ABSTRACT
When working on techniques to address the wide-spread problem of software bugs, one often faces the need for a large number of realistic bugs in real-world programs. Such bugs can either help evaluate an approach, e.g., in form of a bug benchmark or a suite of program mutations, or even help build the technique, e.g., in learning-based bug detection. Because gathering a large number of real bugs is difficult, a common approach is to rely on automatically seeded bugs. Prior work seeds bugs based on syntactic transformation patterns, which often results in unrealistic bugs and typically cannot introduce new, application-specific code tokens. This paper presents SemSeed, a technique for automatically seeding bugs in a semantics-aware way. The key idea is to imitate how a given real-world bug would look like in other programs by semantically adapting the bug pattern to the local context. To reason about the semantics of pieces of code, our approach builds on learned token embeddings that encode the semantic similarities of identifiers and literals. Our evaluation with real-world JavaScript software shows that the approach effectively reproduces real bugs and clearly outperforms a semantics-unaware approach. The seeded bugs are useful as training data for learning-based bug detection, where they significantly improve the bug detection ability. Moreover, we show that SemSeed-created bugs complement existing mutation testing operators, and that our approach is efficient enough to seed hundreds of thousands of bugs within an hour.

CCS CONCEPTS
• Software and its engineering → Software creation and management; Software testing and methodologies → Machine learning;

KEYWORDS
bugs, bug injection, machine learning, dataset, token embeddings

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION
Bugs are one of the key challenges in software development, and various techniques have been proposed for bug detection, bug fixing, and bug prevention. A common problem faced when working on bug-related techniques is the need for large amounts of known, realistic bugs. Such bugs can serve multiple purposes. One of them is to provide a benchmark for evaluating and comparing bug-related tools. For example, static bug detectors and fuzz testing tools are evaluated against sets of known bugs [10, 15, 20], and bugs created via mutations are useful for evaluating the effectiveness of test suites [22]. Unfortunately, real bugs are scarce and without precise knowledge about where exactly a bug is, assessing whether a problem reported by a tool is indeed a bug requires manual effort. As a result, many tool evaluations are limited to a small number (typically, a few dozens) of bugs, e.g., bugs manually gathered from open-source projects [25, 54].

Another purpose of known bugs is to help build a bug-related technique. For example, learning-based bug detectors [36, 37, 49], defect prediction models [63], and repair tools [5, 35] rely on bugs to learn from. These techniques require large amounts of training data, typically in the form of code known to contain a (specific kind of) bug. Since obtaining large amounts of bugs is non-trivial, current techniques either focus on bugs created through simple code transformations [49], on noisy datasets that, e.g., approximate buggy code as any code changed in the next version of a program [63], on manually curated bug datasets [25, 54], or on code changes that are heuristically linked with bug reports [36].

This paper presents SemSeed, which addresses the need for large amounts of known, realistic bugs through a semantics-aware bug seeding technique. The key idea is to generalize a bug observed in the past and to seed variants of the bug at other code locations, either in the same or another program. To reason about the semantics of code, we exploit token embeddings [8, 62], a learned representation of code elements, such as identifier names and literals. To the best of our knowledge, we are the first to use learned embeddings for bug seeding.

SemSeed addresses three important challenges, which, as shown below, are not sufficiently considered in previous work. (C1—Where) Where in a target program to seed bugs that resemble a given bug-to-imitate? We address this challenge by checking which locations in the target program semantically fit the bug-to-imitate. (C2—How) How to adapt the bug-to-imitate to the target program? SemSeed addresses this challenge by semantically adapting identifiers and literals to the target location. (C3—Unbound tokens) How to handle
tokens in the buggy code that do not occur in the correct code, e.g., when the buggy code refers to an application-specific identifier name or literal? We address this challenge, called unbound tokens, through semantic analogy queries in the token embedding space that find a token that resembles the bug-to-imitate but fits the bug seeding location.

Table 1 summarizes and contrasts SemSeed with other work on automatically seeding bugs. First, mutation testing [24, 43] seeds bugs based on pre-defined code transformations. However, mutation operators cover only a small set of the syntactic transformations that occur in the wild and only sometimes represent real-world bugs [19]. Second, some work infers mutation operators from past bug fixes [9]. Both pre-defined and inferred mutation operators are applied in a purely syntactic way, without considering whether a code transformation semantically fits a code location (C1) or how to adapt the transformation to the location (C2). Third, neural models can learn from past bug fixes how to inject bugs [61]. Such approaches implicitly select target locations for seeding bugs and adapt the seeding to these target locations, but the details are hidden within the neural network. Finally, work aimed at evaluating fuzz testing tools [15, 53] seeds bugs along execution paths that are non-trivial to trigger. Even though these bugs may appear realistic from an execution perspective, they are easy to detect statically, making the approach unfit for evaluating or training static bug detectors. None of the above approaches addresses the problem of unbound tokens (C3), which our evaluation shows to prevent existing work from seeding the majority of bugs that appear in the wild.

We evaluate SemSeed by learning from real-world bugs and by seeding hundreds of thousands of new bugs. The evaluation focuses on JavaScript, because it has become one of the most popular languages and is used in various domains, but the approach is not specific to this language. The results show that SemSeed is effective at creating realistic bugs, that the seeded bugs complement bugs created with traditional mutation operators [43], and that our implementation can seed hundreds of thousands of bugs within an hour. Using the seeded bugs as training data for a learning-based bug detector [49] significantly improves the bug detection ability compared to the state of the art.

