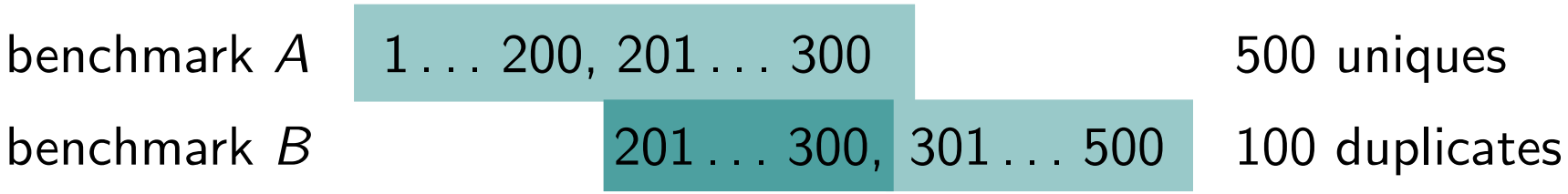
Better Evaluations by Analyzing Benchmark Structure

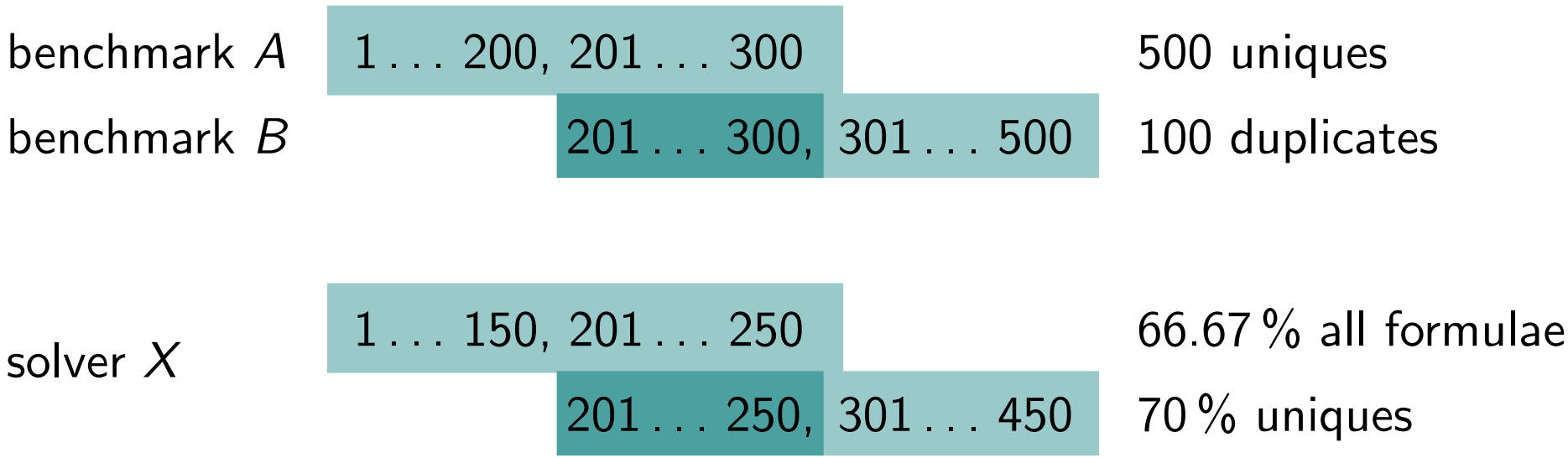
Norbert Manthey and Sibylle Möhle

Knowledge Representation and Reasoning Group
Technische Universität Dresden

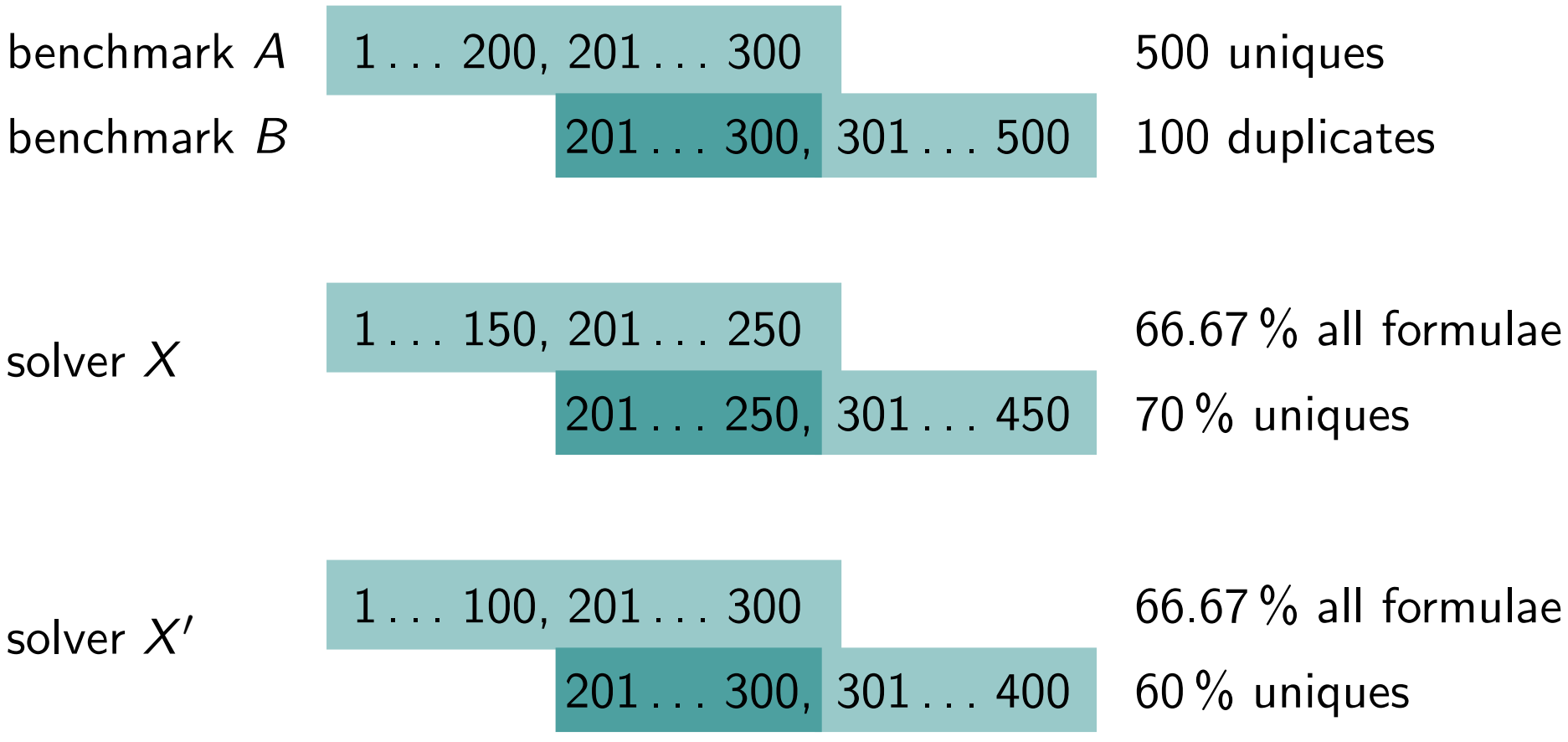
Redundancy Reduces the Significance of Evaluation Results



Redundancy Reduces the Significance of Evaluation Results



Redundancy Reduces the Significance of Evaluation Results



Outline

Redundancy in Benchmarks

Speeding up Solver Evaluation

Experimental Results

Conclusion and Future Work

Redundancy in Benchmarks

Feature $f : CNF \to \Sigma^*$

Feature equivalence $F \equiv_f G$ iff $f(F) = f(G)$

Redundancy-free extract Maximum subset B' of B obtained by iteratively removing G from B if $G \equiv_f F$ for any $F \in B$

