Genetic Improvement for Software Product Lines: An Overview and a Roadmap

Roberto E. Lopez-Herrejon¹, Lukas Linsbauer¹, Wesley K. G. Assunção²,³, Stefan Fischer¹, Silvia R. Vergilio¹, Alexander Egyed¹
¹ISSE - Johannes Kepler University Linz, 4040 Linz, Austria
²DINF - Federal University of Paraná, CP: 19081, 81531-980, Curitiba, Brazil
³COINF - Technological Federal University of Paraná, 85902-490, Toledo, Brazil
{roberto.lopez, lukas.linsbauer, stefan.fischer, alexander.egyed}@jku.at,
{wesleyk, silvia}@inf.ufpr.br

ABSTRACT

Software Product Lines (SPLs) are families of related software systems that provide different combinations of features. Extensive research and application attest to the significant economical and technological benefits of employing SPL practices. However, there are still several challenges that remain open. Salient among them is reverse engineering SPLs from existing variants of software systems and their subsequent evolution. In this paper, we aim at sketching connections between research on these open SPL challenges and ongoing work on Genetic Improvement. Our hope is that by drawing such connections we can spark the interest of both research communities on the exciting synergies at the intersection of these subject areas.

CCS Concepts

• Software and its engineering → Software product lines; Search-based software engineering; • Computing methodologies → Search methodologies;

Keywords

evolutionary algorithms, genetic programming, genetic improvement, software product lines, variability

1. INTRODUCTION

Software Product Lines (SPLs) are families of related software systems [⁴], where each product has a different combination of features - increments in program functionality [⁴⁸]. Over the last two decades, extensive research and application of SPL practices have shown their technological and economical advantages (e.g. [⁴⁶]). However, for an effective adoption of SPLs many challenges still need to be addressed. Salient among them are the reverse engineering of SPLs from existing software system variants, the most prevalent scenario in industry [¹⁰], and the evolution of SPLs [³ ⁷ ²³].

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 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.

GECCO ’15, July 11 - 15, 2015, Madrid, Spain
© 2015 ACM. ISBN 978-1-4503-3488-4/15/07…$15.00
DOI: http://dx.doi.org/10.1145/2739482.2768422

Genetic Improvement (GI) is the process of automatically improving a system’s behavior using genetic programming [²⁴]. In contrast with typical work on genetic programming that evolves programs from scratch [⁴¹], GI starts from existing systems which are evolved so that they are better with respect to given criteria. GI is an expanding research area that has already been successfully used, for instance, to specialize boolean satisfiability (SAT) programs written in C++ for concrete tasks such as combinatorial interaction testing [⁴⁰], to repair broken functionality [¹⁵], or to add new functionality [¹⁸]. In recent publications, Harman and colleagues advocate the potential use of GI for SPL development [¹⁷ ²⁴]. For instance, to generate a Pareto surface of programs that exhibit the same functionality but with different quality attributes such as performance. Another possibility they propose is actually growing new functionality, i.e. creating a new branch or product variant of a SPL.

Our recent systematic mapping study showed the increasing interest in research at the intersection of SPLs and Search Based Software Engineering (SBSE) [⁵⁴]. The driving motivation of this paper is to sketch in further detail the connections between SPLs (reverse engineering and their evolution) and ongoing work on GI. We believe that by drawing these connections we could bring the attention to the similar challenges faced by both research communities, and draft a rough roadmap of open issues that could be addressed in the short to medium term. We hope this paper can help to spark the interest on the exciting synergies at the intersection of these areas.

The paper is organized as follows. Section 2 provides the basic background to understand the SPL challenges that our paper focuses on. Section 3 draws the connections between ongoing work on SPLs and the work of GI as characterized by the Genetic Improvement of Software for Multiple Objective Exploration (GISMOE) project [¹⁹ ²⁴]. Section 4 presents our roadmap of issues that could be addressed as some of the first steps in exploring the synergies between SPLs and GI.

