5.

---

1. http://github.com/uwplse/PUMPKIN-PATCH/tree/cpp18
3.1 Motivating the Core

We identify and isolate five core components critical to finding reusable patches:

1. **Semantic differencing** between terms
2. **Patch specialization** to arguments
3. **Patch abstraction** of arguments or functions
4. **Patch inversion** to reverse a patch
5. **Lemma factoring** to break a term into parts

The semantic differencing component finds the difference between two terms, which produces a candidate for a reusable patch. The other components modify the candidate to try to produce a patch. To motivate this workflow, consider using Pumpkin to search the proofs in Figure 2 for a patch between conclusions. We invoke the plugin using `old` and `new` as the example change:

```
Patch Proof old new as patch.
```

**Pumpkin** first determines the type that a patch from `new` to `old` should have. To determine this, it semantically differs the types and finds this goal type (line 2):

```
V n m p, n <= m -> m <= p -> n <= p + 1
```

It then breaks each inductive proof into cases and determines an intermediate goal type for the candidate. In the base case, for example, it differs the types and determines that a candidate between the base cases of `new` and `old` should have this type (lines 11 and 12):

```
(fun p0 : nat) n <= p0 + 1
```

It then differs the terms (line 13) for such a candidate:

```
fun p0 : nat, n <= p0 -> m <= p0 + 1
```

This candidate is close, but it is not yet a patch. This candidate maps base case to base case (it is applied to `n`), the patch should map conclusion to conclusion (it should be applied to `p`). **Pumpkin abstracts** this candidate by `n` (line 11), which lifts it out of the base case:

```
fun n0 : nat => n <= p0 + 1
```

**Semantic Differencing** The tool should be able to identify the semantic difference between terms. The semantic difference is the difference between two terms that corresponds to the difference between their types. Consider the base case terms in Figure 2 (line 13):

```
le_plus_trans n m 1 H : n <= m + 1
```

The semantic differencing component first identifies the difference in their types, or the **goal type**:

```
V n m p, n <= m -> m <= p -> n <= p + 1
```

It then finds a difference in terms that has that type:

```
fun (H : n <= p) => le_plus_trans n m 1 H
```

This is the **candidate** for a reusable patch that the other components modify to find a patch.

Differencing operates over terms and types. Differencing tactics is insufficient, since tactics and hints may mask patches (line 5). Furthermore, differencing is aware of the

---

Since this is a simple example, replaying an existing tactic happens to work. There are additional examples in the repository (Cex.v).
semantics of terms and types. Simply exploring the syntactic difference makes it hard to identify which changes are meaningful. For example, in the inductive case (line 14), the inductive hypothesis changes:

\[
\begin{align*}
&\ldots (\text{IH}e : n \leq m_0 + 1) \ldots \\
&\ldots (\text{IH}e : n \leq m_0) \ldots
\end{align*}
\]

However, the type of \text{IH}e changes for \textit{any} two inductive proofs over \textit{le} with different conclusions. A syntactic differencing component may identify this change as a candidate. Our semantic differencing component knows that it can ignore this change.

**Patch Specialization** The tool should be able to specialize a patch candidate to specific arguments as determined by the differences in terms. To find a patch for Figure 2, for example, \textsc{Pumpkin} must specialize the patch candidate to \texttt{P} to produce the final patch.

**Patch Abstraction** A tool should be able to abstract patch candidates of this form by the common argument:

\[
\begin{align*}
&\text{candidate : } P' \ t \to P \ t \\
&\text{candidate_abs : } \forall \ t_0, P' \ t_0 \to P \ t_0
\end{align*}
\]

and it should be able to abstract patch candidates of this form by the common function:

\[
\begin{align*}
&\text{candidate : } P' \ t' \to P \ t \\
&\text{candidate_abs : } \forall \ t_0, P' \ t_0 \to P \ t_0
\end{align*}
\]

This is necessary because the tool may find candidates in an applied form. For example, when searching for a patch between the proofs in Figure 2, \textsc{Pumpkin} finds a candidate in the difference of base cases. To produce a patch, \textsc{Pumpkin} must abstract the candidate by the argument \texttt{P}. Abstracting candidates is not always possible; abstraction will necessarily be a collection of heuristics.

**Patch Inversion** The tool should be able to invert a patch candidate. This is necessary to search for isomorphisms. It is also necessary to search for implications between propositionally equal types, since candidates may appear in the wrong direction. For example, consider two list lemmas (we write \texttt{length} as \texttt{len}):

\[
\begin{align*}
&\text{old : } \forall \ l', 1 \leq \text{len} (l' ++ 1) = \text{len} (l' + 1) \text{len} l
\\
&\text{new : } \forall \ l', 1 \leq \text{len} (l' ++ 1) = \text{len} (l' + 1) \text{len} (\text{rev} l)
\end{align*}
\]

If \textsc{Pumpkin} searches the difference in proofs of these lemmas for a patch from the conclusion of \texttt{new} to the conclusion of \texttt{old}, it may find a candidate \textit{backwards}:

\[
\begin{align*}
&\text{candidate' : } H : \text{old' : } 1 \leq \text{len} l' \to \text{new' : } 1 \leq \text{len} l
\\
&\text{candidate_abs : } \forall \ l, H \ l \ x \to \text{new' : } 1 \leq \text{len} l
\end{align*}
\]

The component can invert this to get the patch:

\[
\begin{align*}
&\text{patch : } H : \text{old' : } 1 \leq \text{len} l' \to \text{new' : } 1 \leq \text{len} l
\\
&\text{patch_abs : } \forall \ l, H \ l \ x \to \text{new' : } 1 \leq \text{len} l
\end{align*}
\]

We can then use this patch to port proofs. For example, if we add this patch to a hint database [3], we can port this proof:

\[
\begin{align*}
&\text{Theorem app_rev_len : } \forall \ l, l' \. \\
&\text{len} (\text{rev} (l' ++ 1)) = \text{len} (\text{rev} l) + \text{len} (\text{rev} l').
\end{align*}
\]

\[
\begin{align*}
&\text{Proof. \hspace{1cm}} \\
&\text{intros. rewrite rev_app_distr. apply old.✓}
\end{align*}
\]

to this proof:

\[
\begin{align*}
&\text{intros. rewrite rev_app_distr. apply new.✓}
\end{align*}
\]

Rewrites like candidate are \textit{invertible}: We can invert any rewrite in one direction by rewriting in the opposite direction. In contrast, it is not possible to invert the patch \textsc{Pumpkin} found for Figure 2. Inversion will necessarily sometimes fail, since not all terms are invertible.

