Clustering Test Cases to Achieve Effective & Scalable Prioritisation Incorporating Expert Knowledge

Yuening Hu
University of Maryland, College Park
ynhu@cs.umd.edu

Nov. 24, 2009

Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions

Background

- Test case prioritisation
  - Regression test
  - An efficient ordering of test cases
- Ideal ordering
  - Reveal faults earliest
  - Not known in advance

Background

- Available criteria
  - Structural coverage
  - Requirement priority
  - Mutation score
- Powerful expert judgement
  - Human tester
  - Rich domain knowledge
  - Human guidance to avoid bias
  - Techniques: Analytic Hierarchy Process

Motivation

- Analytic Hierarchy Process (AHP)
  - Assumption: Human involvement
    → prioritisation improvement
  - Pair-wise comparison
  - Scalability challenges: 100 meaningful comparisons
  - Usually much more than 100

AHP-based prioritisation

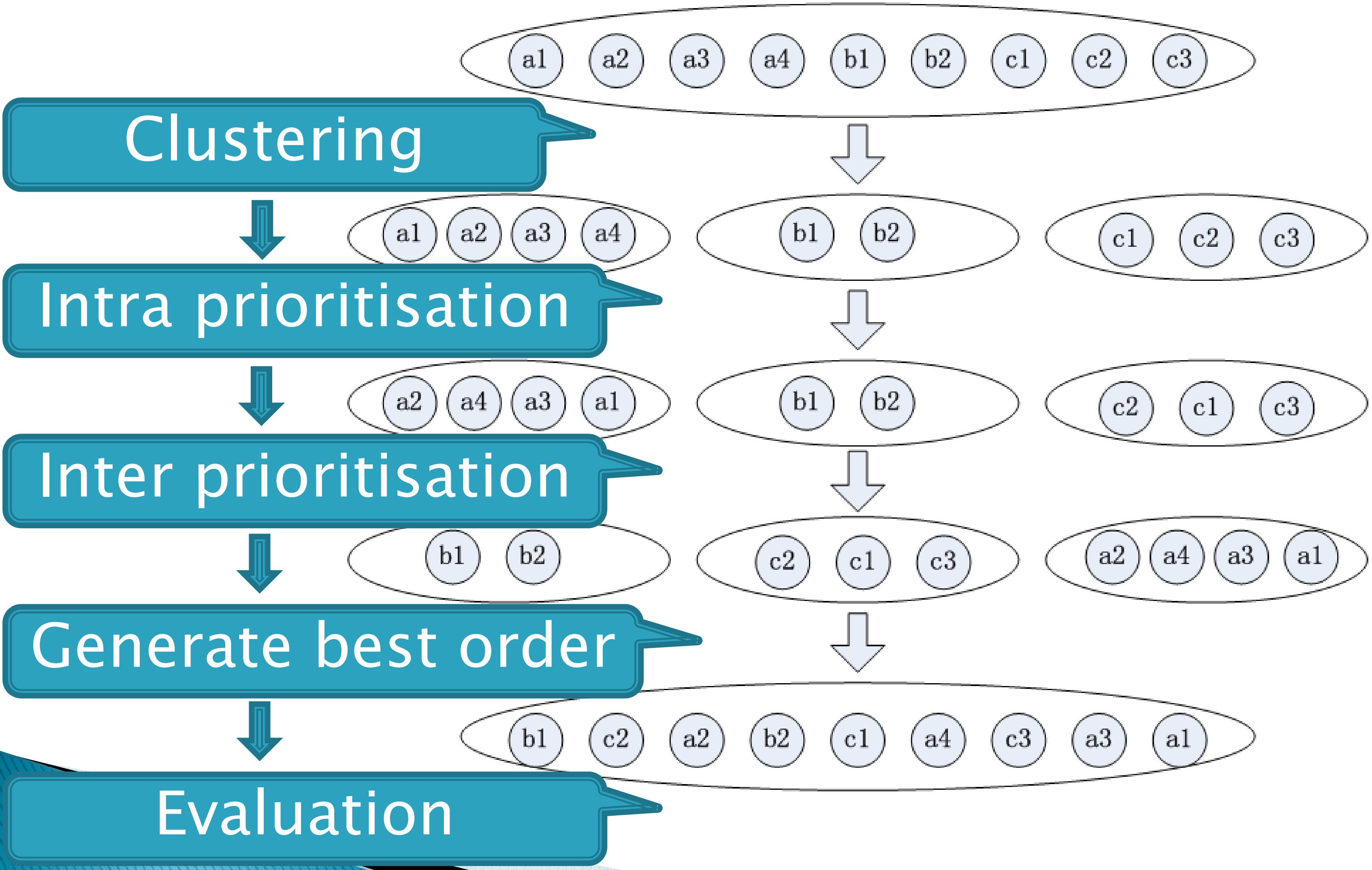
Clustering to control the number of comparisons

