Towards Comparing and Combining Points-to Analyses

Tobias Gutzmann, Antonina Khairova, Jonas Lundberg, Welf Löwe

Växjö University, Sweden

September 20, 2009

Points-to Analysis

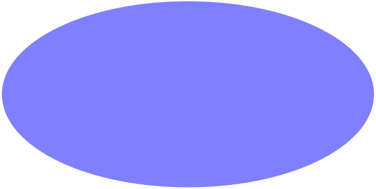
- Static *dataflow analysis*
- Computes reference information, e.g., possible targets of a call and possible objects referenced by a field (in the following, we use abstract *result sets* for illustration)
- Input to, e.g., optimizing compilers, software analysis tools
- Requirements: **accuracy**, speed

Motivation

- Different implementations use different data structures
- Different papers use different metrics
    - Hard to tell, how they compare to each other
- If we can properly *compare* analysis results from different implementations, we can also *combine* them, thus exploiting the "best of" from different approaches/implementations

Types of Analysis

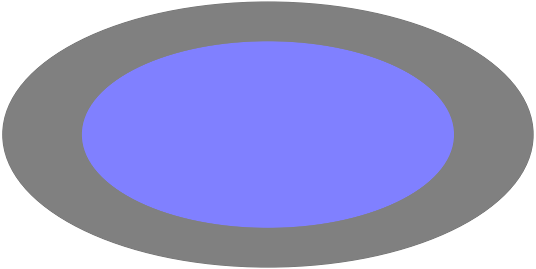
Gold standard (G): The “exact” result set of a given analysis



Types of Analysis

Gold standard (G): The “exact” result set of a given analysis

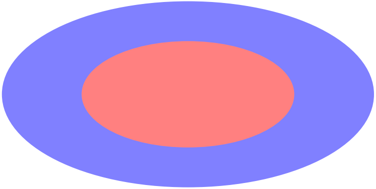
- *conservative* analysis: over-approximation of G



Types of Analysis

Gold standard (G): The “exact” result set of a given analysis

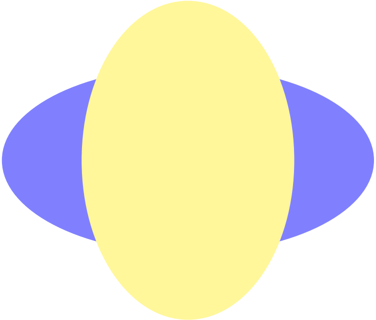
- *conservative* analysis: over-approximation of G
- *optimistic* analysis: under-approximation of G



Types of Analysis

Gold standard (G): The “exact” result set of a given analysis

- *conservative* analysis: over-approximation of G
- *optimistic* analysis: under-approximation of G
- *general* analysis: a mix of conservative and optimistic analysis



Why *general* analysis?

- Static analysis should be conservative
- This is often the case only for *subsets* of a programming language!
  - Dynamic class loading
  - Reflection
- Maybe even on purpose, to improve performance
- Applications:
  - Software understanding tools
  - Optimistic optimizations (with guard)

Assessing accuracy

- Precision P: how much of the analysis result set is in G?
- Recall R: how much of G is found through the analysis?
- Accuracy F (F-score): Weighted measure between Precision and Recall

Comparing

Special case of A_1, A_2	Comparison of A_1, A_2 with respect to		
	P	R	F-score
A_1, A_2 cons.	$P_1 \geq P_2 \Leftrightarrow \lvert A_1 \rvert \leq \lvert A_2 \rvert$	$R_1 = R_2 = 1$	$F_1 \geq F_2 \Leftrightarrow \lvert A_1 \rvert \leq \lvert A_2 \rvert$
A_1, A_2 opt.	$P_1 = P_2 = 1$	$R_1 \geq R_2 \Leftrightarrow \lvert A_1 \rvert \geq \lvert A_2 \rvert$	$F_1 \geq F_2 \Leftrightarrow \lvert A_1 \rvert \geq \lvert A_2 \rvert$
A_1 cons. A_2 opt.	$P_1 \leq P_2 = 1$	$R_1 = 1 \geq R_2$	$F_1 \geq F_2 \Leftrightarrow \frac{\lvert A_1 \rvert}{\lvert G \rvert} \leq \frac{\lvert G \rvert}{\lvert A_2 \rvert}$
A_1 cons.	$P_1 \geq P_2 \Leftarrow \lvert A_1 \rvert \leq \lvert A_2 \rvert$	$R_1 = 1 \geq R_2$	$F_1 \geq F_2 \Leftarrow \lvert A_1 \rvert \leq \lvert A_2 \rvert$
A_1 opt.	$P_1 = 1 \geq P_2$	$R_1 \geq R_2 \Leftarrow \lvert A_1 \rvert \geq \lvert A_2 \rvert$	$F_1 \geq F_2 \Leftarrow \lvert A_1 \rvert \geq \lvert A_2 \rvert$

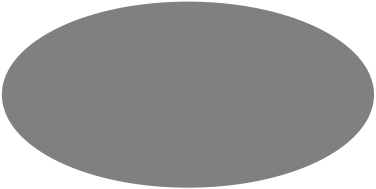
Combining

- Two conservative analyses: Compute intersection
- Two optimistic analyses: Compute union

Improving Analysis

What happens when improving a static analysis?