**Lemma Factoring** The tool should be able to factor a term into a sequence of lemmas. This can help break other problems, like abstractions, into smaller subproblems. It is also necessary to invert certain terms. Consider inverting an arbitrary sequence of two rewrites:

\[
\begin{align*}
&t := \text{eq_ind_r G} \ldots (\text{eq_ind_r F} \ldots)
\end{align*}
\]

We can view \texttt{t} as a term that composes two functions:

\[
\begin{align*}
&g := \text{eq_ind_r G} \ldots \\
&f := \text{eq_ind_r F} \ldots \\
&t := g \circ f
\end{align*}
\]

The inverse of \texttt{t} is the following:

\[
\begin{align*}
&t' := f^{-1} \circ g^{-1}
\end{align*}
\]

To invert \texttt{t}, \textsc{Pumpkin} identifies the factors \{\texttt{f}; \texttt{g}\}, inverts each factor to \{\texttt{f}^{-1}; \texttt{g}^{-1}\}, then folds and applies the inverse factors in the opposite direction.

### 3.2 Composing Components

The components come together to form a proof patch finding procedure:

\[
\begin{align*}
&\text{Pseudocode: find_patch(term, term', direction)}
\\
&\text{1. diff types of term and term' for goals}
\\&\text{2. diff term and term' for candidates}
\\&\text{3. if there are candidates then}
\\&\text{4. factor, abstract, specialize, and/or invert candidates}
\\&\text{5. if there are patches then return patches}
\\&\text{6. if direction is forwards then}
\\&\text{7. find_patch(term', term, backwards) for candidates}
\\&\text{8. factor and invert candidates}
\\&\text{9. if there are patches then return patches}
\\&\text{10. return failure}
\end{align*}
\]

\textsc{Pumpkin} infers a \textit{configuration} from the example change. This configuration customizes the highlighted lines for an entire class of changes: It determines what to diff on lines 1 and 2, and how to use the components on line 4.

For example, to find a patch for Figure 2, \textsc{Pumpkin} used the configuration for changes in conclusions of two proofs that induct over the same hypothesis. Given two such proofs:

\[
\begin{align*}
&\forall \ x, H \ x \to P \ x
\\&\forall \ x, H \ x \to P^* \ x
\end{align*}
\]

\textsc{Pumpkin} searches for a patch with this type:

\[
\begin{align*}
&\forall \ x, H \ x \to P^* \ x \to P \ x
\end{align*}
\]
using this configuration:

1: diff conclusion types for goals
2: diff conclusion terms for candidates
3: if there are candidates then
4: abstract and then specialize candidates

Section 4 describes real-world examples that demonstrate more configurations.

4 Case Studies: Changes in the Wild

We used the PUMPKIN prototype to emulate three motivating scenarios from real-world code:

1. Updating definitions within a project (CompCert, Section 4.1)
2. Porting definitions between libraries (Software Foundations, Section 4.2)
3. Updating proof assistant versions (Coq Standard Library, Section 4.3)

The code we chose for these scenarios demonstrated different classes of changes. For each case, we describe how PUMPKIN configures the procedure to use the core components for that class of changes. Our experiences with these scenarios suggest that patches are useful and that the components are effective and flexible.

Identifying Changes We identified Git commits from popular Coq projects that demonstrated each scenario. These commits updated proofs in response to breaking changes. We emulated each scenario as follows:

1. Replay an example proof update for PUMPKIN
2. Search the example for a patch using PUMPKIN
3. Apply the patch to fix a different broken proof

Our goal was to simulate incremental use of a patch finding tool, at the level of a small change or a commit that follows best practices. We favored commits with changes that we could isolate. When isolating examples for PUMPKIN, we replayed changes from the bottom up, as if we were making the changes ourselves. This means that we did not always make the same change as the user. For example, the real change from Section 4.1 updated multiple definitions; we updated only one.

PUMPKIN is a proof-of-concept and does not yet handle some kinds of proofs. In each scenario, we made minor modifications to proofs so that we could use PUMPKIN (for example, using induction instead of destruction). PUMPKIN does not yet handle structural changes like adding constructors or parameters, so we focused on changes that preserve shape, like modifying constructors. We discuss supporting these features and whether they may necessitate other components in Section 7.

4.1 Updating Definitions

Coq programmers sometimes make changes to definitions that break proofs within the same project. To emulate this use case, we identified a CompCert commit [44] with a breaking change to int (Figure 3). We used PUMPKIN to find a patch that corresponds to the change in int. The patch PUMPKIN found fixed broken inductive proofs.

Replay We used the proof of unsigned_range as the example for PUMPKIN. The proof failed with the new int:

\[
\text{Theorem unsigned\_range:}\quad \forall (i : \text{int}), 0 \leq \text{unsigned} i < \text{modulus}.
\]

Proof.

\[
\text{intros}\ i. \ induction\ i\ using\ int\_ind;\ \text{auto.}
\]

\[
\text{x}
\]

We replayed the change to unsigned\_range:

\[
\text{intros}\ i. \ induction\ i\ using\ int\_ind. \ \text{simp. omega.}
\]

✓

Search We used PUMPKIN to search the example for a patch that corresponds to the change in int. It found a patch with this type:

\[
\forall z : \text{Z}, -1 < z < \text{modulus} \rightarrow 0 \leq z < \text{modulus}
\]

Apply After changing the definition of int, the proof of the theorem repr\_unsigned failed on the last tactic:

\[
\text{Theorem repr\_unsigned:}\quad \forall (i : \text{int}), \text{repr (unsigned} i) = i.
\]

Proof.

\[
\text{... apply 2mod\_small; auto.}
\]

\[
\text{x}
\]

Manually trying omega—the tactic which helped us in the proof of unsigned\_range—did not succeed. We added the patch that PUMPKIN found to a hint database. The proof of the theorem repr\_unsigned then went through:

\[
\text{... apply 2mod\_small; auto.}
\]