Expert-guided prioritisation

Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions

Framework

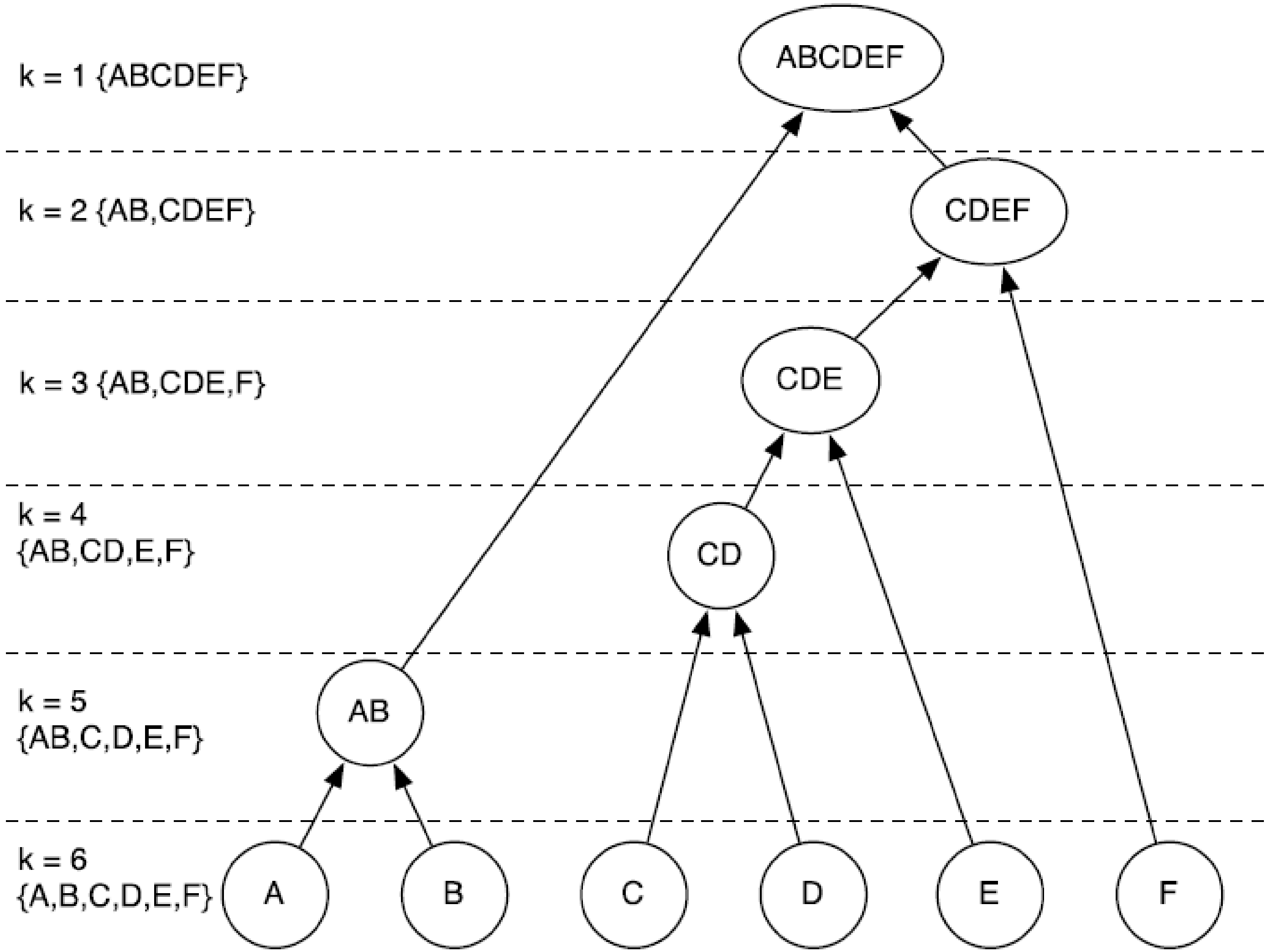


a1 a2 a3 a4 b1 b2 c1 c2 c3
Clustering
a1 a2 a3 a4
b1 b2
c1 c2 c3
Intra prioritisation
a2 a4 a3 a1
b1 b2
c2 c1 c3
Inter prioritisation
b1 b2
c2 c1 c3
a2 a4 a3 a1
Generate best order
b1 c2 a2 b2 c1 a4 c3 a3 a1
Evaluation

