

DREU 2014 - MIT - Progress Report

Santiago Gonzalez

Undergraduate, EECS Department, Colorado School of Mines

July 21, 2014

Abstract

This is mid-way report of the my research experience here at the Massachusetts Institute of Technology (MIT) with Dr. Armando Solar-Lezama under the DREU program. Described herein is the research project I was involved in and its entailing aspects, the project's status, a review of my working environment here at MIT, and additional comments pertaining to the research experience.



1 Research Project

1.1 Research Problem

A common hurdle in the field of constraint-based synthesis involves reasoning about large code bases efficiently. Often, efficient implementations of certain constructs, such as data structures, are more complex and difficult to reason about, for both the synthesizer and the programmer, than equivalent, less-efficient implementations. That is, we want the synthesizer to be able capable of reasoning a complex implementation by representing it in a simpler form that facilitates reasoning.

Consider that a programmer has relatively large and complex program which uses sets (i.e. unordered collections) in several places and is written for a synthesis based compiler. Sets can typically be implemented in a variety of different ways, each suited to different types of problems. In some cases, the programmer knows that a tree-based set implementation would be ideal. However, in other cases, a hash table based implementation would be best, but the programmer is unsure which set implementation would be optimal, so he (erroneously) chooses the tree-based implementation. Upon compilation, the synthesizer takes a very long time to reason about the program, due to the complexity of the tree-based set. A system which could use a simpler, equivalent set implementation for reasoning while still using the efficient implementation, which could potentially be inferred from analysis, for the compiled program would be very beneficial.

1.2 Background

The Computer Aided Programming group at the Massachusetts Institute of Technology (MIT) has been working on a C-like, constraint-based synthesis programming language, named `Sketch` for several years. `Sketch` relies on a SAT solver and a program synthesizer based on a counterexample-guided inductive synthesis (CEGIS) algorithm. CEGIS converges on a correct solution by iteratively generating potential implementations which

are then run through a verifier, a better potential candidate can then be produced from counterexample traces upon failure. [10]

Sketch has demonstrated the applicability of constraint-based synthesis to a variety of programming problems, including automated grading for programming assignments [9], machine learning problems such as recommendations within social media [2], and optimization of database-backed applications which use an object-relational mapping layer [3]. **Sketch**'s versatility and flexibility arise from its easy to understand programming model (i.e. the programmer does not need to be well versed in verification proofs), its ability to run on nearly any system due to its cross-platform Java and C++ codebase, and its syntactic similarity to C.

1.3 Approach

We introduce a solution to the problem of reasoning about complex structures by extending the **Sketch** programming language to include a new programming construct called the package, which has the ability to define subtype relations between related packages. Such subtype relations allow the synthesizer to use an alternative package for analysis which has a simpler, equivalent implementation, while still using the more efficient package for the compiled code. Packages can have hierarchical relationships with each other. As such, packages have similar behavior to conventional classes in object-oriented programming languages, such as Java, with several key differences.

While packages have the concept of inheritance and polymorphism, they are not instantiated like classes. Packages serve to contain code, and as such, packages vend a variety of functions and C-like structs, which *can* be instantiated.

The availability of a function or struct (i.e. whether it is private to the package or public to the entire program) within a package is determined by whether all child packages implement that function or struct. For example, assume there are packages A and B , such that $B <: A$, and A contains a function f_1 , while B contains a function f_2 ; f_1 would be

private to its package unless B also implemented while f_2 is public since B has no children. Note, that child packages can introduce new functionality to a package, much like classes.

Additionally, a package's structs can not be instantiated from outside of a package, necessitating the use of factory functions, and such a struct's fields can only be accessed from within that struct's containing package. This enforces the use of setter and getter functions to access a struct's data, allowing different, replaceable struct implementations. These semantics essentially simplify and allow the use of structs as data objects whose exact implementations are unknown by the programmer. Furthermore, the synthesizer has the assurance that instances of a struct can be replaced with other package's implementations, provided that the packages vend equivalent behavior.

Packages are best used to represent collections of code with very similar behaviors and differing implementations. Essentially, the Liskov Substitution Principle (LSP) [6] is enforced for both parent-child and child-parent package relationships.

Packages with no parent package and no children packages, termed utility packages, are essentially simple containers for code, much like namespaces and packages in C++ and Java, respectively. Utility packages allow full access to structs' fields since it is impossible for ambiguities and type conflicts to arise, due to the lack of descendants.

In order to be able to replace a package with one of its descendants, both packages must be verified to be functionally equivalent for a given program. We plan to accomplish such verification through the use of a most general client that aims to generalize the potential usage instances of a package. We intend to construct the most general client through a hybrid approach: (1) analyzing the program in which the package is used, and (2) creating random function call sequences based on information gathered from program analysis. Such an approach would allow for a looser definition of equivalency, one that is more specific to the use case and less general. Analysis of a program's package use cases would reduce the quantity of function call sequences that need to be tested.