In summary, this paper makes the following contributions:

- We are the first to use learned token embeddings for bug seeding.
- We present a semantics-aware technique to decide where to seed a bug, how to adapt a given bug-to-imitate to the target location, and how to handle unbound tokens.
- We present an efficient algorithm for semantic bug seeding, which chooses from thousands of candidate bugs the semantically most suitable within about 0.01 seconds, on average.
- We show empirical evidence that SemSeed seeds realistic bugs, outperforms a purely syntactic bug seeding technique, complements traditional mutation operators, and yields bugs useful for training more effective bug detection models.

Our implementation and all data to reproduce our results are publicly available:

https://github.com/sola-st/SemSeed/

2 OVERVIEW

This section illustrates the key ideas of our approach with an example. At a high level, SemSeed consists of three main steps: abstraction, semantic matching, and pattern application. Given a set of concrete bug fixes, e.g., gathered from version histories, the first step abstracts away project-specific details, such as the identifier names. This abstraction step results in bug seeding patterns that describe how to syntactically transform a piece of code to introduce a new bug.

The top part of Figure 1 shows one concrete bug fix that the approach takes as an input. The middle part of the figure shows the corresponding bug seeding pattern. The concrete identifiers, e.g., process and platform are abstracted based on their syntactic...
category, e.g., into id1 and id2. Intuitively, the bug pattern could be described as "wrong comparison with wrong literal".

The second step of the approach matches the inferred bug seeding patterns with a given target program, addressing challenge C1. A naive baseline approach would apply a pattern at every syntactically matching location. For our example target program in Figure 1, the "wrong comparison with wrong literal" pattern could be applied at every binary expression that compares some id1.id2 with some lit1 using ===. However, such a purely syntactic approach will lead to a large number of unrealistic bugs.

A key idea of SemSeed is to not apply patterns at every syntactically matching code location, but only at locations that are semantically similar to the locations where a pattern was derived from. For the example in Figure 1, SemSeed's semantic matching component may select the code location process.arch == "x64" because it also is about checking whether some platform matches a string descriptor of a platform. The challenge in finding such a location is that the semantic similarity may not be obvious to a program analysis that is unaware of domain knowledge. For our example, the approach needs to understand that the identifiers platform and architecture refer to similar concepts, and that the literals "darwin" and "x64" both describe platforms.

The third and final step of SemSeed applies bug seeding patterns at the candidate seeding locations identified by the second step, addressing challenges C2 and C3. The key idea is to adapt the syntactic bug seeding pattern to the selected location in the target program. Specifically, the approach instantiates the pattern with identifiers and literals that are semantically similar to a location where the pattern was derived from (C2), and it finds suitable tokens for identifiers not present in the original target program (C3). As a result, the approach semantically generalizes the given concrete bugs to other code locations, which yields fewer, but more realistic seeded bugs than syntactic bug seeding. To determine how similar two code locations are, we rely on neurally learned embeddings of identifier names and literals, which have been used for other program analysis tasks in the past [3, 41, 46, 49], but to the best of our knowledge have not yet been used for bug seeding.

To seed a bug at the selected location, SemSeed transforms the code as shown at the bottom part (target program) of the figure. This transformation not only instantiates the bug seeding pattern, but it also picks a suitable platform descriptor, "x86". The approach picks this literal based on identifiers and literals used in the vicinity of the bug seeding location, mimicking a mistake a developer might also make. As a result, the seeded bug resembles the original bug that the pattern was derived from, while adapting it to the local context.

3 APPROACH

This section presents the three steps of our SemSeed approach in detail. At first, Section 3.1 describes how to extract bug seeding patterns from concrete bug fixes in version histories. Then, Section 3.2 presents how our approach semantically matches these bug seeding patterns against previously unseen code to find candidate locations for seeding new bugs. Finally, Section 3.3 describes how to apply a bug seeding pattern to a given code location by semantically adapting a specific bug-to-imitate.

3.1 Abstraction into Bug Seeding Patterns

The first step of SemSeed analyzes bug-fixing code changes in the version histories of popular code repositories to generalize them into bug seeding patterns. The basic idea is that reverting and generalizing a code change that fixes a bug will yield a pattern that we can then use to introduce this kind of bug into other code.

3.1.1 Selecting Bug-Fixing Commits. To gather examples of bug-fixing code changes, we mine the commit histories of code repositories. For a given repository, SemSeed filters all commits based on four criteria, which are designed to identify simple, bug-fixing commits. First, we select only those commits where the commit message contains any one of the words "bug", "fix", "error", "issue", "problem", and "correct", which we assume to indicate that the commit is fixing a bug. Second, we select only those commits that have a single parent commit, to avoid merged commits. Third, we select commits where the number of changed files is one and where the changed file is written in the target programming language, i.e., JavaScript in this paper. Finally, our fourth criterion is to omit commits where the number of changed lines is higher than one. Both the third and the fourth criterion help with omitting commits that fix more than a single bug or that intermingle a bug fix with other code changes. Prior work shows single-line bugs to be relevant and frequent in practice [16, 28, 51].

Since identifying bug-fixing code changes is a non-trivial problem, our four filters are designed to rather exclude some bug-fixing commits than to include many other commits. Of course, there is no guarantee that the commits obtained using these four filters are bug-fixing code changes. Manual inspection by previous research [61] of commits mined with a similar approach has shown 97% of their commits to be bug-fixing.

3.1.2 Extracting Concrete Changes From Commits. Given a set of bug-fixing code commits, SemSeed next extracts code changes into a format suitable for the remainder of the approach. Due to our filtering of commits, each commit changes exactly one line of code. One option would be to consider the entire changed line, which may, however, include parts unrelated to the bug fix. Including such "noise" would make it harder to identify recurring patterns and to find suitable locations for seeding bugs based on these patterns. Another option would be to consider only those tokens of the line that have been modified, which may, however, miss surrounding tokens important to capture the context of the change. Including some contextual information helps SemSeed identify the most suitable locations for seeding bugs with a given pattern.