Clustering

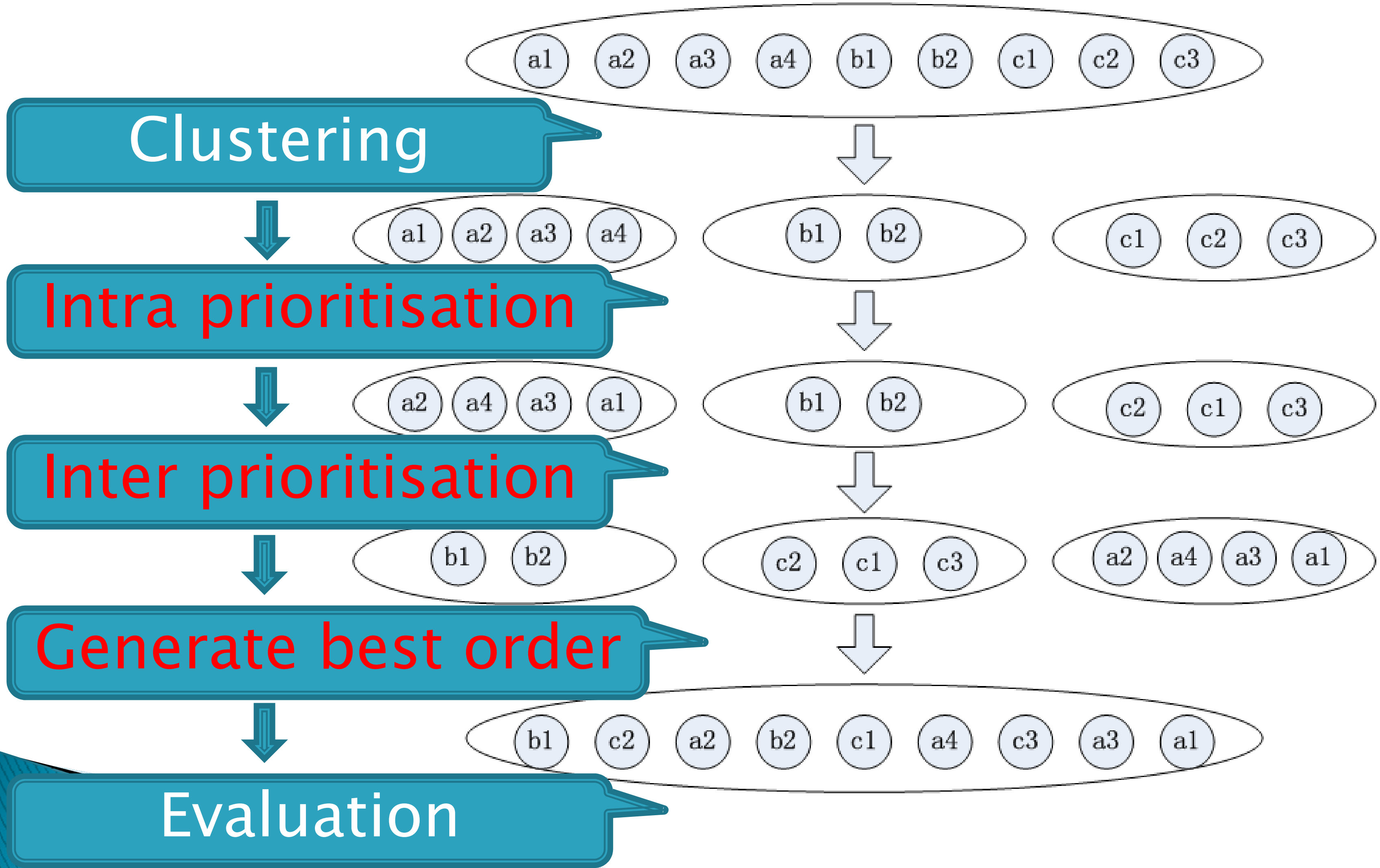
- Ideal clustering criterion
  - Similarity of detected faults
- Used clustering criterion
  - One bit per statement: 1/0
  - Binary string of each test cases
  - Hamming distance

Clustering



k = 1 {ABCDEF}
k = 2 {AB,CDEF}
k = 3 {AB,CDE,F}
k = 4
{AB,CD,E,F}
k = 5
{AB,C,D,E,F}
k = 6
{A,B,C,D,E,F}
ABCDEF
CDEF
CDE
CD
AB
A
B
C
D
E
F