2. SPL BACKGROUND

Some of our recent work has focused on several novel approaches for reverse engineering SPLs and their subsequent evolution [¹ ² ³ ¹² ¹³ ¹⁷ ¹⁹ ²² ²⁸ ³⁰ ³³ ³⁵]. In this section, we illustrate the two main challenges that our work addresses.
2.1 Reverse Engineering SPLs

The most common scenario for reverse engineering SPLs in practice is depicted on the left part of Figure 1. In this scenario, there exist several related software products (i.e. they offer some similar functionality) that were created partially independently (e.g. they may have shared some artifacts at some point in their history) typically with ad hoc techniques; for instance, by forking the new products from other already existing software systems and then manually adapting them to fit the new requirements. Such products follow their own, mostly independent, development lifecycle. Ad hoc practices like these are collectively called Clone and Own (C&O) [10], and support available for them is very limited.

Not surprisingly, even when dealing with a small number of products, C&O approaches lead to multiple maintenance problems like inefficient feature update or bug fixing (i.e. a bug needs to be fixed separately for each product), duplicate functionality, redundant and inadequate testing, etc [10]. In other words, C&O approaches may work fine up to a certain number of distinct products depending on product domains, the development organization and its software engineering practices. However, after such number, simply adding new independently-developed systems is no longer feasible either because of managerial, economical or technical reasons.

It is then that SPL approaches are the premier alternative for effectively coping with the complexity of the variability – the capacity of software artifacts to vary – in the existing products. The transition to a SPL approach is not a minor undertaking because it requires a significant investment of resources to identify, extract, and reify what is common (i.e. commonality) and what varies (i.e. variability) across all the artifacts involved as depicted on the right part of Figure 1. The artifacts with their variability and commonality identified constitute the building blocks used by variability management mechanisms to generate the existing products but should in addition be capable of providing robust support for maintenance and evolution tasks. There are many variability management mechanisms, all of them with advantages and disadvantages (see e.g. [14]). What is common among variability management mechanisms is that they rely on a variability model to express the valid combinations of features desired for the products of a SPL. The most popular form of variability models is feature models depicted at the top of Figure 1. For further details please refer to [21].

Our previous work has explored the use of SBSE techniques for reverse engineering feature models [21][28][33][35].

In this paper, the focus is on reverse engineering the variability and commonality of the artifacts that are fed into the variability management mechanisms. We use as a running example simple software products that draw different geometric shapes on a screen canvas. Even though our example uses Java code for illustration, the challenges that will be highlighted can be easily mapped to other types of artifacts.

In SPLs, each software product is distinguished by the set of features it provides. Let us consider an example of a product P1 which provides a single feature LINE. This product consists of two classes whose code snippets are shown in Figure 3. Class Canvas contains fields for the start and end points of lines (Line 3), a constructor (Line 4), a method paint that paints the lines (Lines 5-8) and an auxiliary method setEnd that sets the end point (Line 9). Class Canvas contains fields lines that holds a list of the lines drawn (Line 12), and method paintComponent (Lines 13-18) which is called every time the canvas is repainted and that in turn draws the lines on the canvas by calling their method paint (Line 16).

Now consider a second product P2 whose code is also sketched on Figure 3. In addition to feature LINE, this product also implements feature WIPE that wipes the canvas clean. This new feature causes changes in the code because it adds a new method wipe (Lines 35-44) shown underlined in the figure. When this method is called, the list that holds the Line objects is cleared (Line 39), so that when the call to repaint (Line 40) is made, the triggered call of paintComponent (Lines 33-34) which is called every time the canvas is repainted and that in turn draws the lines on the canvas by calling their method paint (Line 16).

Let us consider a third product P3 that implements feature RECT which draws rectangles on the canvas. The code snippets are also shown on Figure 3. Similarly to variant P1, it defines a new class, Rect (Lines 44-52), to implement the new shape type. This class contains the fields to draw the rectangle, upper left corner coordinates plus width and height (Line 45), a paint method (Lines 47-50), and a setEnd method (Line 51). This variant also contains its Canvas class (Lines 55-61), which now has a list of rectangles...

Figure 1: Reverse Engineering Software Product Lines Overview

Figure 3: Class Canvas and Feature Models Overview
(Line 54) and its method paintComponent (Lines 55-60) that paints the rectangles (Line 58).

Based on these three products of our running example, the goal is to reverse engineer a SPL that we will refer to as the Draw Product Line (DPL). As mentioned before, there are many variability management mechanisms that have been used to implement SPLs. The simplest ones are based on annotations made to the artifacts that are pre-processed based on the different selection of features for each product that needs to be created. Figure 2 shows an example of how the class Canvas could be annotated. In this simple example, a product that contains feature LINE will include in the final code of class Canvas the pieces of code of Line 4 and Line 12. Similarly for products that provide feature RECT which will include Line 4 and Line 15. In practice, the conditions that guard the inclusion of the fragments of these artifacts can be more elaborate to reflect the complex interactions that features can have. This will be elaborated on in Section 3.1.