As a contrived example, consider that a programmer has a program with packages A

and B , such that $B <: A$, and both A and B contains a function f_1 that performs an operation on a number. Assume that B has a more efficient, and complex, implementation that works with all number inputs, while A 's implementation only works with even number inputs. In the most general sense, A and B are not equivalent since they fail verification with odd number inputs. However, if the program only uses B with even numbers, the two packages can be considered equivalent for the program.

1.4 Results and Contributions

We have fully defined and integrated the type semantics for the package system into the `Sketch` compiler. The package system's type rules have been formalized in standard notation as described in [1]. Package abstract syntax tree (AST) nodes have been extended from simple code containers to support inheritance, public function and struct interfaces, and the parser grammar has been extended to support the Java-like syntax for declaring parent packages. Furthermore, the package system's type-checking has been implemented as a new compiler pass using a visitor pattern as opposed to being integrated into the preexisting type-checking compiler pass.

Due to the slightly unconventional nature of packages, several of the type rules have unique formulations. For example, the type-checking rule for determining the ambiguity of a function call avoids strange, unintuitive situations where the removal of an unused package from the program changes the package from which function is used in an unspecified function call. Such scenarios can arise due to the nature of child packages which allows them to modify a parent package's function's visibility. These would bring about subtle, difficult to debug issues that result in the program's logic changing, hence eliciting the more conservative rule.

Continuing work involves the creation of a most general client for a given program to verify package equivalence. Additionally, efforts will go towards modifying the synthesizer to reason with simpler implementations and extensive testing, possibly with string libraries.

Future work could potentially include modifications to the compiler in order to allow it to deduce which package implementation is most efficient for a given usage instance.

1.5 Related Work

There has been significant work by researchers at ETH Zürich on using the Eiffel programming language with contracts [4] as an alternative approach to this problem; namely, in the field of program verification. [8] The design by contract programming model is a way to achieve modularity and local reasoning by having the programmer write very detailed contracts. Specifically, model-based contracts which support the verification of software modules. [7] Contracts are invariant specifications which cover pre-conditions, known as obligations, and post-conditions, known as benefits. Proponents for design by contract programming argue that rigorous functionality specifications help to bring successful, reusable software to fruition. [5]

2 Project Status

Currently I am six weeks in, and have four weeks remaining in the research experience. As detailed in 1.4, the project has been progressing smoothly. The type-checker compiler pass for the package system has been fully designed, implemented, and tested. Current work involves formulating a most general client to allow the compiler to verify package equivalence.

Dr. Solar-Lezama has discussed plans to potentially publish a research paper to PLDI, VMCAI, or CAV. We are aiming to have a rough draft of the paper – with a potentially incomplete results section – by the end of the 10-week research experience. About two weeks ago, Dr. Solar-Lezama made me aware of the OOPSLA Student Research Competition, a poster competition in the OOPSLA conference later this year, and encouraged me to apply, citing that it would be a good opportunity to present the research and get feedback

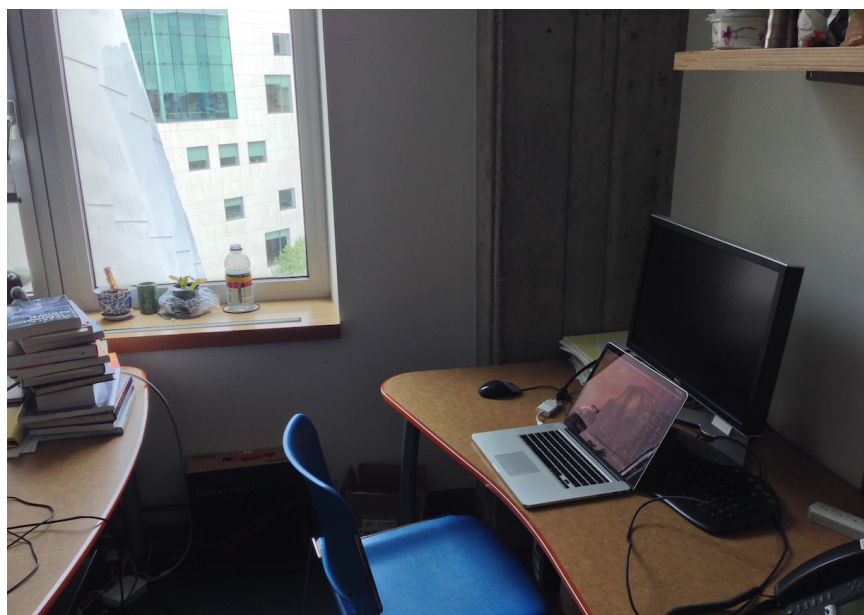


Figure 1: My workspace in the Stata Center

prior to publication. I wrote an extended abstract detailing the project (very similar to section 1) for the application.

3 Work Environment

Dr. Solar-Lezama was able to provide me with very nice workspace accommodations (Figure 1). I was given a spot next to a window in an office shared with three other graduate students on the seventh floor of the Stata Center. I have found that the Stata Center (Figure 2), although difficult to navigate in the beginning, is well organized and very conducive to interactions with other people through its various common areas. For example, there is a nice kitchen area on each floor with coffee, tea, a microwave, and other amenities.