Framework

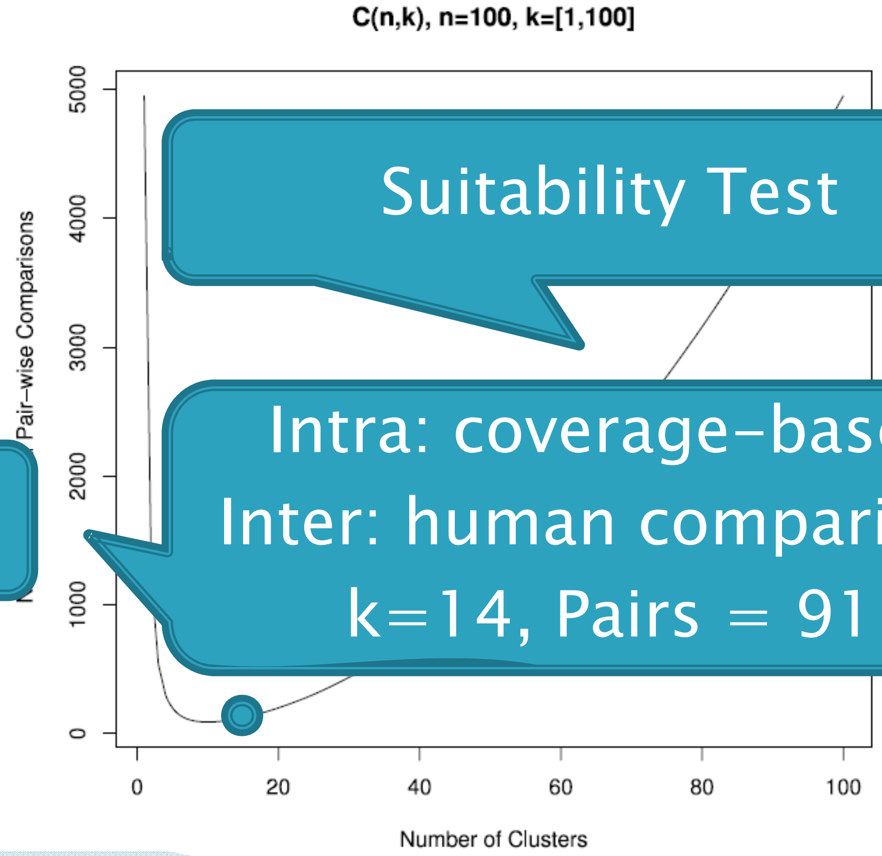

a1 a2 a3 a4 b1 b2 c1 c2 c3
Clustering
a1 a2 a3 a4
b1 b2
c1 c2 c3
Intra prioritisation
a2 a4 a3 a1
b1 b2
c2 c1 c3
Inter prioritisation
b1 b2
c2 c1 c3
a2 a4 a3 a1
Generate best order
b1 c2 a2 b2 c1 a4 c3 a3 a1
Evaluation

Clustering Based Prioritisation

- Interleaved Clusters Prioritisation (ICP)
  - Intra-cluster prioritisation
  - Inter-cluster prioritisation
  - Comparison limit: 100 pairs

Clustering Based Prioritisation

- Interleaved Clusters Prioritisation (ICP)
  - n test cases, k clusters
  - Pairs: k(k−1)/2 + k(n/k)(n/k−1)/2



k=17

Pairs = 381 > 100

Suitability Test

Intra: coverage-based
Inter: human comparison
k=14, Pairs = 91

Analytic Hierarchy Process

- Analytic Hierarchy Process, AHP
  - Not transitive
  - Ratio-based

p_{ij}	Preference
1	i is equally preferable to j
3	i is slightly preferable over j
5	i is strongly preferable over j
7	i is very strongly preferable over j
9	i is extremely preferable over j

Analytic Hierarchy Process

- Comparison Matrix M

$$\forall i(1 \leq i \leq n)\forall j(1 \leq j \leq n \wedge i \neq j), M(i, j) = p_{ij}$$

$$M(i, i) = 1(1 \leq i \leq n).$$

- Column normalized M

$$M'(i, j) = \frac{M(i, j)}{\sum_{1 \leq k \leq n} M(i, k)}$$

- Priority weighting vector

$$E_i = \frac{\sum_{1 \leq k \leq n} M(k, i)}{n}$$

Analytic Hierarchy Process

- Ideal User Model
  - tA detected nA faults
  - tB detected nB faults

Condition	p_{AB}	Description
$n_A = n_B$	1	Equal
$n_A > 0$ and $n_B = 0$	7	Very Strongly prefer t_A
$n_A > 0, n_B > 0, n_A \geq 3n_B$	9	Extremely prefer t_A
$n_A > 0, n_B > 0, n_A \geq 2n_B$	7	Very Strongly prefer t_A
$n_A > 0, n_B > 0, n_A \geq n_B$	5	Strongly prefer t_A

