Improvements for Spectrum-based Fault Localization in Spreadsheets

Elisabeth Getzner

June 18, 2015

Outline

1. Motivation
2. Fault Localization in Spreadsheets
3. Improvements for SFL
4. Evaluation
5. Conclusion

Outline

1. Motivation
2. Fault Localization in Spreadsheets
3. Improvements for SFL
4. Evaluation
5. Conclusion

Motivation

Spreadsheets are . . .

- used privately and in corporate environment
- used for critical computations and decisions[1]
- faulty! (~88 % of all spreadsheets)[2]

Quality assurance in spreadsheets:

- Fault detection, localization, repair

1 James Kwak. The Importance of Excel. The Baseline Scenario. @. Feb. 9, 2013. URL: http://baselinescenario.com/2013/02/09/the-importance-of-excel/ (visited on 03/31/2015)

2 Raymond R. Panko. “Spreadsheet Errors: What We Know. What We Think We Can Do”. In: Proceedings of the European Spreadsheet Risks Interest Group (EuSpRIG). 2000, pp. 7–17. URL: http://arxiv.org/abs/0802.3457 (visited on 04/08/2014)

Example - Bonus Calculation

	A	B	C	D	E
1		**Hours**	**Salary**	**Bonus**	**Sum**
2	Jones	17	272	26	298
3	Smith	13	208	0	208
4	Rogers	20	320	40	360
5	**Total**		800	66	866

	A	B	C	D	E
1		**Hours**	**Salary**	**Bonus**	**Sum**
2	Jones	17	=B2*16	=IF(B2>15; C3 /8;0)	=SUM(C2:D2)
3	Smith	13	=B3*16	=IF(B3>15; C3 /8;0)	=SUM(C3:D3)
4	Rogers	20	=B4*16	=IF(B4>15; C4 /8;0)	=SUM(C4:D4)
5	**Total**		=SUM(C2:C4)	=SUM(D2:D4)	=SUM(E2:E4)

Figure: Faulty bonus calculation

Example - Bonus Calculation

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	272	26	298
3	Smith	13	208	0	208
4	Rogers	20	320	40	360
5	Total		*800*	*66*	*866*

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	=B2*16	=IF(B2>15; C3 /8;0)	=SUM(C2:D2)
3	Smith	13	=B3*16	=IF(B3>15; C3 /8;0)	=SUM(C3:D3)
4	Rogers	20	=B4*16	=IF(B4>15; C4 /8;0)	=SUM(C4:D4)
5	Total		*=SUM(C2:C4)*	*=SUM(D2:D4)*	*=SUM(E2:E4)*

Figure: `D2` is faulty (26 instead of 34)

Outline

1. Motivation

2. Fault Localization in Spreadsheets

3. Improvements for SFL

4. Evaluation

5. Conclusion

Spectrum-based Fault Localization I

Goal: Find root cause of unexpected spreadsheet behavior

- Trace-based (as opposed to model-based)
  - Analyze cell dependencies
  - Return fault likelihoods for each cell
- Process:
  1. Testing Decisions
  2. Analyze dependencies (CONEs)
  3. Compute fault likelihood (similarity coefficient)

Spectrum-based Fault Localization II

1. **Testing Decisions** (TD)
   - User provided
   - Judging the value of cells
     - Expected (✓) $= TD^+$
     - Unexpected (✗) $= TD^-$

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	272	26	298
3	Smith	13	208	0	✓ 208
4	Rogers	20	320	40	360
5	Total		✓ 800	66	✗ 866

Spectrum-based Fault Localization III

2. Create CONEs from the testing decisions

- CONE(c) = Set of cells containing c and all cells referenced by c
    - directly (in formula) and
    - indirectly (recursive)
- CONE(E5) = {E5,E2,E3,E4,D2,D3,D4,C2,C3,C4}

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	272	26	298
3	Smith	13	208	0	208
4	Rogers	20	320	40	360
5	Total		***800***	***66***	✗ ***866***

Spectrum-based Fault Localization IV

3. Similarity coefficient correlates

   - No. TD^{+} and the
   - No. TD^{-} a cell contributes to

Using the **Ochiai**[1] coefficient:

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	0.7	1	1
3	Smith	13	0.6	0.7	0.7
4	Rogers	20	0.7	1	1
5	Total		0	0	1