To extract the changed tokens along with some context, SemSeed uses the AST of the old and the new file to find a subsequence of the changed line’s tokens that forms a complete syntactic entity, i.e., a complete subtree of an AST. Focusing on complete syntactic entities, instead, e.g., on all tokens in a changed line, increases the chance to find recurring patterns. To this end, the approach converts both the buggy and the correct file into ASTs and maps each AST node to its corresponding range of line numbers in the file. Next, the approach prunes all AST nodes that do not comprise any changed line. From the remaining nodes of a file, SemSeed selects the subtree that encompasses the maximum number of source code tokens in the changed line, and then emits the sequence of tokens rooted at
3.2 Matching Bug Seeding Patterns against Code

The problem in that field is to identify semantically similar words, such as part of a pattern against token sequences extracted from the target program. We call the matching token sequences candidate seeding locations. One key contribution of SemSeed is to determine candidate seeding locations not only by syntactically matching a pattern against the target program, but also by semantically reasoning about the similarity of the involved identifiers and literals.

3.2.1 Extracting Token Sequences from Target Program. Given a target program where SemSeed should seed bugs, the approach extracts various token sequences to match against the inferred bug seeding patterns. Similar to the pattern extraction step, SemSeed starts by parsing the target program into an AST and then extracts sequences of tokens that correspond to subtrees of the AST. Given a node and its corresponding token sequence $C = [t_1, \ldots, t_n]$, the approach applies the same abstraction algorithm as in Section 3.1.3 to get the abstracted token sequence $C_{abstr}$.

For example, consider the target program in Figure 1. The approach extracts multiple AST subtrees and corresponding token sequences. Two of the extracted subtrees represent binary expressions, and the corresponding token sequences are:

```
[item, .., errCode, ===, -1]
[process, .., arch, ===, 'x64']
```

For both sequences, abstracting identifiers and literals results in the following abstracted token sequence:

```
[id1, .., id2, ===, lit1]
```

3.2.2 Syntactic Matching. Given a set of token sequences extracted from the target program, SemSeed matches each abstracted token sequence against each bug seeding pattern. For a sequence $C_{abstr}$ and a pattern $(P_{corr}, P_{bug})$, the approach checks whether $C_{abstr}$ matches $P_{corr}$, i.e., the correct part of the pattern. As a first step, SemSeed performs a simple syntactic matching, where $C_{abstr}$ and $(P_{corr}, P_{bug})$ match if $C_{abstr}$ is equal to $P_{corr}$. Note that the syntactic matching is similar to a corresponding step in previous work on seeding bugs with inferred mutation operators [9].

For our example, the two token sequences abstracted into

```
[id1, .., id2, ===, lit1]
```

are both equal to the correct part of the bug seeding pattern from Section 3.1. Hence, both binary expressions in the target program are retained as candidate seeding locations.

3.2.3 Semantic Matching. Syntactically matching bug seeding patterns against a target program yields a large number of candidate seeding locations. Unfortunately, seeding bugs at all these locations would result in many seeded bugs that do not semantically resemble the concrete bugs that SemSeed is learning from. For example, applying the bug seeding pattern of our running example both to `item.errCode === -1` and to `process.arch === 'x64'` would yield two seeded bugs. However, only the second seeded bug would be semantically similar to the concrete bug that the pattern was learned from: The bug is about incorrectly checking the name of a platform against a string that describes a platform, `process.platform === 'darwin'`, and a similar bug could occur in the `process.arch === 'x64'` expression. In contrast, the other possible candidate seeding location `item.errCode === -1` matches the original bug only syntactically, but not semantically.

To ensure that SemSeed seeds realistic bugs, the approach focuses on bugs that semantically resemble a given concrete bug fix. Checking whether two code locations are semantically similar is a hard problem, which we address by borrowing ideas from machine learning-based natural language processing. An important research problem in that field is to identify semantically similar words, such as
as "chicken" and "fowl". A state-of-the-art approach to address this problem is word embeddings learned from a corpus of text, e.g., using Word2vec [42] or FastText [8]. An embedding maps each word into a real-valued vector so that semantically similar words have similar vectors. For example, the word vectors of "chicken" and "fowl" will be close to each other in the embedding space. To determine how similar two word vectors $v, w$ are in an embedding space, we compute their cosine similarity: $\text{simil}(v, w) = \frac{v \cdot w}{\|v\| \|w\|}$.

SemSeed computes embeddings of source code tokens and uses them to reason about the semantic similarity of tokens. As the embedding technique, we build on FastText [8], which we choose for two reasons. First, FastText does not suffer from the out-of-vocabulary problem [29], because it reasons about the n-grams in a token instead of relying on a fixed-size vocabulary. Second, FastText has been shown to more accurately represent the semantic similarities of code tokens than other popular embeddings [62].

Algorithm 1 summarizes how SemSeed checks whether a given token sequence semantically matches a bug seeding pattern. The approach semantically compares the concrete tokens $C_{corr}$ where the bug described by the pattern has occurred with the tokens $C$ in the target program. The algorithm computes for each identifier and literal token in $C$ its semantic similarity to the corresponding token in $C_{corr}$. If the average similarity for all tokens in $C$ exceeds a threshold $m$, which we call the matching threshold, then the algorithm returns True, and SemSeed marks $C$ as a candidate seeding location. Averaging the similarity across embeddings of individual tokens is inspired by work on representing natural language sentences and documents [4, 33] and captures how well a word would fit a target location. For bug seeding patterns derived from multiple concrete bug fixes, the approach invokes the algorithm multiple times, and considers $C$ a candidate seeding location if $C$ resembles at least one of the concrete bug fixes. Our evaluation studies the impact of the matching threshold $m$ in practice.