Working with Dr. Solar-Lezama and the graduate students in his research group has



Figure 2: The Stata Center at MIT

been great. Dr. Solar-Lezama seems to be confident in my abilities since he has given me important work on the compiler, he has not simply figuratively “shoved me off to the side”. When Dr. Solar-Lezama is not traveling, he is able to meet with me several times each week to discuss the project. Every Friday, a “group lunch” is organized for the research group which has allowed me to get to know everyone better.

I have attended several talks and thesis defenses here at the Computer Science and Artificial Intelligence Laboratories (CSAIL) over a variety of topics, including linguistic structure speech modeling, auto-active verification, quantitative program correctness, and more. I have found that these talks are a great opportunity to learn about new developments in the field from various research institutions.

4 Additional Comments

Overall, I have thoroughly enjoyed these past several weeks here at MIT. Everyone here at CSAIL is awesome, and the Boston area is a really cool place to live and work in (except in the winter, so I am told). The Boston area has tons of things to do. Its wide assortments of parks and paths are great for walking, running, and being around during weekends; one of my personal favorites being the Charles River Esplanade which runs along side the Charles river, across from MIT. The subway system (referred to locally as “the T”) is very clean with absolutely no graffiti and allows you to easily reach practically any place in Boston and the surrounding areas.

4.1 Housing

Initially, I logically attempted to secure a dorm at MIT for the summer. However, I was unable to due to myself not meeting the age requirements (I am 16).

I was able to find a relatively compact, single-room apartment near Davis Square (which is near Tufts University) with a private bathroom and a small refrigerator and microwave; Davis Square itself being about two blocks away. The commute is very easy, usually taking

between 20 and 25 minutes to go from home to my office. It is very convenient not having a roommate since I like my living spaces to be clean and tidy.

4.2 Biggest Challenge

This summer has been free of insurmountable challenges, thankfully. I feel that the biggest challenge for me so far this summer has been being away from my family and friends. This has been the first time I have lived alone and been away for a prolonged period of time. My mom and dad are great parents and have always supported me. I did not have any stereotypical “homesickness” since being here has been an amazing experience, to say the least. This mild discomfort has been greatly alleviated thanks to technology. I am able to talk and FaceTime with them frequently.

Overall, I was able to adapt to being here very quickly, especially since my mother was here for the first few days to help me get settled in. By the middle of the first week, I felt fine living alone and was super excited to start working on Dr. Solar-Lezama’s projects. Also, I was not completely cut off from my family, I could still converse with them in real time. Additionally, my parents and sister came to visit at the end of the third week since my sister was going to a camp in New Hampshire. I was able to show them around Boston and MIT.

4.3 Most Exciting Thing

The most exciting thing about this summer has quite simply been working at MIT. Every time I get off the train and walk into the MIT campus I think: “Oh my god! I’m actually in MIT!”. Its exciting to think that I am working alongside many of the greatest minds in the world, even if only for one summer. I consider myself very lucky that I have been able to work on such a cool project, whose underlying ideas and technology will most likely make it into the hands of many programmers in coming years.

It has also been very exciting for me to be living in the Boston area. I have never been

in Boston before so it has been very nice being able to see all the various sites and places of interest in Boston and the surrounding area. It is also interesting seeing such a diverse group of people here, especially in Cambridge.

4.4 Closing Remarks

I am very thankful to the organizers of the DREU program and Dr. Solar-Lezama for allowing me to have this amazing once-in-a-lifetime opportunity.

References

- [1] L. Cardelli. *Handbook of Computer Science and Engineering*. CRC Press, 1997.
- [2] A. Cheung, A. Solar-Lezama, and S. Madden. Using program synthesis for social recommendations. *CIKM*, 2012.
- [3] A. Cheung, A. Solar-Lezama, and S. Madden. Optimizing database-backed applications with query synthesis. *PLDI*, 2013.
- [4] Eiffel Software. *Building bug-free O-O software: An Introduction to Design by Contract*.
- [5] J.-M. Jazequel and B. Meyer. Design by contract: the lessons of ariane. *Computer*, 30(1):129–130, Jan 1997.
- [6] B. H. Liskov and J. M. Wing. A behavioral notion of subtyping. *TOPLAS*, 1994.
- [7] N. Polikarpova, C. A. Furia, and B. Meyer. Specifying reusable components. *VSTTE*, 2010.
- [8] N. Polikarpova, J. Tschannen, C. A. Furia, and B. Meyer. Flexible invariants through semantic collaboration. *International Symposium on Formal Methods*, 19, 2014.

- [9] R. Singh, S. Gulwani, and A. Solar-Lezama. Automated semantic grading of programs. *PLDI*, 2013.
- [10] A. Solar-Lezama, C. G. Jones, and R. Bodík. Sketching concurrent data structures. *PLDI*, 2008.