$$p_{BA} = \frac{1}{p_{AB}}$$

Analytic Hierarchy Process

- Human Error Model
  - Only type 1 ~ 6 considered

Type	Original	Error
1	$p_{AB} > 1$	$p'_{AB} = 1$
2	$p_{AB} < 1$	$p'_{AB} = 1$
3	$p_{AB} > 1$	$p'_{AB} < 1$
4	$p_{AB} < 1$	$p'_{AB} > 1$
5	$p_{AB} = 1$	$p'_{AB} > 1$
6	$p_{AB} = 1$	$p'_{AB} < 1$
7	$p_{AB} > 1$	$p'_{AB} > 1$ and $p'_{AB} \neq p_{AB}$
8	$p_{AB} < 1$	$p'_{AB} < 1$ and $p'_{AB} \neq p_{AB}$

Analytic Hierarchy Process

- Pair-wise comparison

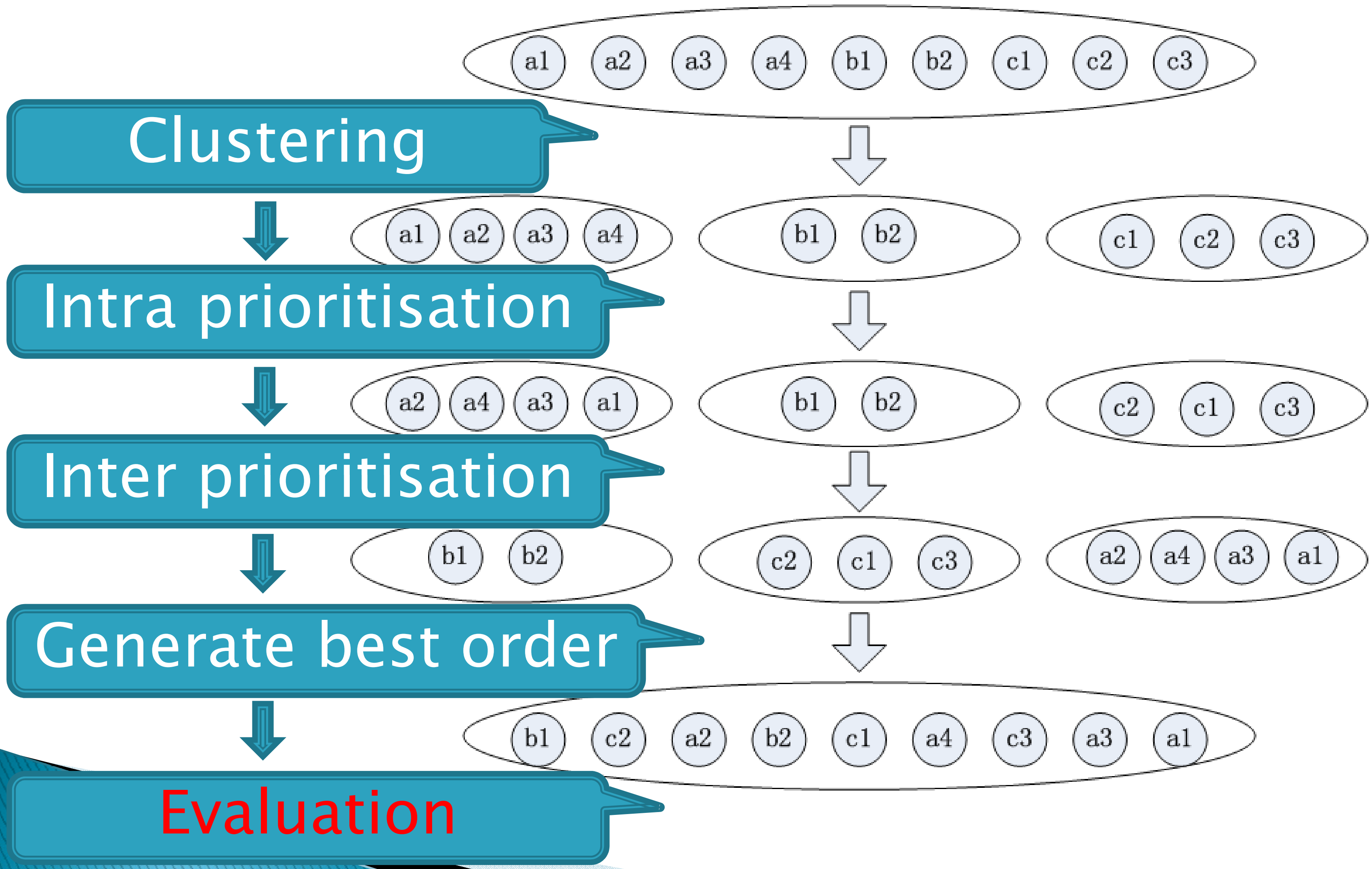
Test Case	Branch 1 (Fault 1)	Branch 2 (Fault 2)	Branch 3 (Fault 3)	Branch 4 (Fault 4)
t_1	x	x	x	
t_2	x	x		
t_3				x

- (t1, t2, t3) or (t1, t3, t2) ?