For the example, SemSeed invokes Algorithm 1 for the two syntactically matching code locations. The first invocation, where $C$ contains the tokens in $\text{item.errCode} = \text{errCode} = -1$, is likely to return False (depending on the matching threshold) because the compared tokens, e.g., $\text{item vs. process}$, or 'darwin' vs. '-1', are dissimilar. In contrast, the second invocation is likely to return True because the tokens in $\text{process.arch} = \text{arch}$ have a high pairwise similarity with the tokens in $\text{process.platform} = \text{darwin}$.

### 3.3 Applying Bug Seeding Patterns

#### 3.3.1 Unbound Tokens

The main challenge in this step is tokens that appear in the buggy part but not in the correct part of the pattern, which we call unbound tokens. For example, recall the bug seeding pattern in Figure 1, and in particular, the 'lit2' token in the buggy code. When applying the pattern to a program, this token is unbound, i.e., it is unclear what concrete token to insert instead of 'lit2'. Prior work on automatically seeding bugs based on inferred bug patterns [9, 61] ignores the problem of unbound tokens, and hence can apply only bug seeding patterns without any such tokens. However, as we show in our evaluation, the majority of all bug seeding patterns contains unbound tokens, i.e., ignoring them would ignore many bug seeding opportunities.

Before presenting our approach for applying bug seeding patterns with unbound tokens, we consider a few alternatives. Suppose we want to apply a bug seeding pattern ($P_{corr}, P_{bug}$), which was inferred from a concrete bug fix ($C_{bug}, C_{corr}$) and has an unbound token $t_i$. The question is which concrete token to use instead of $t_i$ when concretizing $P_{bug}$ in the target program.

One option would be to replace $t_i$ with the concrete token it is bound to in $C_{bug}$. However, this token may not be a natural fit for the target program. For example, when $t_i$ is an identifier, then simply replacing it with the corresponding identifier from $C_{bug}$ is likely to result in an undefined variable, resulting in a rather unrealistic bug. Another option would be to replace $t_i$ with a random token sampled from the vocabulary of all tokens, which again is likely to result in an unrealistic bug. A third option would be to sample a token from all tokens in the same file or same function as the bug seeding location. While this approach increases the chance of resulting in realistic code, it is still likely to yield a token that does not fully fit the context of the bug seeding location, e.g., because it uses a variable of a wrong type.

#### 3.3.2 Binding Tokens via Analogy Queries

To address the challenge of binding an unbound token in a way that fits the bug seeding location, and hence ultimately, create a realistic bug, we again take inspiration from natural language processing. Given a learned word embedding, word analogy tasks intend to answer similarity questions involving two or more pairs of words. For example, one may ask the analogy question What word is to "Trance" what "Tokyo" is to "Japan"?, which is likely to yield the answer "Paris" [42]. Adapting this idea to unbound tokens, SemSeed uses the bound tokens of a bug seeding pattern to resolve any unbound tokens in the same pattern.

Figure 2 illustrates our analogy-based technique for binding unbound tokens using the example in Figure 1. Given the bug seeding pattern, the token lit2 is unbound. In contrast, process and arch are bound, because id1 and id2 occur both in the correct and the buggy part of the pattern. SemSeed searches for a suitable token for lit2 by asking three analogy questions:

- What token is to 'x6a' what 'win32' is to 'darwin'? This analogy is represented by the two upward arrows in Figure 2.
Algorithm 2 Apply bug seeding pattern to candidate token sequence

Input: A candidate token sequence $C$, a concrete bug fix $(C_{bug}, C_{corr})$ and its corresponding bug seeding pattern $(P_{corr}, P_{bug})$, a set $T$ of identifier and literal tokens.

Output: Tokens $C_{seed}$ of seeded bug

1. $C_{seed} \leftarrow \emptyset$
2. for $i \leftarrow 1$ to length($C_{bug}$)
   3. if kind($C_{bug}[i]$) ∈ {Identifier, Literal} then
      4. Append $C_{bug}[i]$ to $C_{seed}$ ▷ Copy token
   5. else if $P_{bug}[i]$ bound to $t_{bound}$ then
      6. Append $t_{bound}$ to $C_{seed}$ ▷ Use bound token
   7. else
      8. $V_{tgt} \leftarrow \emptyset$ ▷ Bind token via analogy queries
      9. for $t_{abstr} \in P_{corr}$ do
         10. $t_{orig} \leftarrow$ token that $t_{abstr}$ is bound to in $C_{bug}$
         11. $t_{seed} \leftarrow$ token that $t_{abstr}$ is bound to in $C$
         12. $v_{tgt} \leftarrow$ emb($t_{seed}$) + emb($C_{bug}[i]$) − emb($t_{orig}$)
         13. Add $v_{tgt}$ to $V_{tgt}$
         14. $t_{tgt} \leftarrow$ arg max $\sim$(emb($t$), avg($V_{tgt}$))
         15. Append $t_{tgt}$ to $C_{seed}$

3.3.3 Algorithm. After providing an intuition of the approach, Algorithm 2 presents in detail how SemSeed applies a bug seeding pattern. The algorithm takes three inputs. First, a candidate token sequence $C$, identified as described in Section 3.2, which the algorithm will mutate to seed a bug. Second, a concrete bug fix $(C_{bug}, C_{corr})$ and its corresponding bug seeding pattern $(P_{corr}, P_{bug})$. The algorithm will seed a new bug by imitating the given bug and by semantically adapting it to the context of the candidate token sequence. Third, a set $T$ of literal and identifier tokens from which the algorithm selects tokens to use for unbound tokens. For example, this set may consist of all identifiers and literals in the file where the bug is seeded or the $n$ most common tokens in a corpus of code. Our evaluation compares different ways of computing the set $T$.