Redundancy $1 - \frac{|B'|}{|B|}$

Analyzing Application Benchmark Structure

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	$\sum$	N	B′
02	829														829	100 %	827
03	0	119													119	100 %	119
04	18	8	298												324	91.97 %	315
05	0	0	0	176											176	100 %	175
06	5	5	14	5	71										100	71 %	100
07	0	0	0	15	20	140									175	80 %	175
08	7	7	30	10	26	11	109								200	54.5 %	187
09	4	11	12	9	28	54	13	161							292	55.1 %	292
10	0	1	0	3	7	17	6	23	43						100	43 %	100
11	12	16	7	10	7	19	11	28	11	179					300	59.67 %	299
12	28	28	35	19	31	69	28	82	9	89	182				600	30.33 %	558
13	0	9	0	8	10	18	3	33	3	32	26	158			300	52.67 %	300
14	1	14	0	4	3	14	9	38	4	29	20	12	152		300	50.67 %	299
15	1	2	0	3	0	9	5	19	1	13	8	3	69	167	300	55.67 %	291
R	76	101	98	86	132	211	75	223	28	163	54	15	69				

Structure of SAT Competition Application Benchmarks

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	$\sum$	N	B′
02	829														829	100 %	827
03	0	119													119	100 %	119
04	*18*	*8*	*298*												*324*	*91.97 %*	*315*
05	0	0	*0*	176											176	100 %	175
06	5	5	*14*	5	71										100	71 %	100
07	0	0	*0*	15	20	140									175	80 %	175
08	7	7	*30*	10	26	11	109								200	54.5 %	187
09	4	11	*12*	9	28	54	13	161							292	55.1 %	292
10	0	1	*0*	3	7	17	6	23	43						100	43 %	100
11	12	16	*7*	10	7	19	11	28	11	179					300	59.67 %	299
12	28	28	*35*	19	31	69	28	82	9	89	182				600	30.33 %	558
13	0	9	*0*	8	10	18	3	33	3	32	26	158			300	52.67 %	300
14	1	14	*0*	4	3	14	9	38	4	29	20	12	152		300	50.67 %	299
15	1	2	*0*	3	0	9	5	19	1	13	8	3	69	167	300	55.67 %	291
R	76	101	*98*	86	132	211	75	223	28	163	54	15	69				

Structure of SAT Competition Application Benchmarks

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	$\sum$	N	B′
02	829														829	100 %	827
03	0	119													119	100 %	119
04	18	8	298												324	91.97 %	315
05	0	0	0	176											176	100 %	175
06	5	5	14	5	71										100	71 %	100
07	0	0	0	15	20	140									175	80 %	175
08	7	7	30	10	26	11	109								200	*54.5 %*	187
09	4	11	12	9	28	54	13	161							292	*55.1 %*	292
10	0	1	0	3	7	17	6	23	43						100	*43 %*	100
11	12	16	7	10	7	19	11	28	11	179					300	*59.67 %*	299
12	28	28	35	19	31	69	28	82	9	89	182				600	*30.33 %*	558
13	0	9	0	8	10	18	3	33	3	32	26	158			300	*52.67 %*	300
14	1	14	0	4	3	14	9	38	4	29	20	12	152		300	*50.67 %*	299
15	1	2	0	3	0	9	5	19	1	13	8	3	69	167	300	*55.67 %*	291
R	76	101	98	86	132	211	75	223	28	163	54	15	69				

Structure of SAT Competition Application Benchmarks

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	$\sum$	N	B′
02	829														*829*	100 %	*827*
03	0	119													*119*	100 %	*119*
04	18	8	298												*324*	91.97 %	*315*
05	0	0	0	176											*176*	100 %	*175*
06	5	5	14	5	71										*100*	71 %	*100*
07	0	0	0	15	20	140									*175*	80 %	*175*
08	7	7	30	10	26	11	109								*200*	54.5 %	*187*
09	4	11	12	9	28	54	13	161							*292*	55.1 %	*292*
10	0	1	0	3	7	17	6	23	43						*100*	43 %	*100*
11	12	16	7	10	7	19	11	28	11	179					*300*	59.67 %	*299*
12	28	28	35	19	31	69	28	82	9	89	182				*600*	30.33 %	*558*
13	0	9	0	8	10	18	3	33	3	32	26	158			*300*	52.67 %	*300*
14	1	14	0	4	3	14	9	38	4	29	20	12	152		*300*	50.67 %	*299*
15	1	2	0	3	0	9	5	19	1	13	8	3	69	167	*300*	55.67 %	*291*
R	76	101	98	86	132	211	75	223	28	163	54	15	69				

