Generation techniques for consistent, realistic, diverse, and scalable graphs

Oszkár Semeráth, Kristóf Marussy,
Aren A. Babikian, Percy Chen, Gábor Szárnyas, Daniel Varró, and others
Graph models are widely used in software engineering system designs, data structures, DB content.

Testing and benchmarking scenarios rely on models.

Generating \textbf{(consistent | realistic | diverse | scalable)} models!

Generic, domain-independent generators.
Model generation setup

- Generation of valid models = challenging mathematical problem
- Generic, domain-independent model generation
Model generation: Requirements & Objectives

the CoREDiSc model
What is consistency?

- Invalid configurations

- Inconsistent models invalidate testing / benchmarking
- Constraints (queries) $\rightarrow$ validate graph models

$$\text{invalidTime}(p1,p2) := \text{replyTo}(p2,p1) \land \text{created}(p1,t1) \land \text{created}(p2,t2) \land (t2 \leq t1)$$

- Model validation + Configure generators
Consistent generators

- **Correct:** all constraints are satisfied

- **Complete:** all consistent models can be derived

- **Consistent:** Correct + Complete

Extremely challenging logic + numeric reasoning problem
Results

• We constructed scalable logic solvers for the generation of valid graph models.

Maximal model size

<table>
<thead>
<tr>
<th></th>
<th>largest model (#Objects)</th>
<th>Graph Solver</th>
<th>Sat4J</th>
<th>MiniSat</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAM+WF</td>
<td>6250</td>
<td>58</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>FAM-WF</td>
<td>7000</td>
<td>87</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Yak+WF</td>
<td>1000</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Yak-WF</td>
<td>7250</td>
<td>86</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>FS</td>
<td>4750</td>
<td>87</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Ecore</td>
<td>2000</td>
<td>38</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

Example comparison

Our solver generates ~two orders of magnitude larger models

When is Realistic?

- Cannot be distinguished from real model

- Set of generated models is close to real ones

- In custom generators, realistic nature ensured manually
Results

• We measured several graph metrics to characterize realistic models


• Configured graph generator to construct models with the same values

  Example evaluation of Multiplex participation coefficient

We were able to derive statistically similar graph models wrt. relevant metrics

Diverse

- Single models are not symmetric
  e.g. copy-paste-models (used frequently)

- The distance between any pairs of models is large
  E.g. all equivalence classes are covered

- Critical for testing graph processing systems
Results

- Proposed diversity metrics that correlate with mutation score
- Better diversity $\Leftrightarrow$ More faults detected

Results

• Proposed diversity metrics that correlate with mutation score
• Better diversity ⇔ More faults detected

Correlation between Diversity and Mutation Score in Alloy+GS+Human

Alloy (def) < Alloy (s=0) < Human < GS

Scalability

- **In size:** ability to generate huge graphs
- **In quantity:** generation time of next model does not grow

- Interactions between CoRe-DiSc elements
- Consistency is challenging
- Large inconsistent models cannot be transformed to consistent
Conclusion
Tool support

• Implemented in the VIATRA Solver Framework
• Standard EMF as input + output | visualization | config. language

• **Structural:** VIATRA Query incremental query engine
• **Numerical:** Microsoft Z3 SMT solver
• Open source: [github.com/viatra/VIATRA-Generator](https://github.com/viatra/VIATRA-Generator)