✓

4.1.1 Core Components

This scenario used the configuration for changes in constructors of an inductive type. Given such a change:

\[
\text{Inductive} T := \ldots | C : \ldots \rightarrow H \rightarrow T
\]

\[
\text{Inductive}\ T' := \ldots | C : \ldots \rightarrow H' \rightarrow T'
\]

PUMPKIN searches two inductive proofs of theorems:

\[
\forall (t : T), P\ t
\]

\[
\forall (t : T'), P\ t
\]

for an isomorphism\(^3\) between the constructors:

\[
\ldots \rightarrow H \rightarrow H'
\]

\[
\ldots \rightarrow H' \rightarrow H
\]

\[^3\]If PUMPKIN finds just one implication, it returns that.
The user can apply these patches within the inductive case that corresponds to the constructor c to fix other broken proofs that induct over the changed type. PUMPKIN uses this configuration for changes in constructors:

1: diff inductive constructors for goals
2: use all components to recursively search for changes in conclusions of the corresponding case of the proof
3: if there are candidates then
4: try to invert the patch to find an isomorphism

4.2 Porting Definitions

Coq programmers sometimes port theorems and proofs to use definitions from different libraries. To simulate this, we used PUMPKIN to port two solutions [10, 14] to an exercise in Software Foundations to each use the other solution’s definition of the fixpoint bin_to_nat (Figure 4). We demonstrate one direction; the opposite was similar.

Replay We used the proof of bin_to_nat_pres_incr from User A as the example for PUMPKIN. User A cut an inline lemma in an inductive case and proved it using a rewrite:

\[ \text{assert (} \forall a, S (a + S (a + 0)) = S (S (a + (a + 0)))) \].

\[ \ldots \text{rewrite } \text{<- plus_0_r} \ldots \text{rewrite } \rightarrow \text{plus_comm}. \]

When we ported User A’s solution to use User B’s definition of bin_to_nat, the application of this inline lemma failed. We changed the conclusion of the inline lemma and removed the corresponding rewrite:

\[ \text{assert (} \forall a, S (a + S a) = S (S (a + a))) \].

\[ \ldots \text{rewrite } \rightarrow \text{plus_comm}. \]

Search We used PUMPKIN to search the example for a patch that corresponds to the change in bin_to_nat. It found an isomorphism:

\[ \forall P, b, P (\text{bin_to_nat } b) \rightarrow P (\text{bin_to_nat } b + 0) \]

\[ \forall P, b, P (\text{bin_to_nat } b + 0) \rightarrow P (\text{bin_to_nat } b) \]

Apply After porting to User B’s definition, a rewrite in the proof of the theorem normalize_correctness failed:

\[ \text{Theorem normalize_correctness:} \]

\[ \forall b, \text{nat_to_bin } (\text{bin_to_nat } b) = \text{normalize } b. \]

\[ \text{Proof.} \]

\[ \quad \text{... rewrite } \rightarrow \text{plus_0_r} \cdot X \]

Attempting the obvious patch from the difference in tactics—rewriting by plus_n_0—failed. Applying the patch that PUMPKIN found fixed the broken proof:

\[ \quad \text{... apply patch_inv} \ldots \text{rewrite } \rightarrow \text{plus_0_r} \cdot X \]

In this case, since we ported User A’s definition to a simpler definition, PUMPKIN found a patch that was not the most natural patch. The natural patch would be to remove the rewrite, just as we removed a different rewrite from the example proof. This did not occur when we ported User B’s definition, which suggests that in the future, a patch finding tool may help inform novice users which definition is simpler: It can factor the proof, then inform the user if two factors are inverses. Tactic-level changes do not provide enough information to determine this; the tool must have a semantic understanding of the terms.

4.2.1 Core Components

This scenario used the configuration for changes in cases of a fixpoint. Given such a change:

\[ \text{Fixpoint } f \ldots := \ldots \mid g X \]

\[ \text{Fixpoint } f' \ldots := \ldots \mid g' X \]

PUMPKIN searches two proofs of theorems:

\[ \forall \ldots, P (f \ldots) \]

\[ \forall \ldots, P (f' \ldots) \]

for an isomorphism that corresponds to the change:

\[ \forall P, P X \rightarrow P X' \]

\[ \forall P, P X' \rightarrow P X \]

The user can apply these patches to fix other broken proofs about the fixpoint.

The key feature that differentiates these from the patches we have encountered so far is that these patches hold for all P; for changes in fixpoint cases, the procedure abstracts candidates by P, not by its arguments. PUMPKIN uses this configuration for changes in fixpoint cases:

1: diff fixpoint cases for goals
2: use all components to recursively search an intermediate lemma for a change in conclusions
3: if there are candidates then
4: specialize and factor the candidate
   abstract the factors by functions
   try to invert the patch to find an isomorphism

For the prototype, we require the user to cut the intermediate lemma explicitly and to pass its type and arguments. In the future, an improved semantic differencing component can infer both the intermediate lemma and the arguments: It can search within the proof for some proof of a function that is applied to the fixpoint.

\[ ^4 \text{User A uses } +; \text{ User B uses } +. \text{ For arbitrary } n, \text{ the term } 2 \ast n \text{ reduces to } n + (n + 0), \text{ which does not reduce any further}. \]
Adapting Proof Automation to Adapt Proofs

4.3 Updating Proof Assistant Versions

Coq sometimes makes changes to its standard library that break backwards-compatibility. To test the plausibility of using a patch finding tool for proof assistant version updates, we identified a breaking change in the Coq standard library [45]. The commit changed the definition of `divide` prior to the Coq 8.4 release (Figure 5). The change broke 46 proofs in the standard library. We used Pumpkin to find an isomorphism that corresponds to the change in `divide`. The isomorphism Pumpkin found fixed broken proofs.