Analytic Hierarchy Process

- Single criterion hierarchy: ICPs
  - Pair-wise comparison from the human expert
- Multi criteria hierarchy: ICPm
  - Pair-wise comparison
  - Coverage-based prioritisation: scale of 3
  - Preference Value: {9, 7, 5, 3, 1, 1/3, 1/5, 1/7, 1/9}

Framework



a1
a2
a3
a4
b1
b2
c1
c2
c3
Clustering
a1
a2
a3
a4
b1
b2
c1
c2
c3
Intra prioritisation
a2
a4
a3
a1
b1
b2
c2
c1
c3
Inter prioritisation
b1
b2
c2
c1
c3
a2
a4
a3
a1
Generate best order
b1
c2
a2
b2
c1
a4
c3
a3
a1
Evaluation

Evaluation

- Average Percentage of Fault Detection (APFD)

$$APFD = 1 - \frac{TF_1 + \ldots + TF_m}{nm} + \frac{1}{2n}$$

  - T: n test cases; F: m faults
  - T': the ordered T
  - TFi: the order of the first test case reveal the ith fault

Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions

Experimental Setups

- Subjects
  - From Software Infrastructure Repository (SIR)

Program	Test Suite	(Avg.) TS Size	LOC
printtokens	4	317.00	726
schedule	4	225.25	412
space	4	158.50	6,199
gzip	1	212	5,680
sed	1	370	14,427
vim	1	975	122,169
bash	1	1061	59,846

Results & Analysis

- RQ1: Effectiveness: ICP V.S. OP, SC

Subject	schedule			
Test Suite	1	2	3	4
OP	0.991	0.995	0.993	0.993
ICP_S	0.824	0.917	0.952	0.913
SC	0.806	0.865	0.782	0.844

  - OP: Optimal Ordering
  - SC: Statement Coverage
  - ICPs: ICP with single crit

OP > ICPs > SC

Results & Analysis

- RQ1: Effectiveness: ICP V.S. OP, SC

Subject		schedule			
Test Suite		1	2	3	4
OP		0.991	0.995	0.993	0.[illegible]
	9	0.825	0.916	0.954	0.912
	7	0.825	0.916	0.954	0.912
	5	0.825	0.915	0.954	0.912
ICP_M	3	0.825	0.914	0.952	0.912
$p_{[H][C]}$	1	0.824	0.915	0.951	0.912
	1/3	0.823	0.905	0.945	0.909
	1/5	0.820	0.903	0.943	0.907
	1/7	0.821	0.901	0.941	0.906
	1/9	0.821	0.901	0.941	0.904
SC		0.806	0.865	0.782	0.844

OP > ICPm > SC

- OP: Optimal Ordering
- SC: Statement Coverage-based ordering
- ICPm: ICP with multi criteria