The main loop of the algorithm goes through all tokens in the bug-to-imitate $C_{bug}$, and iteratively builds a new sequence $C_{seed}$ of buggy tokens. For each token to generate, the algorithm distinguishes three cases. The first case (line 3) handles tokens that are neither identifiers nor literals, but standard tokens of the programming language, such as operators or parentheses. Each such token is directly copied from the bug-to-imitate into $C_{seed}$. The second case (line 5) handles bound identifier and literal tokens, i.e., tokens that appear in the candidate token sequence. The algorithm uses the concrete token $t_{bound}$ from the candidate token sequence and appends it to $C_{seed}$. For our running example, the first two cases handle the tokens `process`, `arch`, and `!!!`. These cases are sufficient to handle bug seeding patterns without any unbound tokens, where it suffices to rearrange the tokens in the candidate token sequence into the buggy token sequence. In contrast, unbound tokens, such as `lit2` in our example, require including a new token into the sequence $C_{seed}$.

The third case (line 7 in Algorithm 2) handles unbound tokens by computing a set $V_{tgt}$ of target points in the vector space of the token embedding. For each abstract token $t_{abstr}$ that appears in the correct part $P_{corr}$ of the bug seeding pattern, the algorithm computes a target point based on the concrete tokens $t_{orig}$ and $t_{seed}$ that $t_{abstr}$ is bound to in the bug-to-imitate and the candidate token sequence, respectively. The target point is computed at line 12, which implements an analogy query. The query starts from the embedding of $t_{abstr}$ and adapts it by adding the vector that leads from the embedding of $t_{abstr}$ to the corresponding token $C_{bug}[i]$ in the bug-to-imitate. For our running example, the algorithm computes three target locations (which correspond to the three analogy questions from above):

$$V_{tgt} = \{\text{emb}(\text{'x64'}) + \text{emb}(\text{'win32'}) - \text{emb}(\text{'darwin'}), \text{emb}(\text{arch}) + \text{emb}(\text{'win32'}) - \text{emb}(\text{platform}), \text{emb}(\text{process}) + \text{emb}(\text{'win32'}) - \text{emb}(\text{process})\}$$

That is, the algorithm finds the difference between the vectors of `darwin` and `win32` and adds it to the vector of `x64`, and similar for the other two queries. The resulting three target locations are the blue points in Figure 2.

Given the set $V_{tgt}$, the algorithm queries $T$ for the token whose embedding is most similar to the average of all target points. Intuitively, this token is the available token that is semantically closest to the token observed in the bug-to-imitate. Once retrieved, the algorithm adds the token to the sequence $C_{seed}$ of result tokens.
We evaluate SemSeed based on bug fixes extracted from the version histories of 100 popular JavaScript projects. The evaluation addresses the following research questions:

- **RQ1**: How effective is SemSeed in reproducing real-world bugs?
- **RQ2**: How does SemSeed compare to a semantics-unaware variant of the approach?
- **RQ3**: What is the impact of the configuration parameters of the approach?
- **RQ4**: How useful are the seeded bugs for training a learning-based bug detector?
- **RQ5**: How do the seeded bugs compare to bugs created with traditional mutation operators?
- **RQ6**: How efficient is SemSeed in seeding bugs?

### 5.1 Experimental Setup

We gather bug-fixing commits from the version histories of the 100 JavaScript projects that have most stars on GitHub. For each repository, we extract all commits and filter them as explained in Section 3.1.1, resulting in 3,600 concrete bug fixes. We split the bugs into 2,880 guiding bugs, used to extract bug seeding patterns and as concrete bugs-to-imitate, and 720 held-out bugs. The split is date-based, using older commits as guiding bugs and newer commits as held-out bugs, so we can evaluate whether imitating bugs from the past creates bugs that have occurred later on. Extracting bug seeding patterns from the guiding bugs yields 2,201 bug seeding patterns. The frequency of the patterns follows a long tail distribution, which shows that real-world bugs are diverse, and that extracting bug seeding patterns from a large dataset is worthwhile.

The approach depends on three configuration parameters that control how many and which bugs get seeded: the matching threshold \( m \), the set \( T \) of tokens to choose unbound tokens from, and the number \( k \) of bugs to seed per code location. As a default, we use \( m = 0.2 \), \( k = 10 \), and \( T \) as all tokens in the file where the bug gets seeded plus the 1,000 most frequent tokens across all files with guiding bugs. RQ3 further evaluates the impact of these parameters.

### 5.2 RQ1: Effectiveness in Reproducing Real-World Bugs

We evaluate SemSeed’s ability to seed realistic bugs by comparing the seeded bugs with the held-out bugs. There are two preconditions for SemSeed to be able to reproduce a specific bug. First, the bug seeding pattern of the bug must occur across the guiding set and the held-out set. Due to the long-tail distribution of bug seeding patterns, many of the patterns occur only once, and we focus on the 151 concrete bugs that have a pattern in the intersection of guiding bugs and held-out bugs. Second, for bugs that involve tokens not present in the correct code, i.e., unbound tokens, the unbound token must be in the set \( T \) of tokens the approach chooses from when applying a bug seeding pattern. For our default configuration of \( T \), 53 out of the 151 bugs that fulfill the first precondition also fulfill the second precondition. We use this set of 53 held-out bugs as the target bugs, and compute how many of them SemSeed reproduces, i.e., the seeded bug is exactly the same as the original bug.