Rank	Cells
1.	D2,E2,D4,E4,E5
2.	C2,D3,E3,C4
3.	C3

1 R. Abreu, P. Zoeteweij, and A.J.C. van Gemund. "An Evaluation of Similarity Coefficients for Software Fault Localization". In: 12th Pacific Rim International Symposium on Dependable Computing, 2006. PRDC '06. Dec. 2006, pp. 39–46. DOI: 10.1109/PRDC.2006.18

SFL Properties

Advantages

- Fast
- Low user requirement
- Intuitive cell ranking

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	272	26	298
3	Smith	13	208	0	✓ 208
4	Rogers	20	320	40	360
5	Total		✓ 800	66	✗ 866

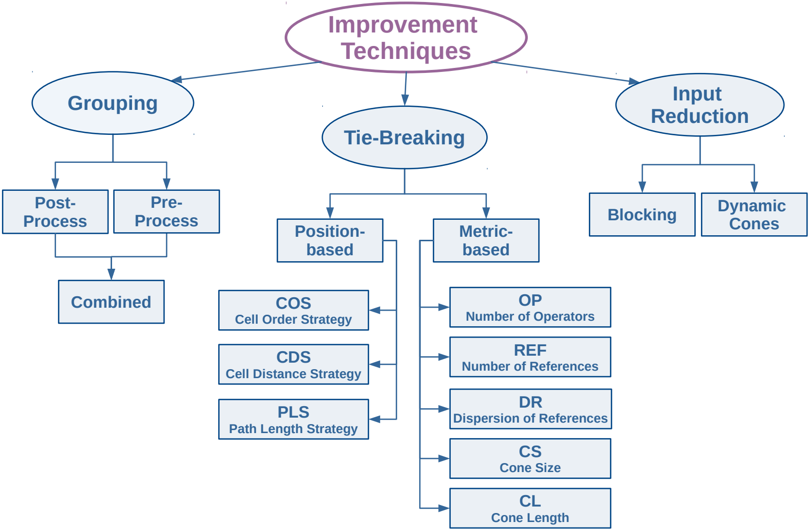
	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	0.7	1	1
3	Smith	13	0.6	0.7	0.7
4	Rogers	20	0.7	1	1
5	Total		0	0	1

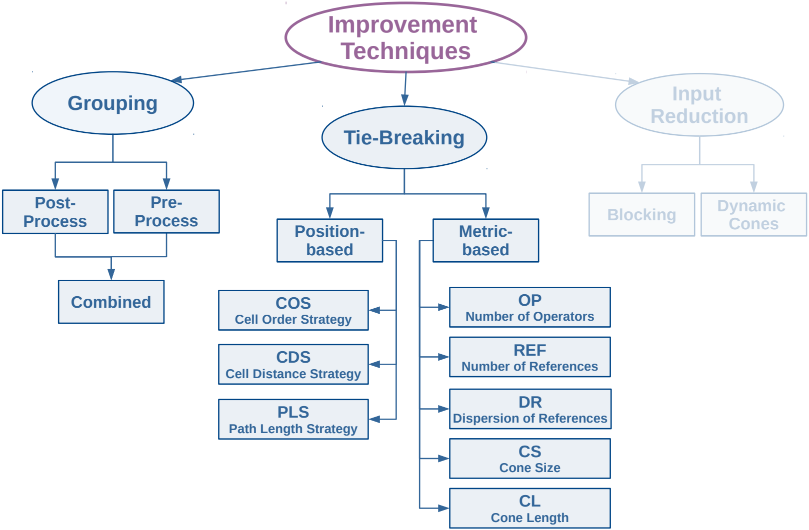
Issues

- Multiple fault interference
- Low rank of the faulty cell
  - Oracle mistakes
  - Coincidental correctness
- Large Ties
  - Lack of prioritization
  - Difficult to compare

Outline

1. Motivation
2. Fault Localization in Spreadsheets
3. Improvements for SFL
4. Evaluation
5. Conclusion


Improvement Techniques
Grouping
Tie-Breaking
Input Reduction
Post-Process
Pre-Process
Combined
Position-based
Metric-based
COS
Cell Order Strategy
CDS
Cell Distance Strategy
PLS
Path Length Strategy
OP
Number of Operators
REF
Number of References
DR
Dispersion of References
CS
Cone Size
CL
Cone Length
Blocking
Dynamic Cones