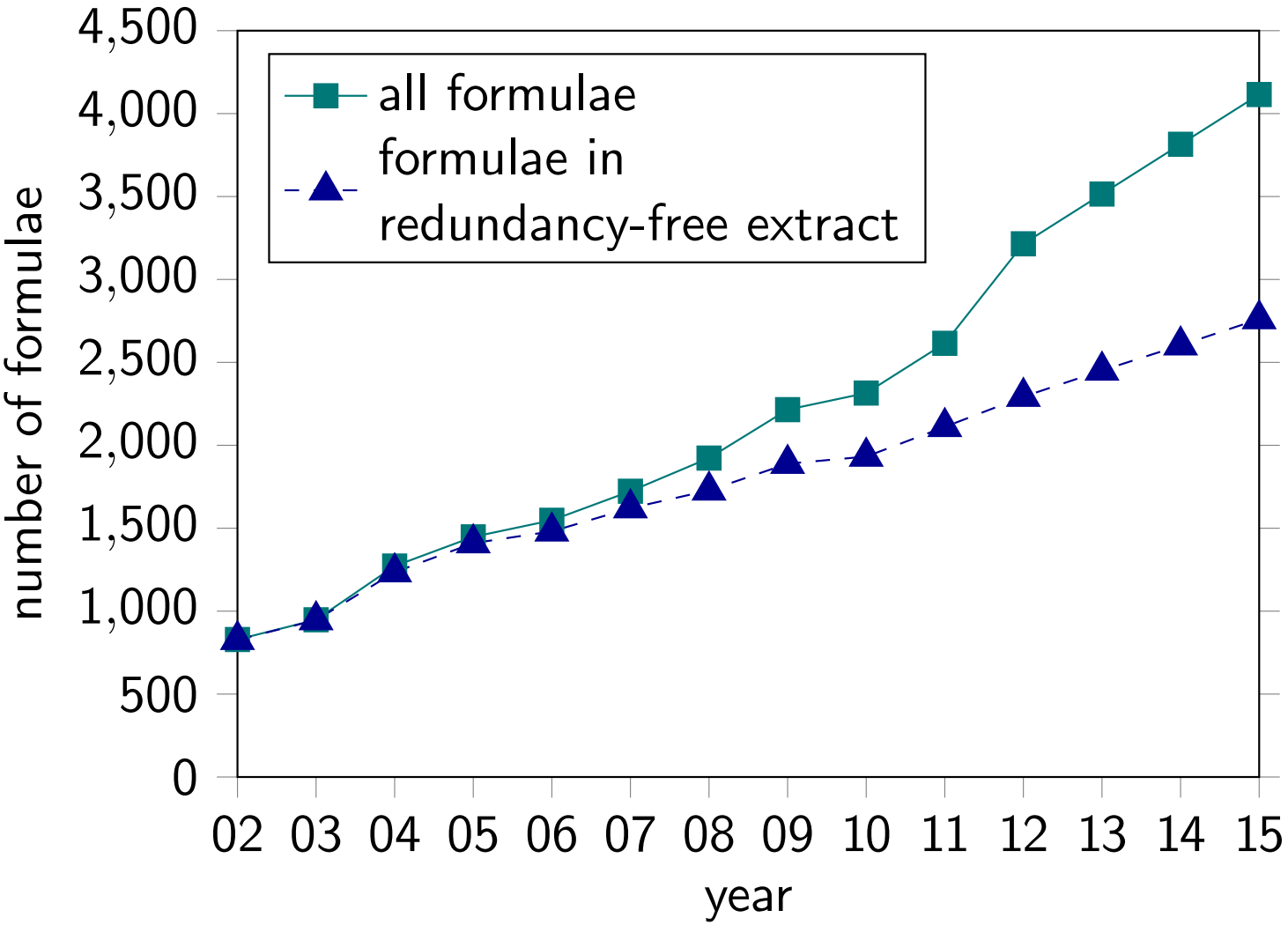
Redundancy in the Combination of two Benchmarks

