SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality

Birgit Hofer, Graz University of Technology
Why spreadsheets?

- Used in nearly every company
- Basis for decisions
- Error prone
  - 3-5 % chance to make a fault in a formula
  - 88 % of spreadsheets contain faults
- Hard to debug
  - Size of spreadsheets
  - Structure hidden

*Spectrum-based Fault Localization (SFL) can help to narrow down the search space.*
Outline

1. SFL for Spreadsheets
2. Initial Situation
3. Research questions
   - RQ1: Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?
   - RQ2: If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?
   - RQ3: Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?
4. Future work
Test Cases for Spreadsheets

- **Input cells:** cells that do not reference other cells
  \[ I = \{B2=23, C2=31, E2=15, B3=35, C3=34, E3=17\} \]

- **Output cells:** any formula cell, determined by user
  \[ O = \{B4=58, C4=65, D4=123, F2=810, F3=1173\} \]
SFL for Spreadsheets (1)

Program debugging: execution traces, slices

Spreadsheets: cones (borrowed from hardware debugging)

\[
\text{CONE}(c) = c \cup \bigcup_{c' \in \rho(c)} \text{CONE}(c')
\]

The function \( \rho(c) \) returns all cells referenced in \( c \).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>week 1</td>
<td>week 2</td>
<td>Total</td>
<td>$/h</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>23</td>
<td>31</td>
<td>=SUM(B2)</td>
<td>15</td>
<td>=D2*E2</td>
</tr>
<tr>
<td>3</td>
<td>Jones</td>
<td>35</td>
<td>34</td>
<td>=SUM(B3:C3)</td>
<td>17</td>
<td>=D3*E3</td>
</tr>
<tr>
<td>4</td>
<td>Total</td>
<td>=SUM(B2:B3)</td>
<td>=SUM(C2:C3)</td>
<td>=SUM(D2:D3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{CONE}(F2) = \{B2, D2, E2, F2\}
\]
SFL for Spreadsheets (2)

**Spectra:** Cones of erroneous and correct output cells

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>week 1</td>
<td>week 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>23</td>
<td>31</td>
<td>=SUM(B2)</td>
<td>15</td>
<td>=D2*E2</td>
</tr>
<tr>
<td>3</td>
<td>Jones</td>
<td>35</td>
<td>34</td>
<td>=SUM(B3:C3)</td>
<td>17</td>
<td>=D3*E3</td>
</tr>
<tr>
<td>4</td>
<td>Total</td>
<td>=SUM(B2:B3)</td>
<td>=SUM(C2:C3)</td>
<td>=SUM(D2:D3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CONE(F2) = {B2,D2,E2,F2}
CONE(D4) = {B2,D2,B3,C3,D3,D4}
CONE(B4) = {B2,B3,B4}
CONE(C4) = {C2,C3,C4}
CONE(F3) = {B3,C3,D3,E3,F3}
SFL for Spreadsheets (3)

\[
\text{CONE}(F2) = \{B2, D2, E2, F2\}
\]

\[
\text{CONE}(D4) = \{B2, D2, B3, C3, D3, D4\}
\]

\[
\text{CONE}(B4) = \{B2, B3, B4\}
\]

\[
\text{CONE}(C4) = \{C2, C3, C4\}
\]

\[
\text{CONE}(F3) = \{B3, C3, D3, E3, F3\}
\]

\[
\begin{align*}
a_{11}(c) &= \{c' \mid c \in \text{CONE}(c') \land c' \text{ is erroneous}\} \\
a_{10}(c) &= \{c' \mid c \in \text{CONE}(c') \land c' \text{ is correct}\} \\
a_{01}(c) &= \{c' \mid c \notin \text{CONE}(c') \land c' \text{ is erroneous}\}
\end{align*}
\]

\[
Ochiai(c) = \frac{a_{11}(c)}{\sqrt{(a_{11}(c) + a_{01}(c))(a_{11}(c) + a_{10}(c))}}
\]
Evaluation Methods

- Best Case
  \[ \text{Rank}_{\text{Best}} = | \{ c \mid Ochiai(c) > f \} | + 1 \]

- Average Case
  \[ \text{Rank}_{\text{AVG}} = | \{ c \mid Ochiai(c) > f \} | + \frac{| \{ c \mid Ochiai(c) = f \} |}{2} + 0.5 \]

- Worst Case
  \[ \text{Rank}_{\text{WORST}} = | \{ c \mid Ochiai(c) \geq f \} | \]
Initial situation – Average SFL Ranking

No user wants to indicate for so many output cells if they are correct.

Outline

1. SFL for Spreadsheets
2. Initial Situation
3. Research questions
   - RQ1: Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?
   - RQ2: If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?
   - RQ3: Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?
4. Future work
RQ1: Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

EUSES my_financial_model
**RQ1:** Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

**ISCAS85 c7552**

![Graph showing the relationship between the number of correct output cells and fault ranking](image-url)

The graph illustrates the distribution of fault rankings for different numbers of correct output cells.
**RQ2:** If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?

- Avoid coincidental correct output cells
  - A-priori definition not possible
  - Too many potential coincidental correct output cells

- Take output cells with largest cones first

<table>
<thead>
<tr>
<th>Rank_{AVG} for one correct output cell</th>
<th>Random selection</th>
<th>Largest cone</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUSES my_financial</td>
<td>5.8</td>
<td>2.5</td>
</tr>
<tr>
<td>ISCAS85 c7552</td>
<td>100.6</td>
<td>69.5</td>
</tr>
</tbody>
</table>
**RQ3:** Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?

→ Balance ratio of correct and erroneous output cells
→ Duplicate cones of erroneous output cells

**ISCAS85 c7552**
Future work

- Extend evaluation
  - All EUSES and ISCAS85 spreadsheets

- Weighting instead of duplication
  - Wong’s Heuristic III

- Several faults
  - Influence on ranking
  - Clustering
Summary

Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

YES

If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?

YES (use largest cones first)

Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?

YES (duplicate cones of erroneous output cells)

<table>
<thead>
<tr>
<th>Rank_{AVG} (1 correct output cell)</th>
<th>Random</th>
<th>Largest cone</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUSES my_financial</td>
<td>5.8</td>
<td>2.5</td>
</tr>
<tr>
<td>ISCAS85 c7552</td>
<td>100.6</td>
<td>69.5</td>
</tr>
</tbody>
</table>