Practical Challenges with Spreadsheet Auditing Tools

Daniel Kulesz
Institute of Software Technology
University of Stuttgart
daniel.kulesz@informatik.uni-stuttgart.de

Jan-Peter Ostberg
Institute of Software Technology
University of Stuttgart
jan-peter.ostberg@informatik.uni-stuttgart.de

4th July 2013
Agenda

- Motivation
- Our case study
- Reflections:
- Conclusion
Motivation

- Spreadsheets fail like traditional programs.
- Spreadsheets are more defect-prone.
- Static analysis can detect defects and help users better understand their programs.
- Static analysis tools are also available for spreadsheets (we refer to them as “spreadsheet auditing tools”)

Terminology borrowed
From IEEE Std. 1044-2009
Practical questions

- How reliable are the results of spreadsheet auditing tools?
- How useful are the tools for casual spreadsheet users?
- Are there areas where spreadsheet auditing tools can learn from Traditional Static Analysis Tools (TSATs)?
- Which is the best tool for scenario X?
Related work

- [Nixon and O'Hara, 2001]: Compared 5 tools on 1 single spreadsheet with 17 seeded errors.

- [Howard, 2007]: Compared 16 tools and put them in three categories. Lists capabilities of these tools without quality assessment.

- [Powell et al., 2007]: Compared 2 tools on 50 spreadsheets against manual inspections.

- [Clermont et al., 2002]: Analyzed 1 tool (their own) on 78 spreadsheets from industry.

- [Hermans, 2013]: Analyzed 1 tool (their own) on a number of spreadsheets from industry.

Has anyone tried several tools on real spreadsheets?
Our case study
Case study: project setting

- Student project with three 5\textsuperscript{th} semester, undergraduate software engineering students (under our supervision)
- Expected duration: 4 - 6 months
- Industry partner
- Systematic approach
Case study: industry partner

- Department of a large enterprise ("DEPT")
- Employ 150 technicians
- Self-developed spreadsheet templates
- Produced around 1500 spreadsheets in the past 10 years
- "four-eyes-principle"
- Few sufferings
## Case study: spreadsheet samples

<table>
<thead>
<tr>
<th>Sample</th>
<th># of Cells</th>
<th># of Worksheets</th>
<th># of Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2714</td>
<td>10</td>
<td>178</td>
</tr>
<tr>
<td>2</td>
<td>3236</td>
<td>10</td>
<td>184</td>
</tr>
<tr>
<td>3</td>
<td>3236</td>
<td>10</td>
<td>186</td>
</tr>
<tr>
<td>4</td>
<td>610</td>
<td>2</td>
<td>117</td>
</tr>
<tr>
<td>5</td>
<td>2151</td>
<td>6</td>
<td>182</td>
</tr>
<tr>
<td>6</td>
<td>2990</td>
<td>4</td>
<td>208</td>
</tr>
<tr>
<td>7</td>
<td>2092</td>
<td>4</td>
<td>119</td>
</tr>
<tr>
<td>8</td>
<td>2356</td>
<td>4</td>
<td>225</td>
</tr>
<tr>
<td>9</td>
<td>1080</td>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>2034</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>11</td>
<td>6625</td>
<td>7</td>
<td>2432</td>
</tr>
</tbody>
</table>
Case study: the results

- Found 14 tools
- Reduced to 3 using K.O.-criteria
- Remaining tools evaluated in-depth:

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Weight</th>
<th>Tool 1</th>
<th>Tool 2</th>
<th>Tool 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configurability</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Check color conventions</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Findings presentation</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UI Language</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Easy of use</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Learning curve</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Usefulness of results</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>22</strong></td>
<td><strong>9</strong></td>
<td></td>
<td><strong>16</strong></td>
</tr>
</tbody>
</table>

Reflections
Why reflect?

- Case study: one case
- Reflection goal: Stress widespread issues and foster discussion about possible solution approaches
- Reflection based on impressions and experiences gained from the case study (including all 14 tools)
- Learn from TSATs?
Positive impressions

- Easy installation
- Trivial smell patterns work
- Feature unhiding helps
Less positive impressions

1. Presentation of findings
2. (Handling) False positives
3. Non-intrusiveness
4. Understandability of implemented rules
5. Configurability and result inter-comparability
6. Lack of unification
7. Licensing costs versus risk perception
8. Compatibility and portability
9. Localization
9. Localization

- German-speaking department
- Not a single tool in German
8. Compatibility and portability

- Execution environment glues
- Excel 2003 vs. Excel 2007
- In TSATs: environment decoupling
7. Licensing costs vs. risk perception

- Tools are pretty expensive.
- Industry partner didn't perceive spreadsheets as a high risk.
- “Wrong” industry partner, because they don't employ “frankensheets” like some other enterprises?

- Largest spreadsheet less than 7000 cells
  - => extremely small
- Our experience: customers have sheets with sizes between 80-250MB
- These sheets contain upwards of one hundred sheets and ten thousands of cells.
6. Lack of unification

- No “swiss army knife”
- TSATs use aggregation approach:

Example: “conQAT”
5. Configurability, inter-comparability

- Different tool, different numbers:

<table>
<thead>
<tr>
<th>Konstanten in Formeln</th>
<th>SP</th>
<th>RA</th>
<th>ETC</th>
<th>ETC(c)</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>spreadsheet_sample_01</td>
<td>67</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>spreadsheet_sample_02</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
</tr>
<tr>
<td>spreadsheet_sample_03</td>
<td>27</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>spreadsheet_sample_04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>spreadsheet_sample_05</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>spreadsheet_sample_06</td>
<td>18</td>
<td>6</td>
<td>18</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>spreadsheet_sample_07</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>spreadsheet_sample_08</td>
<td>33</td>
<td>0</td>
<td>34</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>spreadsheet_sample_09</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>spreadsheet_sample_10</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>spreadsheet_sample_11</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>spreadsheet_sample_12</td>
<td>16</td>
<td>5</td>
<td>16</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

SP = “Spreadsheet Professional”
RA = “Rainbow Analyst”
ETC = our tool
ETC(c) = our tool with different setup
NMI = manual inspection


- TSATs are more configurable and comparable!
4. Understandability of impl. rules

- Tools: pattern match = “errors”
  → inappropriate term
- TSATs have the same problem.
- Risk: Users tend to ignore warnings or dismiss them if they don't understand them!
3. Non-intrusiveness

- Tools often try to report results in one of these ways:
  - Coloring cells
  - Adding comments
  - Adding additional “report” worksheets
  - Adding macros

- Risk:
  - “Enriched” spreadsheet not alterable
  - Potential side-effects

- Many TSATs do a better job by using IDE-overlays.
2. (Handling) False positives

- High false positive rates! (78-90%)
- But unavoidable, thus:
  - Tools must allow for handling false positives!
- False positive handling missing in most tools!
- TSATs are (slightly) better:
  - Lower false positive rates
  - Rudimentary support for handling false positives
1. Presentation of findings

- Example:

  Before coloring

  After coloring

- Issues: Overuse, customization, transport wrong message
1. Presentation of findings (cont'd)

- Issues with “finding lists”:
  - wrong abstraction level
  - hardly navigable
  - too distant from the source
- TSATs have similar problems but better approaches.
Conclusion

- Spreadsheet auditing tools have great potential, but there is still lots of room for improvement:
  - Reliability (Detection techniques)
  - Understandability (Presentation of results not very useful for end-users.)
  - Usability (Handling of false positives)
- Looking at TSATs might be a good idea (but does not automatically guarantee good solutions).