Improvement Techniques
Grouping
Tie-Breaking
Input Reduction
Post-Process
Pre-Process
Combined
Position-based
Metric-based
Blocking
Dynamic Cones
COS
Cell Order Strategy
CDS
Cell Distance Strategy
PLS
Path Length Strategy
OP
Number of Operators
REF
Number of References
DR
Dispersion of References
CS
Cone Size
CL
Cone Length

Grouping I

Goal: Group cell areas with duplicate formulas to a single unit

- Formulas must be identical in `R1C1`

	1	2	3	4	5
1		Hours	Salary	Bonus	Sum
2	Jones	17	=RC[-1]*16	=IF(RC[-2]>15; R[1]C[-1] /8;0)	=SUM(RC[-2]:RC[-1])
3	Smith	13	=RC[-1]*16	=IF(RC[-2]>15; RC[-1] /8;0)	=SUM(RC[-2]:RC[-1])
4	Rogers	20	=RC[-1]*16	=IF(RC[-2]>15; RC[-1] /8;0)	=SUM(RC[-2]:RC[-1])
5	Total		=SUM(R[-3]C:R[-1]C)	=SUM(R[-3]C:R[-1]C)	=SUM(R[-3]C:R[-1]C)

Figure: Four groups with the faulty cell isolated

- Post- vs. Pre-Processing

Grouping II

- **Post-Process** Grouping
  - Analyze spreadsheet after SFL is applied
  - Groupable cells must have the same similiarity coefficient
- **Pre-Process** Grouping
  - Analyze spreadsheet before SFL is applied
  - Copy testing decisions to all cells in group
  - Cells can only be grouped if they work on the same type of data

Pre-Process Grouping Example

Type-safe group is an area containing

- Constants of the same type (i.e. int, string, . . .)
- Formula cells
    - Share the same formula **and**
    - All references share same type

	1	2	3	4	5
1		Hours	Salary	Bonus	Sum
2	Jones	17	=RC[-1]*16	=IF(RC[-2]>15; R[1]C[-1] /8;0)	=SUM(RC[-2]:RC[-1])
3	Smith	13	=RC[-1]*16	=IF(RC[-2]>15; RC[-1] /8;0)	=SUM(RC[-2]:RC[-1])
4	Rogers	20	=RC[-1]*16	=IF(RC[-2]>15; RC[-1] /8;0)	=SUM(RC[-2]:RC[-1])
5	Total		*=SUM(R[-3]C:R[-1]C)*	*=SUM(R[-3]C:R[-1]C)*	*=SUM(R[-3]C:R[-1]C)*

Figure: Three groups, isolating the row with the faulty cell D2.

Tie Breaking I

Goal: Rank faulty cell higher than non-faulty cells

- Position- vs. Metric-based Tie-Breaking
- **Position-based** TB measures distances / path lengths between cells
  - **COS** (Cell Order Strategy):
    Euclidean distance from top-left corner `A1`
  - **CDS** (Cell Distance Strategy):
    Euclidean distance from nearest TD^{-}
  - **PLS** (Path Length Strategy):
    Number of cell references to reach TD^{-}

Tie Breaking II

- **Metric-based** TB analyzes formulas, using heuristics to find fault likelihood
  - **OP, REF**: Number of Operators / References
  - **DR** (Dispersion of References): Referenced cells/areas where coordinates do not overlap with the referencing cell → higher fault likelihood

	A	B	C	D	E
1		Hours	Salary	Bonus	Sum
2	Jones	17	=B2*16	=IF(B2>15; C3 /8;0)	=SUM(C2:D2)
3	Smith	13	=B3*16	=IF(B3>15; C3 /8;0)	=SUM(C3:D3)
4	Rogers	20	=B4*16	=IF(B4>15; C4 /8;0)	=SUM(C4:D4)
5	Total		=SUM(C2:C4)	=SUM(D2:D4)	=SUM(E2:E4)

  - **CS, CL** (Cone Size/Length): Number of cells needed to compute the cell value

Outline

1. Motivation
2. Fault Localization in Spreadsheets
3. Improvements for SFL
4. Evaluation
5. Conclusion

Measuring Success I

Compare position of faulty cell in the ranking for SFL alone and with our strategies