**Replay** We used the proof of `mod_divide` as the example for Pumpkin. The proof broke with the new `divide`:

```coq
Theorem mod_divide:
  ∀ a b, b~≠0 -> (a mod b == 0 <-> (divide b a)).
Proof.
  ... rewrite (div_mod a b Hb) at 2.
```

We replayed changes to `mod_divide`:

```coq
... rewrite mul_comm. symmetry.
rewrite (div_mod a b Hb) at 2.
```

**Search** We used Pumpkin to search the example for a patch that corresponds to the change in `divide`. It found an isomorphism:

```coq
∀ r p q, p * r = q -> q = r * p
∀ r p q, r * p = p * r = q
```

**Apply** The proof of the theorem `Zmod_divides` broke after rewriting by the changed theorem `mod_divide`:

```coq
Theorem Zmod_divides:
  ∀ a b, b~≠0 -> (a mod b == 0 <-> (∃ c, a = b * c)).
Proof.
  ... split; intros (c,Hc); exists c; auto.
```

Adding the patches Pumpkin found to a hint database made the proof go through:

```coq
... split; intros (c,Hc); exists c; auto.
```

4.3.1 Core Components

This scenario used the configuration for changes in dependent arguments to constructors. Pumpkin searches two proofs that apply the same constructor to different dependent arguments:

```coq
... (C (P x)) ...
... (C (P x)) ...
```

for an isomorphism between the arguments:

```coq
∀ x, P x -> P x
∀ x, P x -> P x
```

The user can apply these patches to patch proofs that apply the constructor (in this case study, to fix broken proofs that instantiate `divide` with some specific `r`).

So far, we have encountered changes of this form as arguments to an induction principle; in this case, the change is an argument to a constructor. A patch between arguments to an induction principle maps directly between conclusions of the new and old theorem without induction; a patch between constructors does not. For example, for `divide`, we can find a patch with this form:

```coq
∀ x, P x -> P x
```

However, without using the induction principle for `exists`, we can’t use that patch to prove this:

```coq
(∃ x, P x) -> (∃ x, P x)
```

This changes the goal type that semantic differencing determines. Pumpkin uses this configuration for changes in constructor arguments:

1: `diff` constructor arguments for goals
2: `use all components` to recursively search those arguments for changes in conclusions
3: `if there are candidates then`
4: `abstract the candidate factor and try to invert` the patch to find an isomorphism

For the prototype, the model of constructors for the semantic differencing component is limited, so we ask the user to provide the type of the change in argument (to guide line 2). We can extend semantic differencing to remove this restriction.

5 Inside the Core

We have shown that the core components are useful for finding proof patches. This section describes our implementation of the components (Section 5.1) and the procedure (Section 5.2) so that it is possible to implement them in other systems. While our system is a very early prototype under active development, we have made the source code available on Github. Our prototype has no impact on the trusted computing base (Section 5.3).

5.1 Inside the Components

The interested reader can follow along in the repository.

5.1.1 Semantic Differencing

We implement semantic differencing over `trees`: Pumpkin compiles each proof term into a tree (`evaluation.ml`). In these trees, every node is a type context, and every edge is an extension to that type context with a new term. Correspondingly, type differencing (to identify goal types) compares nodes, and term differencing (to find candidates) compares edges.
The component (differencing.ml) uses these nodes and edges to prioritize semantically relevant differences. At the lowest level, it calls a primitive differencing function which checks if it can substitute one term within another term to find a function between their types.

The key benefit to this model is that it gives us a natural way to express inductive proofs, so that differencing can efficiently identify good candidates. Consider, for example, searching for a patch between conclusions of two inductive proofs of theorems about the natural numbers:

\[
\text{nat_ind } P \ldots (\text{fun } (IH : n0) \Rightarrow \ldots) : \forall n, P n
\]

In each case, the component diffs the terms in the dotted edges of the tree for nat_ind (Figure 6) to try to find a term that maps between conclusions of that case:

\[
P' \quad \emptyset \Rightarrow P \emptyset \quad (\text{base case candidate } \ast)
\]

\[
P' \quad (S n) \Rightarrow P \quad (S n) \quad (\text{inductive case candidate } \ast)
\]

The component also knows that the change in the type of \(IH\) is inconsequential (it occurs for any change in conclusion). Furthermore, it knows that \(IH\) cannot show up as a hypothesis in the patch, so it attempts to remove any occurrences of \(IH\) in any candidate.

When the component finds a candidate, it knows \(P'\) and \(P\) as well as the arguments \(\emptyset\) or \((S n)\). This makes it simple to query abstraction for the final patch:

\[
\forall n, P' n \Rightarrow P n
\]

The differencing component is lazy: It only compiles terms into trees one step at a time. It then expands each tree as needed to find candidates (expansion.ml). For example, consider searching two functions for a patch between conclusions:

\[
\text{fun } (t : T) \Rightarrow b
\]

\[
\text{fun } (t' : T) \Rightarrow b'
\]

Differencing introduces a single term of type \(\forall\) to a common environment, then expands and recursively diffs the bodies \(b\) and \(b'\) in that environment.

The tool always maintains pointers to easily switch between the tree and AST representations of the terms. This representation enables extensibility; we discuss a broad range of extensions in Sections 6.2 and 7.

5.1.2 Patch Specialization

Specialization (specialize.ml) takes a patch candidate and some arguments, all of which are Coq terms. It applies the candidate to the arguments, then it \(\beta\)-reduces [25] the result using Coq’s Reduction.nf_betaiota function. It is the job of the patch finding procedure to provide both the candidate and the arguments.

5.1.3 Patch Abstraction

Abstraction (abstraction.ml) takes a patch candidate, the goal type, and the function arguments or function to abstract. It first generalizes the candidate, wrapping it inside of a lambda from the type of the term to abstract. Then, it substitutes terms inside the body with the abstract term. It continues to do this until there is nothing left to abstract, then filters results by the goal type. Consider, for example, abstracting this candidate by \(m\):

\[
\text{fun } (H : n <= m) \Rightarrow \text{le_plus_trans n m } H
\]

\[
: n <= m -> n <= m + 1
\]

The generalization step wraps this in a lambda from some \(nat\), the type of \(m\):

\[
\text{fun } (n0 : nat) \Rightarrow
\]

\[
(\text{fun } (H : n <= m) \Rightarrow \text{le_plus_trans n m } H)
\]

\[
: \forall n0, n <= m -> n <= m + 1
\]

The substitution step replaces \(m\) with \(n0\):

\[
\text{fun } (n0 : nat) \Rightarrow
\]

\[
(\text{fun } (H : n <= n0) \Rightarrow \text{le_plus_trans n n0 } H)
\]

\[
: \forall n0, n <= n0 -> n <= n0 + 1
\]

Abstraction uses a list of abstraction strategies to determine what subterms to substitute. In this case, the simplest strategy works: The tool replaces all terms that are convertible to the concrete argument \(m\) with the abstract argument \(n0\), which produces a single candidate. Type-checking this candidate confirms that it is a patch.