Results & Analysis

- RQ2: Configuration: human V.S. coverage

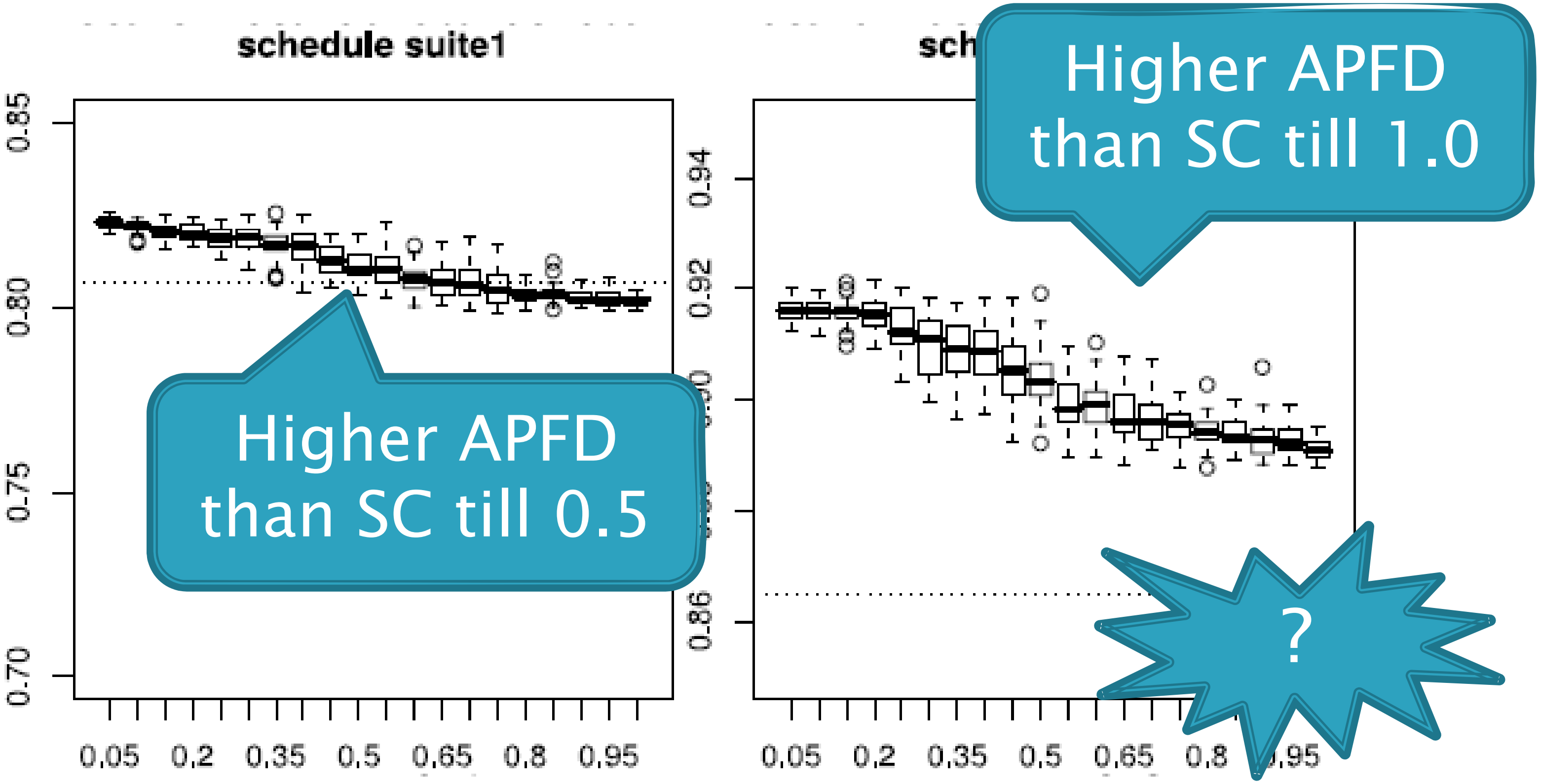
Preference value = 9

Subject		schedule			
Test Suite		1	2	3	
OP		0.991	0.995	0.993	0.993
	9	0.825	0.916	0.954	0.912
	7	0.825	0.916	0.954	0.912
	5	0.825	0.915	0.954	0.912
ICP_M	3	0.825	0.914	0.952	0.912
$p_{[H][C]}$	1	0.824	0.915	0.951	0.912
	1/3	0.823	0.905	0.945	0.909
	1/5	0.820	0.903	0.943	0.907
	1/7	0.821	0.901	0.941	0.906
	1/9	0.821	0.901	0.941	0.904
SC		0.806	0.865	0.782	0.844