The main challenge for reverse engineering SPLs is thus to automate as much as possible the transition from the artifacts of individual products to the corresponding artifacts that, together with a chosen variability management mechanism, correctly and completely capture all possible and valid combinations of features. This challenge becomes more daunting when hundreds of products, hundreds (if not thousands) of features, and multiple artifacts other than source code are considered.

```java
1 class Canvas {
2     #ifdef $LINE
3         List<Line> lines = new LinkedList<Line>();
4     #endif
5     #ifdef $RECT
6         List<Rect> rects = new LinkedList<Rect>();
7     #endif
8
9     void paintComponent(Graphics g) {
10         #ifdef $LINE
11             for (Line l : lines) { l.paint(g); }
12         #endif
13         #ifdef $RECT
14             for (Rect r : rects) { r.paint(g); }
15         #endif
16         ...  
17     }
18 }
```

Figure 2: Example DPL using preprocessor annotations

### 2.2 Evolution of SPLs

Broadly speaking we define SPL evolution as the progress of the changes made to a SPL towards a better adaptation to the environment where it is intended to be used. There are multiple evolution scenarios for SPLs [26]. For illustration purposes here we focus on a single one, the creation of a new product based on the combinations of existing features. More concretely, let us now consider the case of developing a new product P4 that provides both rectangles (RECT) and wiping capabilities (WIPE), both features already available in the SPL from two previous products P2 and P3. The challenge now becomes how to build the new product based on the knowledge already available, as captured, for instance, in the pre-processor implementation of Figure 2.

```java
19 }
```

```java
20 /* Product 1 (LINE) */
21 class Line {
22     Point startPoint, endPoint;
23     Line(Point start) { ... }
24     void paint(Graphics g) {
25         g.setColor(Color.BLACK);
26         g.drawLine(startPoint.x, startPoint.y, endPoint.x, endPoint.y);
27     }
28     void setEnd(Point end) {...}
29 }
30 class Canvas {
31     List<Line> lines = new LinkedList<Line>();
32     void paintComponent(Graphics g) {
33         ...  
34         // Paints the figures
35         for (Line l : lines) { l.paint(g); }
36         ...  
37     }
38     void wipe() {
39         lines.clear();
40         repaint();
41 }
42 }
```

Figure 3: Example products P1, P2, P3

```java
43 /* Product 3 (RECT) */
44 class Rect {
45     int x, y, width, height;
46     Rect(int x, int y) {...}
47     void paint(Graphics g) {
48         g.setColor(Color.BLACK);
49         g.drawRect(x, y, width, height);
50     }
51     void setEnd(int newX, int newY) {...}
52 }
53 class Canvas {
54     List<Rect> rects = new LinkedList<Rect>();
55     void paintComponent(Graphics g) {
56         ...  
57         // Paints the figures
58         for (Rect r : rects) { r.paint(g); }
59         ...  
60     }
61 }
```
### 3. SPLS MEET GISMOE

In this section we start drawing the connections, albeit with broad strokes, between recent and ongoing research on reverse engineering SPLs and their evolution with the Genetic Improvement of Software for Multiple Objective Exploration (GISMOE) approach [19][20][24]. What drove our interest into the subject was the following quote presented at a keynote talk at ASE 2012 [19]:

"The GISMOE approach may also offer solutions to some of the issues raised by SPLs. For example, using GISMOE, we can create new branches automatically: the GP engine will evolve the new versions of the product family from existing members of the family. We may also be able to merge versions when the product family becomes large or unwieldy."

Inspired by this quote, in the next subsections we make an attempt at aligning the two challenges described in Section 2 with recent GISMOE publications.

#### 3.1 Feature-level functional sensitivity analysis

GISMOE advocates the use of sensitivity analysis to determine the places in code artifacts that should be evolved and which should remain untouched according to the property analyzed. We argue that to address the SPL challenges with GISMOE, a key premise is to be able to perform a sensitivity analysis at the feature level because features are the building blocks of the products of a SPL. This entails establishing traceability links between features and feature interactions and the artifacts that realize them. These links would enable targeting the evolution to concrete parts of the artifacts to achieve the desired functionality goals.