The files in which the 53 target bugs should be seeded, the semantic matching identifies all 53 locations as a target location. 16 of the target bugs are rearrangements of existing tokens, i.e., similar to the inferred mutation operators of prior work [9]. Seeding these bugs is straightforward and SemSeed reproduces all of them. The remaining 37 involve unbound tokens, and SemSeed’s algorithm for binding these tokens is successful for all but six of the bugs. Overall, the approach reproduces 47 out of the 53 target bugs.

Table 2 shows three examples of successfully reproduced real-world bugs. For each example, we show the correct and buggy variant of both the bug-to-imitate and the seeded bug. The first example is a bug without unbound tokens, but which requires rearranging existing tokens only. The second example is a bug with unbound identifiers, and SemSeed’s algorithm for binding these tokens is successful for all but six of the bugs.

SemSeed reproduces 47 out of 53 bugs that are in scope for the approach.
### 5.3 RQ2: Comparison with Semantics-Unaware Bug Seeding

SemSeed relies on the semantic information embedded in identifiers and literals in two ways: (i) to select the locations for imitating a given bug, and (ii) to bind unbound tokens. To show the importance of these ideas, we compare our approach against a semantics-unaware variant of SemSeed, which (i) applies a bug pattern at every syntactically matching location, and (ii) binds unbound tokens by randomly picking from all tokens in the set $T$. This baseline approach reproduces only 16 out of the 53 target bugs. All of these 16 bugs do not have any unbound tokens. For bugs that need unbound token, we repeat the experiment for ten times with different seed values and randomly select a token from $T$. In none of the ten repetitions does the random selection pick the correct token required to seed a bug. The reason why randomly binding unbound tokens is ineffective is that picking the right token by chance from $T$ is unlikely. In our default configuration, $T$ contains more than 1,000 tokens, and even when $T$ consists only of tokens that appear in the same function, there are typically several dozens of identifiers and literals to choose from.

To further illustrate the importance of handling unbound tokens, Table 3 lists some bug seeding patterns that SemSeed finds, along with their frequency in our dataset. All of the five most common bug seeding patterns (top-5 shown in table) and 62% of all bug seeding patterns contain at least one unbound token. Prior work on bug seeding based on past bug fixes does not handle unbound tokens [9, 61], and hence, could not benefit from these patterns.

### 5.4 RQ3: Impact of Configuration Parameters

#### 5.4.1 Matching Threshold $m$

The matching threshold $m$ determines in Algorithm 1 whether to apply a bug pattern to a code location. A threshold of 0 means that the seeding location need not be similar to the bug-to-imitate at all, i.e., the decision to apply a bug seeding pattern is purely syntactic. In contrast, a threshold of 1 requires the tokens to perfectly match the original bug.

Figure 3 shows how the matching threshold influences the bugs that SemSeed creates. The upper two curves show two percentages: in blue, the percentage of seeded bugs out of all bugs that a purely syntactic
approach, i.e., with \( m = 0 \), would seed; in red, the percentage of reproduced target bugs (RQ1) among the seeded bugs. As expected, both percentages decrease when the matching threshold increases. Intuitively, if filtering potential matching locations based on the matching threshold is effective at guiding SemSeed towards realistic bugs, then the blue curve should decrease faster than the red curve. Indeed, the figure shows a clear gap between the two curves, i.e., the semantic matching of target locations is effective. For example, with a matching threshold of 0.4, the approach seeds a bug at only 20% of all possible locations, but still reproduces 60% of all bugs that SemSeed can reproduce.

Compared to purely syntactic matching of bug patterns, the semantic matching increases the chance to seed realistic bugs.

5.4.2 Token Set \( T \) and Number \( k \) of Bugs to Seed. Another parameter is the set \( T \) of tokens to consider when binding unbound tokens (Section 3.3.3). We experiment with three variants of \( T \):

1. \( T_{\text{fct}} \): Search for unbound tokens only in the function where the bug gets seeded.
2. \( T_{\text{file}} \): In addition to \( T_{\text{fct}} \), search among all tokens in the file where the bug gets seeded.
3. \( T_{\text{common}} \): In addition to \( T_{\text{file}} \), search among the 1,000 most frequent tokens across all files in the guiding set.

A larger search space increases the chance that the token required to reproduce a bug exists in \( T \), but also makes it more difficult to choose the “right” token, i.e., a token that indeed produces a realistic bug. A related parameter is how many bugs to seed for a given code location and bug seeding pattern. Our approach seeds one bug for each of the \( k \) most likely tokens found by Algorithm 2, and we evaluate values of \( k \) ranging from 1 to 10.

Figure 4 illustrates the effect that the token set \( T \) and the number \( k \) of bugs to seed have on the number of real-world bugs that SemSeed reproduces. We see that using a more restricted search space of tokens yields fewer reproduced bugs than a larger search space. Regarding the influence of \( k \), considering more than the single most likely token significantly increases the number of reproduced bugs, in particular for larger \( T \). Our default configuration of \( T = T_{\text{common}} \) and \( k = 10 \) yields 47 reproduced bugs.

Depending on the token set \( T \) and the number \( k \) of bugs to seed, SemSeed reproduces between 29 and 47 of the target bugs.

5.5 RQ4: Usefulness for Training a Learning-Based Bug Detector

To evaluate the usefulness of semantic bug seeding, we explore one of the applications of SemSeed: seeded bugs as training data for learning-based bug detection. We build on DeepBugs [49], which learns from examples of correct and incorrect code, and then predicts bugs in previously unseen code. DeepBugs supports several bugs patterns, of which we focus on two that are particularly challenging to seed bugs for:

- Wrong assignment bugs, where the right hand side of an assignment is incorrect, e.g., writing \( i = 0 \); instead of \( i = 0 \).
- Wrong binary operands, where a developer uses an incorrect operand in a binary expression, e.g., accidentally writing \( \text{length} \ast \text{height} \) instead of \( \text{length} \ast \text{breadth} \).