- Worst Case Scenario
  - Faulty cell c_f is reached last in the tie

$$\text{RELRANK}_{\text{worst}} = \frac{|\{c \in \text{CELLS} : SC(c) \geq SC(c_f)\}|}{|C_F \subseteq \text{CELLS}|}$$

  - C_F = Formula cells in the Spreadsheet
  - SC = Similarity Coefficient
  - Used for cumulative histogram
  - Emphasizes even small improvements

Measuring Success II

- Average Case Scenario
  - User inspects half of the equally ranked, non-faulty cells before reaching the faulty cell
  - Comparison to “pure chance”
  - Risk analysis with Impact

$$\mathit{Impact} = \textsc{RelRank}_{\text{avg}}^{\mathit{before}} - \textsc{RelRank}_{\text{avg}}^{\mathit{after}}$$

- positive Impact: fault is ranked in the first half of the tie
- negative Impact: fault is ranked in the second half

Evaluation Corpora

1. EUSES: many, diverse, real spreadsheets
2. INFO: student submissions for an Excel course
3. BURNETT: user study with two small spreadsheets

Feature	EUSES	INFO	BURNETT
Spreadsheet size	diverse	large	small
TD origin	injected	injected	authentic
Fault origin	injected	authentic	injected
Grouping	★★	★★★	-
Tie-Breaking	★	★★	★

Grouping Strategies (INFO)

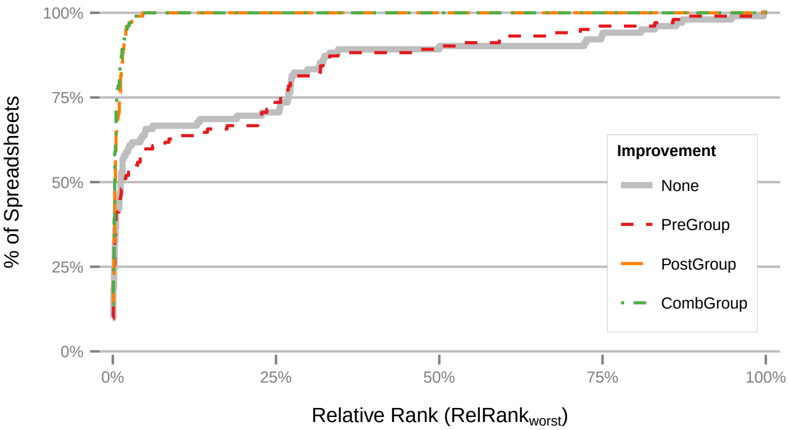


Figure: Cumulative Histogram for the RELRANK$_{\text{worst}}$ in INFO

Grouping Strategies (EUSES)