We now describe this form of sensitivity analysis based on our traceability work [27], which is part of our Extraction and Composition for Clone-and-Own (ECCO) approach [12][13][29]. ECCO works under two simple premises. First, that each product provides its artifact base (e.g. Java source code) and a list of features that it implements (e.g. feature LINE for product P1, and features LINE and WIPE for product P2). Second, that features with the same name implement the same functionality in similar ways. For instance, in our running example, feature LINE has the same functionality in both products P1 and P2. In addition, for sake of illustration, LINE is shown in Figure 3 implemented with exactly the same code in products P1 and P2; however, we will discuss later the implications of this assumption.

Let us now define the two kinds of modules in the software products that ECCO considers [12][27]:

**Definition 1. Base Module.** A base module implements a feature regardless of the presence or absence of any other features and is denoted with the feature’s name written in lowercase.

**Definition 2. Derivative Module.** A derivative module $m = \delta^i(c_0, c_1, \ldots, c_n)$ implements feature interactions, where $c_i$ is $F$ (if feature $F$ is selected) or $\neg F$ (if not selected), and $n$ is the order of the derivative.

ECCO’s extraction algorithm automatically produces a set of traces between both types of modules and the artifacts (e.g. source code fragments) that implement them [27]. These code fragments can be of any granularity level, from entire classes down to individual statements. In essence, this algorithm works by incrementally refining the traces based on integrating the knowledge that can be derived from each product. Let us illustrate the gist of this idea with the code of Figure 3. As a first step, if we initially consider only product P1, all its code would trace to module line. With the knowledge added by product P2, the algorithm can infer now, based on the different features provided by the two products, that all the code in Lines 21-29 Lines 25-26 and Line 12 traces to module line. This is so because both products, on one hand, provide the same feature LINE...
and, on the other, implement this feature in a similar way. In addition, the algorithm now traces Lines 38-41 to module \texttt{wipe} and derivative module \( \delta'(\texttt{line\_wipe}) \). Note that if we were to add another product with feature \texttt{Wipe} but without feature \texttt{LINE}, the algorithm could further refine the traces to indicate that the feature interaction denoted by derivative \( \delta'(\texttt{line\_wipe}) \) is actually realized by the statement in Line 39. This short description omitted many relevant details such as dealing with hierarchy and ordering issues among the artifacts. For further information, please refer to [12, 13, 27].

The relevance for GISMOE of the traceability provided by ECCO is that this new form of sensitivity analysis provides the means by which to focus the places where to target the evolution of the artifacts, for instance to graft new functionality (i.e. grow and insert a new feature), repair a bug (i.e. the faulty feature interaction in Figure 4), or improve a non-functional property of a feature.

Nonetheless, there remain several challenges that still need to be addressed to fully support this form of sensitivity analysis. Currently, the similarity among artifacts is computed purely on the structural comparison of their elements. We envision that clone detection technologies will play an important role addressing this need (e.g. [6, 12, 13]).

### 3.2 Automated test case generation

In GISMOE, the test cases are used to evaluate the fitness of the artifacts that result from evolution. There is an extensive body of work on SPL testing ([39][11]). Among the most popular approaches for SPL testing is \textit{Combinatorial Interaction Testing (CIT)}, a fact that has been corroborated by our recent mapping study [32]. However, to the best of our knowledge, the automated test generation for the scenarios described in Section 2 remains an open question.

ECCO was conceived from the outset to be able to support any type of artifact provided that it can be differentiated, i.e. that a mechanism is available for identifying the common and different parts of the artifacts based on given criteria (e.g. artifact structure). From this perspective, test artifacts (e.g. models or code) are simply another type of artifact for which their traceability to features and feature interactions, as described in Section 2, can be obtained. Consequently, we envisage that the new test cases can themselves be derived and genetically improved in the same way as other artifacts can, i.e. by focusing on the precise artifact fragments identified by ECCO’s traceability algorithm. In addition, we speculate that work on computing CIT covering arrays applied to SPLs (e.g. [31, 32]) that instead focuses on the coverage of base and derivative modules could serve to reduce the testing effort, in a similar way to the notion of output bins that has been employed by GISMOE [21].