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**
03	0.21												
04	2.52	3.84											
05	0.30	0.34	2										
06	0.75	2.28	6.84	2.17									
07	0.20	0	1.80	4.56	9.09								
08	2.04	5.96	10.88	5.85	16.33	10.40							
09	0.54	2.68	5.36	2.14	10.97	16.49	11.59						
10	0.22	0.46	2.36	1.45	4.50	9.45	9	9.69					
11	1.33	4.06	3.53	2.52	3	5.26	6.2	9.80	6				
12	4.48	9.04	10.70	7.22	13.43	17.03	16.25	22.65	12.71	22.56			
13	0.18	2.15	1.76	1.89	3.50	5.68	5.40	8.78	3.75	10	18.89		
14	0.35	3.58	2.56	1.26	2	4	5.80	10.64	4.75	9.67	18.22	15.33	
15	1.06	2.63	3.21	2.73	2.25	4.42	5.80	6.59	4.50	6.33	12	8.17	23.83

Redundancy in the Combination of Two Benchmarks

	02	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**
03	0.21												
04	2.52	3.84											
05	0.30	0.34	2										
06	0.75	2.28	6.84	2.17									
07	0.20	*0*	1.80	4.56	9.09								
08	2.04	5.96	10.88	5.85	16.33	10.40							
09	0.54	2.68	5.36	2.14	10.97	16.49	11.59						
10	0.22	0.46	2.36	1.45	4.50	9.45	9	9.69					
11	1.33	4.06	3.53	2.52	3	5.26	6.2	9.80	6				
12	4.48	9.04	10.70	7.22	13.43	17.03	16.25	22.65	12.71	22.56			
13	0.18	2.15	1.76	1.89	3.50	5.68	5.40	8.78	3.75	10	18.89		
14	0.35	3.58	2.56	1.26	2	4	5.80	10.64	4.75	9.67	18.22	15.33	
15	1.06	2.63	3.21	2.73	2.25	4.42	5.80	6.59	4.50	6.33	12	8.17	*23.83*

Combining Benchmarks for Intervals



Speeding up Solver Evaluation on Combined Benchmarks

1. perform evaluation on a redundancy-free extract of the combined benchmarks
2. duplicate results for feature-equivalent formulae

Experiment – Setup

Solvers

- Glucose 3.0 (2015)
- Lingeling baq (2015)
- Riss 6 (2016)

Hardware

- cluster, 2 Intel Xeon CPU E-2680 v3 with 12 cores per node, 2.50 GHz
- timeout 1 h
- memory limit 6.50 GB

Benchmarks

- *application* track of all benchmarks from 2002 to 2015
  (4115 formulae, 2761 redundancy-free formulae, redundancy 32.9 %)

Experiment – Saving in Penalized Average Run Time

solver	redundancy-free extract [h]	all formulae [h]	Δ [%]
Glucose 3.0	769.87	1106.43	30.42
Lingeling baq	692.83	1006.77	31.18
Riss 6	756.10	1105.59	31.61

Conclusion

- When benchmarks are combined, a bias in the evaluation result may be introduced.
- In the presence of redundancy in combined benchmarks, experiments should be carried out on a redundancy-free extract.
- Evaluation on a redundancy-free extract of a combination of benchmarks saves run time.

Future Work

- investigate the suitability of different features for identifying redundancy in benchmarks
- adapt the presented method to other problems, e.g., QBF, MaxSAT, PB, CSP, and AIG

Better Evaluations by Analyzing Benchmark Structure

Thank you for your attention

Experiment – Results

experiment	solver	usc	# solved	time [s]	PAR [h]
redundancy-free extract	Glucose 3.0	14	2117	453120	769.87
	Lingeling baq	88	2213	521379	692.83
	Riss 6	44	2155	540352	756.10
results mapped on full benchmark	Glucose 3.0	31	3261	899943	1103.98
	Lingeling baq	152	3393	1029790	1008.05
	Riss 6	69	3293	1024330	1106.54
evaluation on full benchmark	Glucose 3.0	31	3256	890734	1106.43
	Lingeling baq	151	3388	1007180	1006.77
	Riss 6	68	3297	1035320	1105.59