The other bug patterns [49], e.g., swapping function arguments, are simpler to seed and do not require to select unbound tokens.

We train DeepBugs using two configurations that differ in the way the incorrect code examples are generated. One configuration, called “artificial”, uses DeepBugs’s default generation of incorrect code examples, which randomly applies purely syntactic transformations and binds unbound tokens at random from \( T_{\text{file}} \). The other configuration generates incorrect examples with SemSeed, which we configure to seed only bugs that match the two bug patterns targeted by DeepBugs. We apply both configurations to the same code corpus: a de-duplicated version [1] of a JavaScript corpus [50], which consist of 120K files. Generating for each correct example at most one incorrect example, the “artificial” configuration yields 1.1 million wrong assignments and 2.6 million wrong binary operands. Since SemSeed focuses on locations that have a semantic match with one of the guiding bugs, it creates fewer incorrect examples, namely 248K wrong assignments and 267K wrong binary operands.

Once trained, we measure the ability of DeepBugs to detect real-world bugs. As the bug patterns targeted by DeepBugs are relatively rare, we gather the bugs in three ways: (i) Those 8 of the held-out bugs that match the two bug patterns; (ii) Additional bugs gathered from 900 popular GitHub JavaScript projects using the methodology in Section 3.1.1; and (iii) Bugs from the JavaScript variant of an existing dataset of single-statement bugs [28]. This process yields 412 bugs (35 wrong assignments and 377 wrong binary operands).

We measure precision, i.e., how many of all reported warnings are among the known bugs, and recall, i.e., how many of all known bugs DeepBugs finds. For each warning, DeepBugs returns a probability for the location to be buggy. Figure 5 shows the precision and recall of DeepBugs depending on the probability threshold used to decide which warnings to report. Overall, using SemSeed-generated bugs instead of artificial bugs significantly increases the effectiveness of DeepBugs, with clearly improved recall and roughly the same precision. For example, using a threshold of 0.5, SemSeed increases the detected bugs from 7% to 53%.

To understand why SemSeed improves the bug detection ability of learned bug detectors, consider two bugs seeded into the following code:
As illustrated by this example, a model trained on the artificial bugs tends to identify obvious yet unrealistic mistakes. Instead, training DeepBugs with SemSeed’s bugs teaches the model to identify subtle yet more realistic mistakes. More broadly, the results also illustrate a quality-versus-quantity tradeoff in bug seeding. The SemSeed-generated bugs yield more effective bug detectors, despite being an order of magnitude fewer than the artificially created bugs.

Using semantically seeded bugs as training data for a learning-based bug detector allows for finding significantly more bugs.

### 5.6 RQ5: Comparison with Traditional Mutation Operators

Existing code mutation approaches, such as Mutandis [43] for JavaScript and Major [24] for Java, use pre-defined mutation operators. We compare the mutation operators in Mutandis with the bugs created by SemSeed. Based on the 2,880 guiding bugs, we seed 677,217 bugs into a random sample of 1,000 JavaScript files and then compare the seeded bugs to the 23 mutation operators in Mutandis.¹

98.2% of the SemSeed-generated bugs go beyond the 23 predefined mutation operators. The 1.8% of the bugs that are shared with Mutandis correspond to 165 out of the 2,880 guiding bugs. For example, one of the Mutandis patterns is about changing a literal in a condition, a change SemSeed also performs. Another example is about removing the var keyword from a variable declaration, a pattern that SemSeed also learns and applies. Inversely, Mutandis also creates some bugs that SemSeed cannot seed. Out of the 23 Mutandis operators, SemSeed has a corresponding bug seeding pattern for 16. For 13 out of these 16, SemSeed seeds at least one bug, while for the remaining three no suitable bug seeding location is found. Among the remaining 23 − 16 = 7 Mutandis operators, two are out of scope for SemSeed because the code transformation affects more than one line, e.g., swapping two nested loops. For the other five operators, SemSeed could in principle seed bugs, but there is no corresponding guiding bug. These are mostly about changes to JavaScript APIs, e.g., removing the integer base argument 10 from calls like `parseInt('09/11/08', 10).`

SemSeed complements traditional mutations by seeding many bugs beyond a fixed set of pre-defined mutation operators.

### 5.7 RQ6: Efficiency

We measure the efficiency of SemSeed by keeping track of the time it needs to seed the 677,217 bugs into the 1,000 files from RQ5. This experiment is performed on a machine with 48 Intel Xeon E5-2650 CPU cores and 64GB of memory. Because some files allow for thousands of seeded bugs, we set a time limit of 30 minutes per file. Out of the 1,000 files, SemSeed could seed bugs into 902 files where it found at least one matching bug seeding pattern. In total, seeding 677,217 bugs takes 140 minutes. Analyzing what part of the approach takes the most time, we find that the analogy queries are the biggest bottleneck.

SemSeed takes, on average, 0.01 seconds to seed a bug and hence can generate a large number of bugs in very little time.

### 6 LIMITATIONS AND THREATS TO VALIDITY

SemSeed focuses on single-line bugs, for two reasons: (i) we can gather a large set of these bugs automatically, which facilitates the evaluation, and (ii) these bugs are relevant and important in practice [16, 28, 51]. For example, Karampatsis and Sutton [28] show that there is an instance of one out of 16 common patterns of single-line bugs every 1,600 to 2,500 lines of code. To generalize SemSeed to more complex bugs, one would consider token sequence that span multiple lines. One challenge we anticipate is that the probability that a complex bug-to-imitate syntactically matches code in another program is smaller than for single-line bugs. Addressing this challenge, e.g., by approximately matching bug seeding patterns to code locations, remains for future work.

