Analysis and Synthesis with "Big Code"

Eran Yahav
The Techion, Haifa, Israel

The vast amount of code available on the web is increasing on a daily basis. Open-source hosting sites such as GitHub contain billions of lines of code. Community question-answering sites provide millions of code snippets with corresponding text and metadata. The amount of code available in executable binaries is even greater. In this lecture series, I will cover recent research trends on leveraging such "big codefor program analysis, program synthesis and reverse engineering.

We will consider a range of semantic representations based on symbolic automata [4, 7], tracelets [3], numerical abstractions [5, 6], and textual descriptions [10, 1], as well as different notions of code similarity based on these representations. To leverage these semantic representations, we will consider a number of prediction techniques, including statistical language models [8, 9], variable order Markov models [2], and other distance-based and model- based sequence classi

cation techniques. Finally, I will show applications of these techniques including semantic code search in both source code and stripped binaries, code completion and reverse engineering.

References

[1] `http://like2drops.com`

[2] R. Begleiter, R. El-Yaniv, G. Yona. *On Prediction using Variable Order Markov Models.* Journal of Artificial Intelligence Research; pp. 385–421; 2004.

[3] Y. David, E. Yahav. *Tracelet-based Code Search in Executables.* In: Procs. of PLDI '14; pp. 349–360; 2014.

[4] A. Mishne, S. Shoham, E. Yahav. *Typestate-based Semantic Code Search over Partial Programs.* In: Procs. of OOPSLA'12; 2012.

[5] N. Partush, E. Yahav. *Abstract Semantic Differencing via Speculative Correlation.* In: Procs. of OOPSLA'14; 2014.

[6] N. Partush, E. Yahav. *Abstract Semantic Differencing for Numerical Programs.* In: Procs. of SAS'13; pp. 238–258; 2013.

[7] H. Peleg, S. Shoham, E. Yahav, H. Yang. *Symbolic Automata for Representing Big Code.* In: Procs. of STTT'15; 2015.

[8] V. Raychev, M. Vechev, E. Yahav. *Code Completion with Statistical Language Models.* In: Procs. of PLDI'14; 2014.

[9] R. Rosenfeld. *Two Decades of Statistical Language Modeling: Where do we go from here?* In: Procs. of the IEEE, Vol. 88; pp. 1270–1278, 2000.

[10] M.B. Sinai, E. Yahav. *Code Similarity via Natural Language Descriptions.* In: OBT'15: POPL *Off the Beaten Track*; 2014.