### 3.3 Feature-level non-functional sensitivity analysis

Research on non-functional properties in SPLs has attracted recent interest. For instance, Noorian et al. propose a taxonomy to classify the different approaches [38]. A first line of research is the work by Siegmund et al. who estimate memory footprint and main-memory consumption [45]. They propose an analytical model that computes, from a CIT perspective, covering arrays of different strengths (i.e. \( t=1, t=2 \), and ad hoc) which they use to perform actual measurements of a single non-functional property. Based on the measurements, their model estimates the impact that features and feature interactions can have on the non-functional property.

A second line of research works by adding additional attributes to the features of the feature model that models the variability of a SPL. Based on these values, multi-objective evolutionary algorithms are used to obtain configurations (i.e. products with valid combinations of features) that meet the property requirements. For instance, the work of Sayyad et al. focuses on information related to cost and defects [44]. Our previous work considered also multi-objective evolutionary algorithms but instead applied to relevant properties (i.e. usability, battery consumption, and memory footprint) for the dynamic configuration of SPLs deployed in mobile phone environments [39]. In both cases, the evaluation was performed basically with synthetic values.

We argue that the solution to achieving the goal of non-functional sensitivity analysis advocated by GISMOE when applied in the context of SPLs lies somewhere at the intersection of these two lines of research. On one hand it is important that the optimization is based on actual measured values at the right level of granularity (i.e. not at the coarse grain level of attributes of single features) and that the optimization considers multiple objectives, hence multiple non-functional properties, simultaneously.

### 3.4 The need of co-evolution

The GISMOE approach relies on co-evolution of the programs and their corresponding tests. SPLs make it a more challenging task because, in a real-life scenario, multiple types of artifacts must be concurrently considered and hence co-evolved. In addition, a key requirement is to keep the artifacts not only for a single product but simultaneously for all the products of the SPL. Work on consistency checking across multiple variants (e.g. [29]) could be brought to bear as a starting point to address this need.

### 3.5 The human in the loop

GISMOE advocates the involvement of developers of the systems, in part to employ domain knowledge perhaps only available in the heads of the domain experts but also to help reduce the adoption barrier of its automatically generated systems.

Our ECCO approach also benefits from the involvement of developers. ECCO provides hints, for instance, to indicate that the implementation of some new modules cannot be fully automated [12, 13, 25]. Based on the input received manually, ECCO updates and refines its traces.

Another source of knowledge could be, for example, employing ontologies to provide information about the relationships between features in a given domain. For instance, in our running example, knowing that a \texttt{LINE} and a \texttt{RECT} are both geometrical forms could help to trim the search space for finding a repair to the faulty product \texttt{P4}. More concretely, to use the code of product \texttt{P3} as source to obtain code fragments (i.e. field \texttt{rects}) from which to evolve or graft the repair.

One of the most salient challenges here is how to convey the large amount of information to the human, such that the interactions are both simple and efficient. Research and tooling available for software visualization is definitively the starting point [47]. Except for few examples, such as [39][36], SPL visualization remains an area largely unexplored.
4. THE ROAD AHEAD

In this section we broadly describe some of the first exploration steps we plan to undertake in the short-term future.

- **Exploit clone detection.** We plan to extend our traceability algorithm to detect similar and different code beyond the current level that focuses on identical names and structure [6,42,43].

- **Grafting variability annotations into UML models.** We plan to explore selecting artifacts from a product and through GI graft annotations that capture the variability present, for the selected artifacts, in all the products of a SPL. We intend to use information retrieval metrics as done in our previous work [35], and perform a comparison with ECCO’s traceability extraction algorithm.

- **Adapting GenProg ideas into the ECCO approach.** We want to explore how the automated repair work proposed by GenProg [16] can be adapted to the context of SPLs, in particular to solve repair scenarios such as the ones illustrated in Section 2. Doing so would require to frame and assess the plastic surgery hypothesis (see [4]) within the realm of SPLs, and integrate the ECCO tool with ASTOR [37], the Java based implementation of GenProg principles.

We hope that the overview we provided can help to draw the attention to the great synergies existing between GI and SPLs, and entice the readers to pursue and tackle the challenges identified.

Acknowledgments

This work was supported by Austrian Science Fund (FWF): P 25289-N15 and the Brazilian Agencies CAPES 007126/2014-00 and CNPQ.

5. REFERENCES