In some cases, the simplest strategy is not sufficient, even when it is possible to abstract the term. It may be possible to produce a patch only by abstracting some of the subterms convertible to the argument or function (we show an example of this in Section 6.2), or the term may not contain any subterms convertible to the argument or function at all. We implement several strategies to account for this. The combinations strategy, for example, tries all combinations of substituting only some of the convertible subterms with the abstract argument. The pattern-based strategy substitutes subterms that match a certain pattern with a term that corresponds to that pattern.

It is the job of the patch finding procedure to provide the candidate and the terms to abstract. In addition, each configuration includes a list of strategies. The configuration for changes in conclusions, for example, starts with the simplest strategy, and moves on to more complex strategies only if
that strategy fails. This design makes abstraction simple to extend with new strategies and simple to call with different strategies for different classes of changes; we identify new strategies in Section 6.2.

5.1.4 Patch Inversion
Patch inversion (inverting.ml) exploits symmetry to try to reverse the conclusions of a candidate patch. It first factors the candidate using the factoring component, then calls the primitive inversion function on each factor, then finally folds the resulting list in reverse. The primitive inversion function exploits symmetry. For example, equality is symmetric, so the component can invert any application of \(\text{eq\_ind}\) or \(\text{eq\_ind\_r}\) (any rewrite). Indeed, \(\text{eq\_ind}\) and \(\text{eq\_ind\_r}\) are inverses, and are related by symmetry:

\[
\text{eq\_ind\_r} \ A \times \ P \ (H : \ P \ x) \ y \ (H0 : \ y = x) := \text{eq\_ind} \ x \ (\text{fun} \ y0 : \ A \Rightarrow \ P \ y0) \ H \ y \ (\text{eq\_sym} \ H0)
\]

If inversion does not recognize that the type is symmetric, it swaps subterms and type-checks the result to see if it is an inverse. In the future, it may be possible to use this swapping functionality in a reflexive application of the induction principle to infer symmetry properties like \(\text{eq\_sym}\) for other types. The component can then use these properties to generate reverse induction principles like \(\text{eq\_ind\_r}\). We further detail inversion in Section 6.2.

5.1.5 Lemma Factoring
The lemma factoring component (factoring.ml) searches within a term for its factors. For example, if the term composes two functions, it returns both factors:

\[
t : X \rightarrow Z \quad \text{(* term *)}
\]

\[
f : X \rightarrow Y, \ g : Y \rightarrow Z \quad \text{(* factors *)}
\]

In this case, the component takes the composite term and \(X\) as arguments. It first searches as deep as possible for a term of type \(X \rightarrow Y\) for some \(Y\). If it finds such a term, then it recursively searches for a term with type \(Y \rightarrow Z\). It maintains all possible paths of factors along the way, and discards any paths that cannot reach \(Z\).

The current implementation can handle paths with more than two factors, but it fails when \(Y\) depends on \(X\). Other components may benefit from dependent factoring; we leave this to future work.

5.2 Inside the Procedure
The implementation (patcher.ml) of the procedure from Section 3.2 starts with a preprocessing step which compiles the proof terms to trees (like the tree in Figure 6). It then searches for candidates one step at a time, expanding the trees when necessary.

The \textsc{Pumpkin} prototype exposes the patch finding procedure to users through the Coq command \texttt{Patch Proof}. \textsc{Pumpkin} automatically infers which configuration to use for the procedure from the example change. For example, to find a patch for the case study in Section 4.1, we used this command:

\texttt{Patch Proof Old.unsigned\_range unsigned\_range as patch.}

\textsc{Pumpkin} analyzed both versions of \texttt{unsigned\_range} and determined that a constructor of the \texttt{int} type changed (Figure 3), so it initialized the configuration for changes in constructors.

Internally, \textsc{Pumpkin} represents configurations as sets of options, which it passes to the procedure. The procedure uses these options to determine how to compose components (for example, whether to abstract candidates) and how to customize components (for example, whether semantic differencing should look for an intermediate lemma). To implement new configurations for different classes of changes, we simply tweak the options.

5.3 Trusted Computing Base
A common concern for Coq plugins is an increase in the trusted computing base. The Coq developers provide a safe plugin API in Coq 8.7 to address this [30]. Our prototype takes this into consideration: While \textsc{Pumpkin} does not yet support Coq 8.7, it only calls the internal Coq functions that the developers plan to expose in the safe API [40]. Furthermore, Coq type-checks terms that plugins produce. Since \textsc{Pumpkin} does not modify the type checker, it cannot produce an ill-typed term.

6 Testing Boundaries
In the case studies in Section 4, we showed how the core components are useful for real scenarios. In this section, we explore the boundary between what the \textsc{Pumpkin} prototype can and cannot handle. It is precisely this boundary that informs us how to improve the implementations of the core components.

To evaluate this boundary, we tested the core components of \textsc{Pumpkin} on a suite of 50 pairs of proofs (Section 6.1). We designed 11 of these pairs to succeed, then modified their proofs to produce the remaining 39 pairs that try to stress the core functionality of the tool. We learned the following from the pairs that tested \textsc{Pumpkin}’s limitations:

1. The failed pairs drive improvements.
   \textsc{Pumpkin} failed on 17 of 50 pairs. These pairs tell us how to improve the core. (Section 6.2)

2. The pairs unearth abstraction strategies.
   \textsc{Pumpkin} produced an exponential number of candidates in 5 of 50 pairs. New abstraction strategies can dramatically reduce the number of candidates. (Section 6.2)

3. \textsc{Pumpkin} was fast, and it can be faster.
   The slowest successful patch took 48 ms. The slowest failure took 7 ms. Simple changes can make \textsc{Pumpkin} more efficient. (Section 6.3)
6.1 Patch Generation Suite

We wrote a suite\(^7\) of 50 pairs of proofs. We wrote these proofs ourselves since searching for proof patches is a new domain, so there was no existing benchmark suite to work with. We used the following methodology:

1. Choose theorems old and new
2. Write similar inductive proofs of old and new
3. Modify the proof of old to produce more pairs
4. Search for patches from new to old
5. If possible, search for patches from old to new

In total, we chose 11 pairs of theorems old and new, and we wrote 50 pairs of proofs of those theorems.

