Analyzing Millions of Lines in Less than a Second

Amir Goharshady

HKUST CSE Research and Technology Forum
Scalable Formal Methods

Motivating Challenges

Big Code
Safety Critical Programs
Unpredictable Interdependency
Real-World Examples

Facebook
More than 40 million lines of code in a monorepo

Google
Close to a billion lines in a dozen repos

Boeing
Even the simplest analyses are intractable
Parameterization

Exploiting Sparsity in Graphs associated to Programs

```c
1 void f(int *x, int *y) {
2     y = new int(1);
3     y = new int(2);
4 }
5 int main() {
6     int *x, *y;
7     f(x, y);
8     *x += *y;
9 }
```
Inter-procedural Null Pointer Analysis

DaCapo Benchmarks

- **Our Algorithm**
- **On-demand IFDS**
- **Classical IFDS**

Average pair query time vs. Instance size graph.
Other Analyses and Optimizations

- Data-flow (e.g. live variables, reaching definitions)
- Points-to analysis and memory leak detection
- Compiler optimizations (e.g. register allocation)
- Optimization of smart contracts' gas usage
- Optimal cache management (data packing)
- Mu-calculus and LTL/HyperLTL verification
Industrial Collaborations and Grants

Strong Relations to American and European Companies Looking for Connections to Local and Chinese Industry Leaders

Facebook
SF bay area / Seattle

IBM
Europe / India

SBA
Austria

If you have a big codebase of safety-critical software, write to goharshady@cse.ust.hk