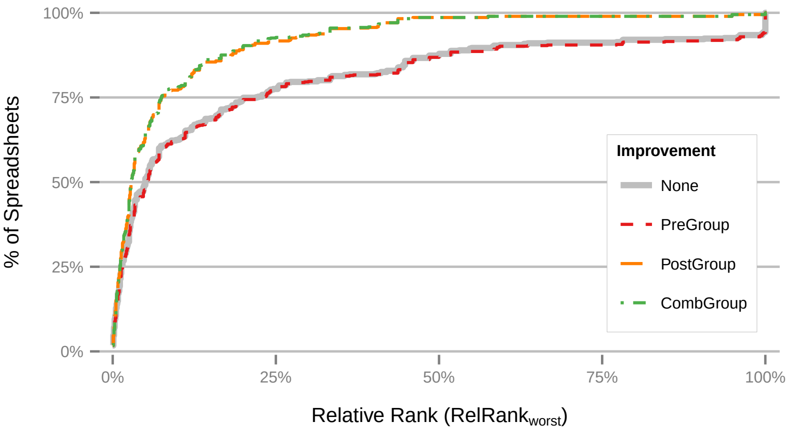


Figure: Cumulative Histogram for the RELRANK$_{\text{worst}}$ in EUSES

Impact Analysis

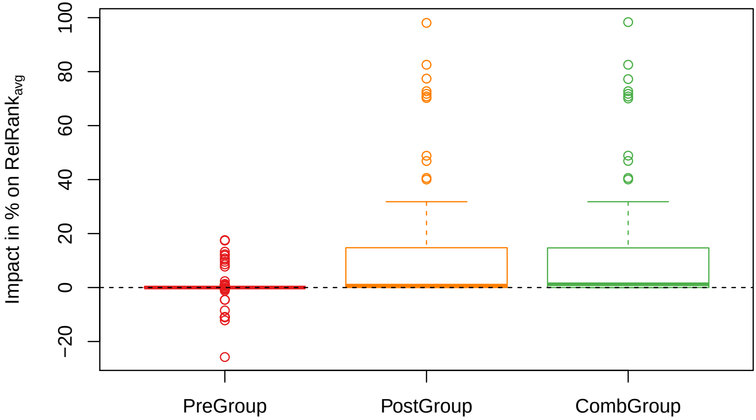


Figure: Boxplot on the Impact on the INFO corpus

Position-based Tie-Breaking

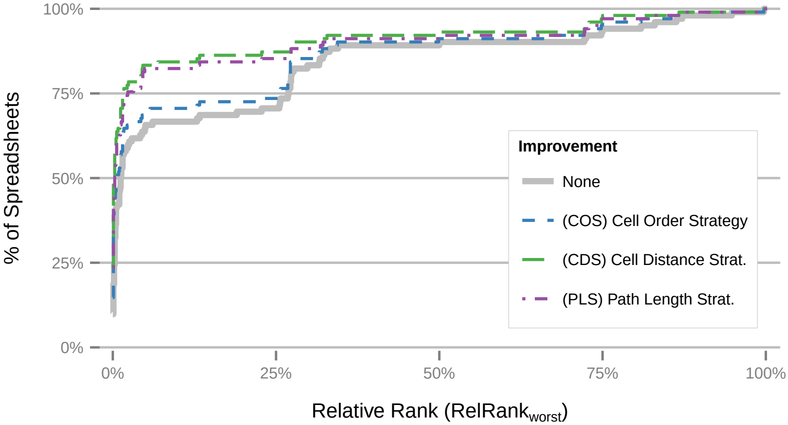


Figure: Cumulative Histogram for the RELRANK$_{worst}$ in INFO

Metric-based Tie-Breaking (OP, REF, DR)

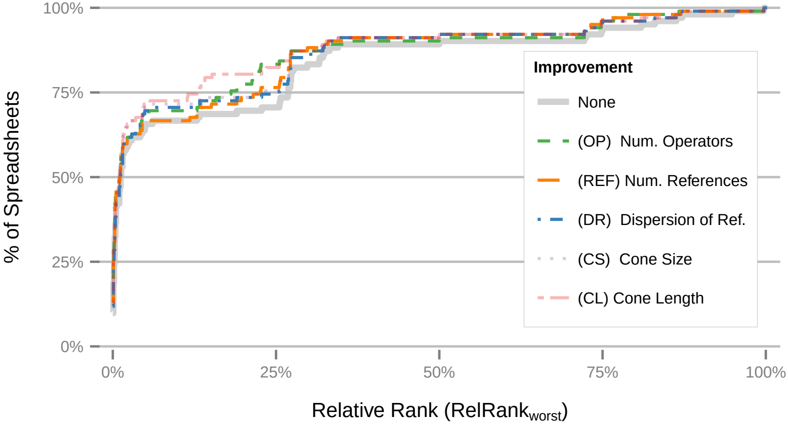


Figure: Cumulative Histogram for the RELRANK$_{worst}$ in INFO

Metric-based Tie-Breaking (CS, CL)