For example, one pair of theorems old and new was a simplification of the auxiliary lemmas that we encountered in the case study in Section 4.2. For the first proof of old, we added a rewrite, like in the case study:

\[
\text{rewrite } \leftarrow \text{plus}_0. \text{rewrite } \rightarrow \text{plus}\_\text{comm.}
\]

For the second proof of old, we commuted the rewrites:

\[
\text{rewrite } \rightarrow \text{plus}\_\text{comm.} \quad \text{rewrite } \leftarrow \text{plus}_0.
\]

We then searched for patches in both directions, since the conclusions of old and new were propositionally equal.

Our goal was to determine what changes to proofs stress the components and how to use that information to drive improvements. We focused on differences in conclusions, the most supported configuration. Since Pumpkin operates over terms, we removed redundant proof terms, even if they were produced by different tactics. We controlled the first pair of proofs of each pair of theorems for features we had not yet implemented, like nested induction, changes in hypotheses, and abstracting omega terms. These features sometimes showed up in later proofs (for example, after moving a rewrite); we kept these proofs in the suite, since isolated changes to supported proofs that introduce unsupported features can inform future improvements.

6.2 Three Challenges

Pumpkin found patches for 33 of the 50 pairs. 28 of the 33 successes did not stress Pumpkin at all; Pumpkin found the correct candidate immediately and was able to abstract it in one try. The pairs that Pumpkin failed to patch and the successful pairs that stressed abstraction reveal key information about how to improve the core components. We walk through three examples. Future tools can use these as challenge problems to improve upon Pumpkin.

\[\text{fun}\ n\ m\ p\ \text{(H : n }\le\text{ m)}\ (H0 : m \le p) \Rightarrow\]

\[\text{le}\_\text{S}\ n\ p\ (* \text{... proof of stronger lemma }\ast)\]

\[
: \forall\ n\ m\ p,\ n \le m \rightarrow m \le p \rightarrow n \le S\ p
\]

**Figure 7.** Two proof terms old (left) and new (right) that contain the same proof of a stronger lemma.

A Challenge for Differencing  For one pair of proofs of theorems with propositionally equal conclusions (Figure 7), the differencing component failed to find candidates in either direction. These proofs both contain the same proof of a stronger lemma; Pumpkin found patches from this lemma to both old and new, but it was unable to find a patch between old and new. A patch may show up deep in the difference between le\_\text{plus\_trans} and le\_\text{S}, but even if we \(\delta\)-reduce (unfold the definition of \([25]\) le\_\text{plus\_trans}, this is not obvious:

\[
\text{le}\_\text{plus\_trans}\ n\ m\ p\ (H : n \le m) :=
\]

\[
(\text{fun }\ \text{lemma} : m \le p + m \Rightarrow\
\text{trans\_contra\_inv}\_\text{impl\_morphism}\n\text{PreOrder\_Transitive}\n(m + p)\\
\text{m}\n\text{lemma})\\
(\text{le}\_\text{add}\_\text{r}\ m\ p)
\]

This points to two difficulties in finding patches: Knowing when to \(\delta\)-reduce terms is difficult; exploring the appropriate time for reduction may produce patches for pairs that Pumpkin currently cannot patch. Furthermore, finding patches is more challenging when neither theorem has a conclusion that is as strong as possible.

A Challenge for Inversion  For one pair of proofs with propositionally equal conclusions, Pumpkin found a patch in one direction, but failed to invert it:

\[
\text{fun}\ n\ m\ p\ (_ : n \le m)\ (_ : m \le p)\ (H! : n \le p) \Rightarrow\
gt\_\text{le}\_\text{S}\ n\ (S\ p)\ (\text{le}\_\text{lt}\_n\_\text{Sm}\ n\ p\ H!)
: \forall\ n\ m\ p,\ n \le m \rightarrow m \le p \rightarrow n \le p \rightarrow S\ n \le S\ p
\]

The inversion component was unable to invert this term, even though an inverse does exist. To invert this, the component needs to know to \(\delta\)-reduce gt\_\text{le}\_\text{S}:

\[
\text{gt}\_\text{le}\_\text{S}\ n\ m\ :=
\]

\[
(\text{fun }\ (H : \forall\ n0\ m0,\ n0 \le m0 \Rightarrow S\ n0 \le m0) \Rightarrow H\ n\ m)
\]

It then needs to swap the hypothesis with the conclusion in \(H\) to produce the inverse:

\[
\text{gt}\_\text{le}\_\text{S}^{-1}\ n\ m\ :=
\]

\[
(\text{fun }\ (H : \forall\ n0\ m0,\ S\ n0 \le m0 \Rightarrow n0 \le m0) \Rightarrow H\ n\ m)
\]

Inversion currently swaps subterms when it is not aware of any symmetry properties about the inductive type. However, it does not know when to \(\delta\)-reduce function definitions. Furthermore, there are many possible subterms to swap; for inversion to know to only swap the subterms of \(H\), it must have a better understanding of the structure of the term. Both of these are ways to improve inversion.
A Challenge for Abstraction  Abstraction produced an exponential number of candidates when abstracting a patch candidate with this type:

\[ \forall n \ n_0, \]
\[ (fun \ m \Rightarrow n \leq max \ m \ n_0) \ n \Rightarrow \]
\[ (fun \ m \Rightarrow n \leq max \ n_0 \ m) \ n \]

The goal was to abstract by \( n \) and produce a patch with this type:

\[ \forall m_0 \ n \ n_0, \]
\[ n \leq max \ n_0 \ m_0 \Rightarrow \]
\[ n \leq max \ n_0 \ m_0 \]

The difficulty was in determining which occurrences of \( n \) to abstract. The component needed to abstract only the highlighted occurrences:

\[
\begin{align*}
\text{fun} & \ n \ n_0 \ (H_0 : n \leq max \ n_0 \ n) \Rightarrow \\
& \ eq\_ind\_r \\
& \ nat \\
& \ (max \ n_0 \ n) \\
& \ (fun \ n_1 \Rightarrow n \leq n_1) \\
& \ H_0 \\
& \ (max \ n \ n_0) \\
& \ (max\_\text{comm} \ n \ n_0)
\end{align*}
\]