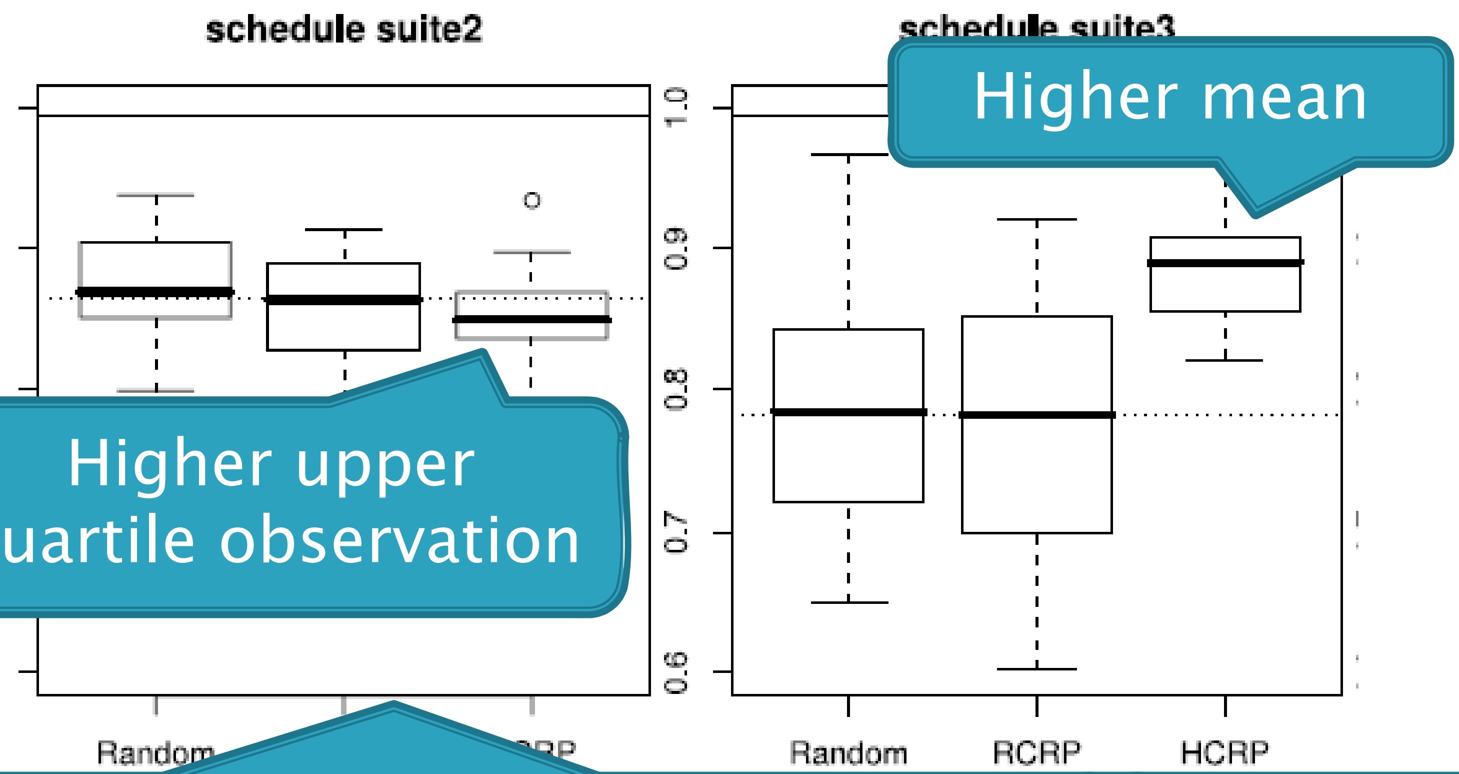
- OP: Optimal Ordering
- SC: Statement Coverage-based ordering
- ICPm: ICP with multi criteria

Results & Analysis

- RQ3: Tolerance: highest tolerated error rate



Results & Analysis



Clustering with14 clusters works

Any prioritisation better than random → improvement

- HCRP: hierarchical clustering random prioritisation

Suitability Test

- Suitability Test – Automated ICP
  - Fault set: AR (Already Revealed)
    TBR (To Be Revealed)
  - Intra & Inter cluster prioritisation on AR set
    - structural coverage
    - Fault information in AR
  - Result > = traditional way
  - Pair-wise comparison will do better on TBR

Suitability Test

- Suitability Test configuration

Program	Size of AR	Size of TBR	Mult. Ver.
printtokens	3	4	No
schedule	4	5	No
space	18	20	No
gzip	2	3	Yes
sed	6	4	Yes
vim	4	3	Yes
bash	4	9	Yes

Suitability Test

- RQ4: Suitability: how accurately does the automated suitability test predict the successful result of ICP?

Subject	schedule			
Test Suite	1	2	3	4
OP	0.991	0.995	0.993	0.993
NCSP AR	0.899	0.974	0.922	0.949
HCSP AR	0.984	0.970*	0.972	0.986
NCSP TBR	0.831	0.880	0.854	0.883
ICP_M TBR	0.994	0.992	0.992	0.992

  - OP: Optimal Ordering
  - NCSP: No clustering/Statement Prioritisation
  - HCSP: Hierarchy clustering with Statement Prioritisation
  - ICPm: ICP with multi criteria

Experiment summary

- Effectiveness
- Configuration
- Tolerance
- Suitability

Successful

Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions

Related Work

- Other prioritisation techniques -- Rothermel
  - Branch-total/additional, Statement-total/additional Fault-Exposing Potential-total/additional
  - No single dominating criterion
- Other prioritisation + clustering usage -- Leon
  - Prioritizing by clustering execution profile
  - Better than coverage-based
- Other AHP applications – human preference
  - Karlsson: requirement prioritisation
  - Finnie: project management
  - Douligeris: Quality of Service
  - Tonella: Case-Base Ranking in test case prioritisation

Conclusion

- Contributions
  - A novel use of clustering
  - A novel AHP-based prioritisation technique
  - A more realistic user model by an error model
  - An automated process of verifying effectiveness
- Future work
  - Different clustering criteria

Thanks for your attention!

Questions?