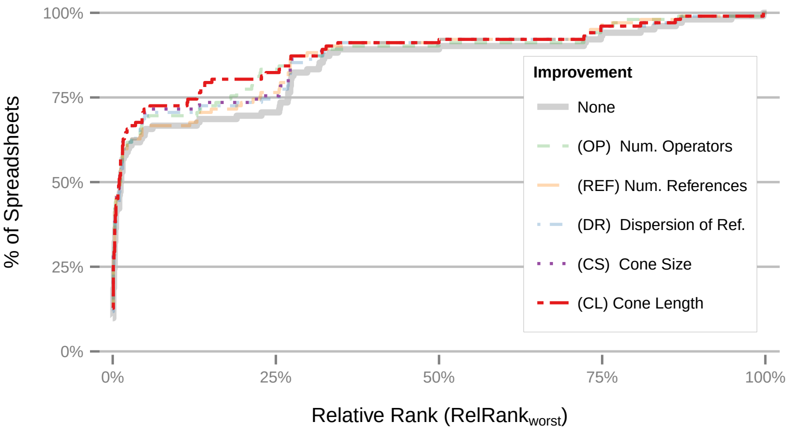


Figure: Cumulative Histogram for the RELRANK$_{worst}$ in INFO

Tie-Breaking Impact

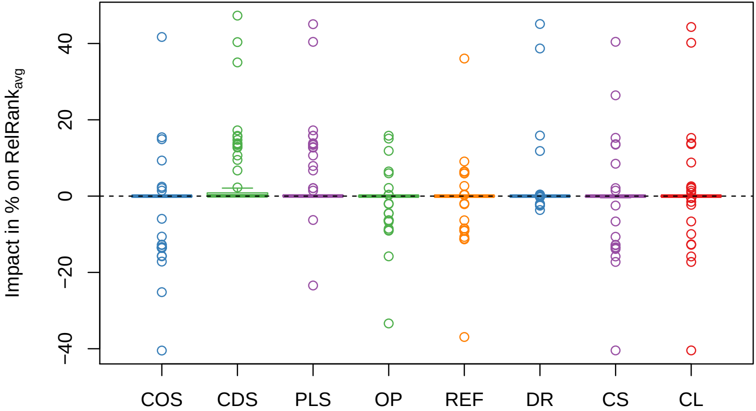
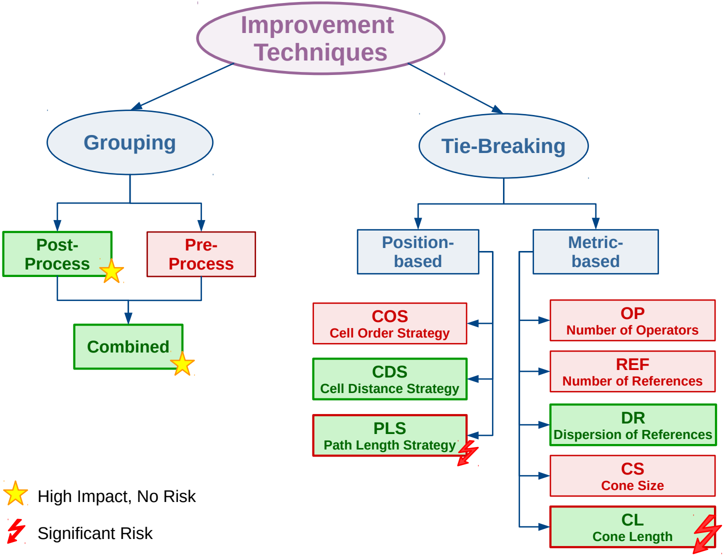


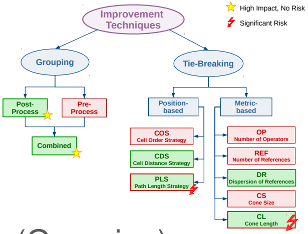
Figure: Boxplot on the Impact on the INFO corpus


Improvement Techniques
Grouping
Tie-Breaking
Post-Process
Pre-Process
Combined
Position-based
Metric-based
COS
Cell Order Strategy
CDS
Cell Distance Strategy
PLS
Path Length Strategy
OP
Number of Operators
REF
Number of References
DR
Dispersion of References
CS
Cone Size
CL
Cone Length
High Impact, No Risk
Significant Risk

Outline

1. Motivation
2. Fault Localization in Spreadsheets
3. Improvements for SFL
4. Evaluation
5. Conclusion

Conclusion



- SFL issues (e.g. large ties)
- Improvement techniques
  - Realistic representation of the tie (Grouping)
  - Rank faulty cell high within tie (Tie-Breaking)
    - Correlation to TDs (CDS)
    - Specialized metrics offer lower risk (DR)
  - Need authentic faults/TDs to evaluate
    - Structural properties influence result
    - 2 new, publicly available spreadsheet corpora[1]
- Future Work
  - Improving Pre-Process Grouping
  - Combination of techniques

1 `spreadsheets.ist.tugraz.at`