The simplest abstraction strategy failed, and a more complex strategy succeeded only after producing exponentially many candidates. While this did not have a significant impact on time, this case gives rise to a new class of abstraction strategies: semantics-aware abstraction. In this case, we know from the type of the candidate and the type of \( eq\_ind\_r \) that these two hypothesis types are equivalent (similarly for the conclusion types):

\[ (fun \ m \Rightarrow n \leq max \ m \ n_0) \ n \]
\[ (fun \ n_1 \Rightarrow n \leq n_1) \ (max \ n_0 \ n) \]

The tool can search recursively for patches to find two patches that bridge the two equivalent types:

\[
\begin{align*}
p_1 & := \text{fun} \ n \Rightarrow max \ n_0 \ n \\
p_2 & := \text{fun} \ n \Rightarrow max \ n \ n_0
\end{align*}
\]

Then the candidate type is exactly this:

\[ \forall n \ n_0, \]
\[ (fun \ n_1 \Rightarrow n \leq n_1) \ (p_2 \ n) \Rightarrow \]
\[ (fun \ n_1 \Rightarrow n \leq n_1) \ (p_1 \ n) \]

Abstraction should thus abstract the highlighted subterms and the terms that have types constrained by those subterms. This produces a patch in one candidate:

\[
\begin{align*}
\text{fun} & \ n \ n_0 \ (H_0 : n \leq max \ n_0 \ n_0) \Rightarrow \\
& \ eq\_ind\_r \\
& \ nat \\
& \ (max \ n_0 \ n_0) \\
& \ (* \ p_1 \ n_0 \ *) \\
& \ (fun \ n_1 \Rightarrow n \leq n_1) \\
& \ (* \ P \ *) \\
& \ H_0 \\
& \ (* \ : \ P \ (p_1 \ n_0) \ *) \\
& \ (max \ n_0 \ n_0) \\
& \ (* \ p_2 \ n_0 \ *) \\
& \ (max\_\text{comm} \ n_0 \ n_0) \\
& \ (* \ : \ p_1 \ n_0 = p_2 \ n_0 \ *)
\end{align*}
\]

This same strategy would find a patch for one of the pairs that \textsc{Pumpkin} failed to abstract. This is a natural future direction for abstraction.

6.3 Performance

\textsc{Pumpkin} performed well for all pairs. The slowest success took 48 ms.\(^8\) When \textsc{Pumpkin} failed, it failed fast. The slowest failure took 7 ms. While we find this promising, proof terms were small (\( \leq 67 \) LOC); we leave evaluating performance on larger terms to future work.

\textsc{Pumpkin} was slowest when the patch showed up inverted in the difference of proofs, since \textsc{Pumpkin} had to search twice, once in each direction. A future procedure may determine which direction to search first ahead of time; proof term size may be a simple heuristic for this.

7 Conclusions & Future Work

Our vision is a future of proof automation in ITP that adapts proofs to breaking changes. This will unify the spirit of ITP—collaboration between the programmer and the tool—with the realities of modern proof engineering: Verification projects are large, specifications evolve over time, and dependencies change and break backward-compatibility. Too much of the burden of change rests on the programmer; not enough rests on the tool.

We conclude with a discussion of improvements that can help bring this vision to life. These improvements are driven by our experiences using the \textsc{Pumpkin} prototype and by conversations within the ITP community.

Supporting structural changes. Coq programmers often make structural changes. It is common, for example, to add new hypotheses, constructors, or parameters to a type. The ideal tool should find patches for these changes. Existing work in proof reuse \cite{17, 54} and type-directed diffing \cite{52} may help guide these improvements.

Exploring new components. The core components of \textsc{Pumpkin} are critical to searching for patches in Coq, but they may not be sufficient for the ideal tool. While we find the flexibility of these components promising thus far, in implementing new features, we may discover new components. For example, patches for certain structural changes may be program transformations as opposed to terms; supporting these patches may reveal new components.

Modeling diverse proof styles. Coq programmers use diverse proof styles; the ideal tool should support many different styles. Proofs about decidable domains that apply the term \texttt{dec_not_not} pose difficulties for abstraction and inversion; the ideal tool should support these. \textsc{Pumpkin} has limited support for changes in hypotheses, fixpoints, constructors, pattern matching, and nested induction; the ideal tool should implement these features.

Improving user experience. The ideal proof patching tool should be natural for programmers to use. A future patching tool can produce tactics from the patches it finds, that way

\(^8\)i7-4790K, at 4.00 GHz, 32 GB RAM
programmers can remove references to old specifications. A future patching tool can integrate with an IDE (such as Proof General [8]) or continuous integration (CI) system (such as Travis [9]) to suggest patches at natural steps in the development process.

**Handling version updates.** Updating Coq versions is an ideal use case for finding proof patches: When the client updates Coq, a tool can automatically search commits to the standard library or to `coq-contribs` for patches. In this case, the example comes from the Coq developers, not from the library client—the client never has to look at the changes that the Coq developers make. The ideal tool should fully support updates in Coq versions. **PUMPKIN** can patch certain changes in the standard library, but it does not yet search for those changes automatically and determine which configuration to use depending on what has changed. Furthermore, as proof assistant versions change, so may the AST and the plugin interface. To fully support version updates, the tool should support different language versions.

**Isolating changes.** Programmers sometimes make multiple changes to a verification project in the same commit; the ideal tool should break down large changes into small, isolated changes to find patches. This will help in finding benchmarks, supporting library and version updates, and integrating with CI systems. We may draw on work in change and dependency management [11, 22, 38] to identify changes, then use the factoring component to break these changes into smaller parts.

**Supporting other proof assistants.** Coq is just one of many proof assistants; ideal tools should support different proof assistants. While **PUMPKIN** focuses on Coq, the underlying concepts extend to other proof assistants with constructive logics (for example, Lean [6] and Agda [1]). Proof assistants with non-constructive logics (for example, Isabelle/HOL [5]) may benefit from a different approach; this is similar to the problem of finding patches for proofs of decidable domains in Coq, since classical properties provably hold for decidable propositions [7].