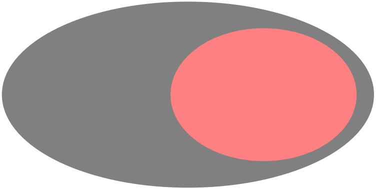
- Given a baseline analysis



Improving Analysis

What happens when improving a static analysis?

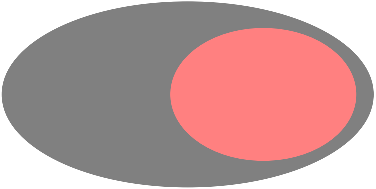
- Given a baseline analysis
- improve its precision (e.g., by context sensitivity) - result set becomes smaller



Improving Analysis

What happens when improving a static analysis?

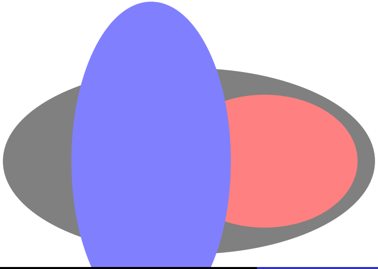
- Given a baseline analysis
- improve its precision (e.g., by context sensitivity) - result set becomes smaller
- this is sound for conservative analysis. What happens if the baseline analysis is *general*?



Improving Analysis

What happens when improving a static analysis?

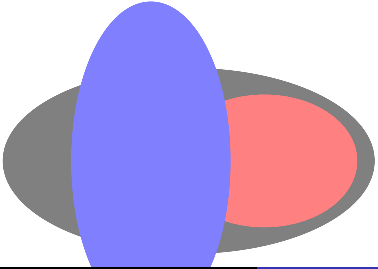
- Given a baseline analysis
- improve its precision (e.g., by context sensitivity) - result set becomes smaller
- this is sound for conservative analysis. What happens if the baseline analysis is *general*?
- Assume we knew the Gold Standard



Improving Analysis

What happens when improving a static analysis?

- Given a baseline analysis
- improve its precision (e.g., by context sensitivity) - result set becomes smaller
- this is sound for conservative analysis. What happens if the baseline analysis is *general*?
- Assume we knew the Gold Standard
- → We cannot assess the benefits of an improved analysis for a *non-conservative* baseline analysis



Experiments - Setup

- Points-to analyses for *Java*:
  - *Spark* from the Soot-framework; inclusion-based, context-insensitive
  - *Points-to SSA*; context-insensitive (CI) and two context-sensitive variants (CS_1 and CS_2) (SCAM 2008)
- Compare the results with results from dynamic analysis
- A set of 12 benchmark programs:
  - for 4 of them, all points-to analyses are conservative
  - for one of them, Spark is not conservative (missing support for a native method)
  - for the rest, all points-to analyses are general

Experiments - Results

- Combining conservative analyses:
    - Spark and CI
    - CS_1 and CS_2
    - measurable, yet very small improvements

Experiments - Results

- Combining conservative analyses:
  - Spark and CI
  - CS_1 and CS_2
  - measurable, yet very small improvements
- For general analyses:
  - Yield (unspecified) approximations of precision, recall, accuracy
  - Improving analyses: When going from CI to CS_1 and CS_2
    - In one case, a method that is reachable in the dynamic analysis and identified as such in CI, is no longer identified as reachable in CS_1 and CS_2
    - For some fine-grained metrics, we also find "more misses"

Experiments - Results

- Combining conservative analyses:
    - Spark and CI
    - CS_1 and CS_2
    - measurable, yet very small improvements
- For general analyses:
    - Yield (unspecified) approximations of precision, recall, accuracy
    - Improving analyses: When going from CI to CS_1 and CS_2
        - In one case, a method that is reachable in the dynamic analysis and identified as such in CI, is no longer identified as reachable in CS_1 and CS_2
        - For some fine-grained metrics, we also find “more misses”
- Optimistic vs. general analysis: For one project, the result sets of many metrics are bigger for the optimistic analysis than for the static ones

Conclusion

- Combining Analyses not worth it in practice

- Static analysis is not always conservative (although often assumed)
- Comparing two analyses wrt. accuracy is possible only in special cases, or when a Gold Standard is at hand
- When improving a non-conservative baseline analysis, we need to be very careful to interpret the results
- Sometimes, dynamic analysis has bigger result sets than static analysis
  → strictly more accurate

Controversial Statement

Ostrich Algorithm applied to evaluating improvements of general analysis is ok – we "stick our head in the sand and pretend that there is no problem". Our experimental results show that the improved analyses suffer only a negligible loss of Recall.

Controversial Statement

Ostrich Algorithm applied to evaluating improvements of general analysis is ok – we “stick our head in the sand and pretend that there is no problem”. Our experimental results show that the improved analyses suffer only a negligible loss of Recall.
(I do not agree)