SATE V Synthetic
Method and Results

Aurelien Delaitre
SATE V Organizer
aure at nist dot gov

March 14, 2014
Disclaimer

Certain instruments, software, materials, and organizations are identified in this paper to specify the exposition adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the instruments, software, or materials are necessarily the best available for the purpose.
Input

Tool Warning
- CWE List
- Path to Sink

Test Case
- CWE
- Code Blocks
- Sinks
Location Matching

Test Case

- Bad
- Good
- Bad

Tool Warning 1
- Location 1
- Location 2
- ...

Tool Warning 2
- Location 1
- Location 2
- ...

CWE + Location Supercombo

Tool Warning
- Location 1
- Location 2
- ...

CWE ID

Disregard

CWE ID

Test Case
- Bad
- Good
- Bad

Different CWE groups

CWE Group 1
CWE Group 2
CWE Group 3
CWE + Location Supercombo

Tool Warning
- Location 1
- Location 2
- ...

Test Case
- Bad
- Good
- Bad

Located in bad code

True Positive

Same CWE group
CWE + Location Supercombo

Tool Warning
- Location 1
- Location 2
- ...

CWE ID

Test Case
- Bad
- Good
- Bad

CWE ID

Located in good code

False Positive

Same CWE group
CWE + Location Supercombo

Tool Warning
- Location 1
- Location 2
- CWE ID

Test Case
- Bad
- Good
- Bad

CWE ID

Nothing in bad code
False Negative

CWE Group 1
CWE Group 2
CWE Group 3
CWE + Location Supercombo

Test Case

- Bad
- Good
- Bad

CWE ID

Tool Warning
- Location 1
- Location 2

CWE Group 1

CWE Group 2

CWE Group 3

Nothing in good code

True Negative
Analysis Quality Trend (C)

Analysis Accuracy Trend for C

Analysis Stage

Accuracy

- Tool A
- Tool B
- Tool C
- Tool D
- Tool E
- Tool F
- Tool G
- Tool H
Analysis Quality Trend (J)

Analysis Accuracy Trend for Java

- Tool A
- Tool B
- Tool C
- Tool D

Accuracy

Analysis Stage

Stage 1
Stage 2
Stage 3
<table>
<thead>
<tr>
<th>Metrics</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>How much can I trust a tool?</td>
</tr>
<tr>
<td>Recall</td>
<td>What proportion of flaws can a tool find?</td>
</tr>
<tr>
<td>Coverage</td>
<td>What kind of flaws can a tool find?</td>
</tr>
<tr>
<td>Discrimination</td>
<td>How smart is a tool?</td>
</tr>
</tbody>
</table>
How much can I trust a tool? (C)
Proportion of flaws found by tools? (C)
Tentative Overall Performance (C)

F-Score w/ and w/o Discrimination for C

- F-Score
- Disc. F-Score
- Discrimination

F-Score (x100)

Tool A Tool B Tool C Tool D Tool E Tool F Tool G Tool H

Discrimination
How much can I trust a tool? (J)

**Precision w/ and w/o Discrimination for Java**

- **Precision**
- **Disc. Precision**
- **Discrimination**

Graph showing precision and discrimination for different tools.
Proportion of flaws found by tools? (J)
Tentative Overall Performance (J)

F-Score w/ and w/o Discrimination for Java

F-Score (x 100) vs. Discrimination

Tool A Tool B Tool C Tool D
Global vs. Condensed (C)

Global vs. Local Metrics for C

- Precision
- Gl. Recall
- Loc. Recall
- Discrimination

Graph showing the comparison of global vs. local metrics for different tools, with metrics on the y-axis and tools on the x-axis.
Global vs. Condensed (C)

Global vs. Local F-Score for C

- Global F-Score
- Local F-Score
- Discrimination

F-Score (x 100)

Discrimination

Tools: A, B, C, D, E, F, G, H
Global vs. Condensed (J)

Global vs. Local Metrics for Java

- Green: Precision
- Blue: Global Recall
- Red: Local Recall
- Orange: Discrimination

Tools:
- Tool A
- Tool B
- Tool C
- Tool D
Global vs. Condensed (J)

Global vs. Local F-Score for Java

- Gl. F-Score
- Loc. F-Score
- Discrimination

- Tool A
- Tool B
- Tool C
- Tool D
Best Covered Weakness Classes (C)

6 Best and Least Covered CWE Group for C

Cumulated Coverage

- Function call
- Expired memory
- Memory leak
- Initialization
- Type-related
- Malware-related
- Error condition
- Access control
- Cleanup
- Encapsulation

Coverage Spectrum (C)

Coverage per Tool for C

- Function call
- Expired memory
- Memory leak
- Initialization
- Type-related
- Strings
- Memory release
- Buffer operation
- Invalid pointer
- Denial of Service
- Expression
- Free of stack memory
- Code quality
- Return value
- Input validation
- Comparison
- Calculation
- Numeric errors
Best Covered Weakness Classes (J)
Coverage Spectrum (J)

Coverage per Tool for Java

Coverage

0% 25% 50% 75% 100%

Tool A Tool B Tool C Tool D

Legend:
- Comparison
- API
- Dynamic code
- Web
- Concurrency
- Expression
- Confidentiality
- Input validation
- Code quality
- Access control
- Control flow
- Invalid pointer
- Environment induced
- Path-related
- Resource management
- Credentials management
- Denial of Service
- Error condition
Overall Performance v2 (C)
Overall Performance v2 (J)
Overall Performance v2 (J)
Future Work

Metrics improvement

Introduce other aspects
  ○ Test case complexity
  ○ Overlap

More Cycles!
Conclusion

Tools differ in several dimensions

Metrics require careful development