**Applying patches.** The ideal tool should not only find patches, but also apply the patches it finds automatically to fix broken proofs. In some cases, this may be as simple as adding the patches as hints to a hint database, so that proofs go through with no changes. However, hint databases in Coq cannot support certain terms [3], and adding too many hints may impede performance. More generally, we can integrate **PUMPKIN** with a transfer tactic [37, 65], which is a perfect fit: Transfer tactics automatically adapt proofs between isomorphic types and implications, but they do not find these functions; **PUMPKIN** finds these functions, but it does not apply them.

## 8 Related Work

Our work builds upon prior research in proof reuse, proof automation, proof engineering, refactoring, differencing & incremental computation, programming by example, and program repair.

**Proof Reuse** Our approach reimagines the problem of proof reuse in the context of proof automation. While we focus on changes that occur over time, traditional proof reuse techniques can help improve our approach. Existing work in proof reuse focuses on transferring proofs between isomorphisms, either through extending the type system [15] or through an automatic method [47]. This is later generalized and implemented in Isabelle [37] and Coq [61, 65]; later methods can also handle implications. Integrating a transfer tactic with a proof patch finding tool will create an end-to-end tool that can both find patches and apply them automatically.

Proof reuse for extended inductive types [17] adapts proof obligations to structural changes in inductive types. Later work [54] proposes a method to generate proofs for new constructors. These approaches may be useful when extending the differencing component to handle structural changes. Existing work in theorem reuse and proof generalization [31, 39, 58] abstracts existing proofs for reusability, and may be useful for improving the abstraction component. Our work focuses on the components critical to searching for patches; these complementary approaches can drive improvements to the components.

**Proof Automation** We address a missed opportunity in proof automation for ITP: searching for patches that can fix broken proofs. This is complementary to existing automation techniques. Nonetheless, there is a wealth of work in proof automation that makes proofs more resilient to change. Powerful tactics like `crush` [24] can make proofs more resilient to changes. Hammers like Isabelle’s sledgehammer [56] can make proofs agnostic to some low-level changes. Recent work [26] paves the way for a hammer in Coq. Even the most powerful tactics cannot address all changes; our hope is to open more possibilities for automation.

Powerful project-specific tactics [23, 24] can help prevent low-level maintenance tasks. Writing these tactics requires good engineering [33] and domain-specific knowledge, and these tactics still sometimes break in the face of change. A future patching tool may be able to repair tactics; the debugging process for adapting a tactic is not too dissimilar to providing an example to a tool.

Rippling [20] is a technique for automating inductive proofs that uses restricted rewrite rules to guide the inductive hypothesis toward the conclusion; this may guide improvements to the differencing, abstraction, and specialization components. The abstraction and factoring components address specific classes of unification problems; recent developments to higher-order unification [51] may help improve
these components. Lean [60] introduces the first congruence closure algorithm for dependent type theory that relies only on the Uniqueness of Identity Proofs (UIP) axiom. While UIP is not fundamental to Coq, it is frequently assumed as an axiom; when it is, it may be tractable to use a similar algorithm to improve the tool.

GALILEO [19] repairs faulty physics theories in the context of a classical higher-order logic (HOL); there is preliminary work extending this style of repair to mathematical proofs. Knowledge-sharing methods [32] can adapt some proofs across different representations of HOL. These complementary approaches may guide extensions to support decidable domains and classical logics.

Proof Engineering Existing proof engineering work addresses brittleness by planning for changes [64] and designing theorems and proofs that make maintenance less of an issue. Design principles for specific domains (such as formal metatheory [13, 28, 29]) can make verification more tractable. CertiKOS [34] introduces the idea of a deep specification to ease verification of large systems. Ornaments [27, 63] separate the computational and logical components of a datatype, and may make proofs more resilient to datatype changes. These design principles and frameworks are complementary to our approach. Even when programmers use informed design principles, changes outside of the programmer’s control can break proofs; our approach addresses these changes.

There is a small body of work on change and dependency management for verification, both to evaluate impact of potential changes and maximize reuse [11, 38] and to optimize build performance [22]. These approaches may help isolate changes, which is necessary to identify future benchmarks, integrate with CI systems, and fully support version updates.

Refactoring Our approach is close in spirit to refactoring [50]. The Haskell refactoring tool HaRe [4] automatically lifts definitions, and may be useful for improving abstraction. There is a growing body of work on refactoring in the context of ITP [18, 62]. The IDE CoqPIE [59] and the verification platform Why3 [16] can both adapt Coq proofs to simple syntactic changes. It may be possible to use our lemma factoring component to improve proof refactoring tools. Proof refactoring tools are semantics-preserving; unlike these tools, our approach handles semantic changes.

Differencing & Incremental Computation Existing work in differencing and incremental computation may help improve our semantic differencing component. Type-directed differ [52] finds differences in algebraic data types. Semantics-based change impact analysis [12] models semantic differences between documents. Differential assertion checking [41] analyzes different versions of a program for relative correctness with respect to a specification. Incremental λ-calculus [21] introduces a general model for program changes. All of these may be useful for improving semantic differencing.

Programming by Example Our approach generalizes an example that the programmer provides. This is similar to programming by example, a subfield of program synthesis [35]. This field addresses different challenges in different logics, but may drive solutions to similar problems in a dependently typed language.

Program Repair Adapting proofs to changes is essentially program repair for dependently typed languages. Program repair tools for languages with non-dependent type systems [42, 46, 48, 53, 57] may have applications in the context of a dependently typed language. Similarly, our work may have applications within program repair in these languages: Future applications of our approach may repurpose it to repair programs for functional languages.

Acknowledgments

We thank Valentin Robert and Zach Tatlock for valuable conversations about proof patching. We thank Leonardo de Moura and the UW PLSE lab for helpful suggestions. We thank the students and lecturers from the first DeepSpec Summer School for motivating future work. We thank Dan Licata for pointers in understanding the mathematical properties of inductive types. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


Thibault Gauthier and Cezary Kaliszyk. 2014. Matching concepts across HOL libraries. In CICM ‘14 (LNCs), Steven Watt, James Davenport, Alan Sexton, Petr Sojka, and Josef Urban (Eds.), Vol. 8543. Springer Verlag, 267–281. DOI: http://dx.doi.org/10.1007/978-3-319-08434-3_20


Talia Ringer, Nathaniël Yazdani, John Leo, and Dan Grossman