Among the many applications of bug seeding, we select learning-based bug detection to evaluate SemSeed’s usefulness. Based on our¹

¹Mutandis can also use runtime information to decide which bugs to seed, which we ignore here because our focus is on static bug seeding.
comparison with traditional mutation operators, we are optimistic that the approach could also be useful, e.g., for mutation testing, and envision future work on this and other applications.

Our evaluation measures the realism and the usefulness of the seeded bugs, but it does not confirm that each seeded bug indeed changes the behavior of the target program. Future work could measure to what extent the seeded bugs affect a program’s semantics, e.g., by executing the program’s test suite.

We implement the approach for JavaScript and cannot draw conclusions about how well it would work for other languages. The fundamental challenges that SemSeed addresses, i.e., where to seed bugs, how to adapt a given example bug to a target location, and how to handle unbound tokens, are language-independent.

7 RELATED WORK

Bug Seeding. One approach to bug seeding is to apply mutations, e.g., based on a predefined set of transformation patterns [21, 24]. Similar to our work, [9] propose to infer such patterns from code changes. In contrast to these approaches, SemSeed decides where to apply a bug seeding pattern and how to adapt it to the local code context based on semantic similarities of code elements.

Tufano et al. [61] use neural machine translation (NMT) to learn and apply mutations. Their approach requires hundreds of thousands of bug-fixing commits to be trained properly. In contrast, SemSeed learns from few examples – in the extreme case, one can use a single example bug to seed similar bugs at various target locations. In general, an NMT-based approach could also seed bugs in a semantics-aware way, and comparing such an approach against SemSeed will be interesting future work. One advantage of our work over NMT-based approaches is that SemSeed enables seeing and possibly filtering the bug seeding patterns. For example, to train a DeepBugs [49]-like, learning-based bug detector, we focus on bug patterns supported by DeepBugs by selecting the corresponding bug patterns. An important difference between SemSeed and both [9] and [61] is that our approach handles unbound tokens, seeding bugs even if this requires an application-specific identifier or literal.

Tailored mutation operators [2, 13, 24, 39, 43], e.g., insert code fragments that occur elsewhere in a project. In contrast to such approaches, the mutations applied by us are based on previously seen bug fixing patterns and not project-specific as in [2] or hard coded as in [24, 43]. IBIR also learns from past bugs how to seed new bugs [31]. It uses natural language in a bug report to decide where to seed a bug, whereas SemSeed focuses on the tokens (including natural language identifiers) in the code. IBIR neither adapts bugs to a target location and nor addresses the unbound token problem, which we show to be crucial for the majority of bug patterns.

Motivated by the abundance of fuzz testing tools [6, 18, 48, 56], automatically seeded bugs have been proposed for evaluating fuzz testing [15, 53]. These seeded bugs aim at being non-trivial to trigger in an execution, but are easy to detect on the source code level, e.g., because the seeded bug relies on magic numbers.

Mining Code Change Patterns. Osman et al. [47] describe an empirical study of frequent bug fixing code changes. Negara et al. [44] identify repetitive code changes from fine-grained sequences of code changes recorded in an IDE. In contrast, our approach mines only concrete changes that correspond to a bug fix rather than any change made by a developer. Nguyen et al. [45] mine semantic change patterns by converting the correct and buggy files to program dependence graphs. Instead of a graph, we leverage embeddings of tokens as the semantic representation and extract changes as a sequence of tokens. Kim et al. [32] manually inspect human-written patches to infer common fix patterns. Neural machine translation can learn to apply bug fixes [60]. SemSeed addresses the inverse problem of seeding bugs, instead of fixing them.

Bug Benchmarks. Several bug benchmarks have been proposed, including SIR [14], Defects4J [25], BugSwarm [38] BugBench [40], BegBunch [11], iBugs [12], ManyBugs [34], Codeflaws [37], DbgBench [7]. Our work complements such manually curated bug datasets by automatically seeding bugs into a target program.

Finding Matching Code. Code clone detection [23, 26, 38, 52, 55] relates to the semantic matching part of SemSeed. These approaches find matching code pieces via string-based, parse tree-based, or token-based comparisons. Our semantic matching relates to the token-based techniques, but differs by using token embeddings to find a match.

Token Embeddings. Recent work shows that token embeddings enable learning-based program analysis, e.g., to detect bugs [49], to predict types [41], to de-obfuscate code [3], or to map APIs across programming languages [46]. Our work is the first to use pre-trained token embeddings for bug seeding. Future work could adapt SemSeed to other kinds of token embeddings, e.g., contextual embeddings [17, 27, 30].

8 CONCLUSION

This paper presents SemSeed, an approach for seeding bugs in a semantics-aware way. Given a possibly small set of example bugs, the approach infers bug seeding patterns and then imitates the given bugs at various code locations in a target program. The key novelty is to go beyond purely syntactic bug seeding by (i) checking if a code location semantically matches the bug-to-imitate, (ii) adapting the bug seeding pattern to the local code context, and (iii) binding unbound tokens based on semantic analogy queries. To reason about the semantics of code elements, SemSeed builds on learned token embeddings, which have not been used for bug seeding before. Our evaluation with thousands of real-world bugs shows that the approach effectively seeds realistic bugs, while being efficient enough for creating hundreds of thousands of bugs within an hour. The created bugs complement traditional mutation operators and are useful as training data for learning-based bug detectors, allowing them to find many otherwise missed bugs.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC, grant agreement 851895), and by the German Research Foundation within the ConcSys and Perf4JS projects.

REFERENCES


Semantic Bug Seeding: A Learning-Based Approach for Creating Realistic Bugs  


