Table 1. Results of the number of winning benchmarks, where *LS-Sampling-Plus* achieves significantly higher t-wise coverage than each of *Baital*, *NS*, *PLEDGE* and *LS-Sampling* (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over 123 public benchmarks.

	LS-Sampling-Plus vs. *Uniform*	*LS-Sampling-Plus* vs. *Baital*	*LS-Sampling-Plus* vs. *NS*	*LS-Sampling-Plus* vs. *PLEDGE*	*LS-Sampling-Plus* vs. *LS-Sampling*
	#win	#win	#win	#win	#win
2-wise (k=50)	122	123	122	122	122
2-wise (k=100)	122	122	122	122	122
2-wise (k=500)	122	122	122	121	121
3-wise (k=50)	123	123	123	123	123
3-wise (k=100)	123	123	123	123	123
3-wise (k=500)	122	122	122	122	122
4-wise (k=50)	123	123	122	123	123
4-wise (k=100)	123	123	122	123	123
4-wise (k=500)	122	122	121	122	122
5-wise (k=50)	123	123	123	123	123
5-wise (k=100)	123	123	122	123	123
5-wise (k=500)	122	122	121	122	122
6-wise (k=50)	123	123	123	123	123
6-wise (k=100)	123	123	123	123	123
6-wise (k=500)	123	122	122	123	122

Table 2. Results of the number of winning benchmarks, where *LS-Sampling-Plus* with λ = 1000 achieves significantly higher t-wise coverage than each of *LS-Sampling-Plus* with λ = 10, 50, 100 and 500 (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over 123 public benchmarks.

	λ = 1000 vs. λ=10	λ = 1000 vs. λ=50	λ = 1000 vs. λ=100	λ = 1000 vs. λ=500
	#win	#win	#win	#win
2-wise (k=50)	121	121	120	119
2-wise (k=100)	120	84	26	10
2-wise (k=500)	18	16	11	8
3-wise (k=50)	122	122	122	121
3-wise (k=100)	122	121	122	119
3-wise (k=500)	27	108	81	16
4-wise (k=50)	123	123	123	121
4-wise (k=100)	122	122	122	120
4-wise (k=500)	120	119	119	116
5-wise (k=50)	123	123	123	121
5-wise (k=100)	123	122	122	122
5-wise (k=500)	122	122	122	120
6-wise (k=50)	123	123	123	122
6-wise (k=100)	123	123	123	121
6-wise (k=500)	122	122	121	121

Table 3. Results of the number of winning benchmarks where *LS-Sampling-Plus* with δ=2 · 10^6 achieves significantly higher t-wise coverage than each of *LS-Sampling-Plus* with $\delta = 5 \cdot 10^5$, $1 \cdot 10^6$ and $1.5 \cdot 10^6$ (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($3 \leq t \leq 6$) and k (k = 50, 100 and 500) over 123 public benchmarks.

	δ=2 · 10^6 vs. δ=5 · 10^5	δ=2 · 10^6 vs. δ=1 · 10^6	δ=2 · 10^6 vs. δ=1.5 · 10^6
	#win	#win	#win
3-wise (k=50)	46	8	11
3-wise (k=100)	121	93	20
3-wise (k=500)	122	121	120
4-wise (k=50)	32	12	9
4-wise (k=100)	123	41	11
4-wise (k=500)	121	121	121
5-wise (k=50)	49	12	9
5-wise (k=100)	121	44	10
5-wise (k=500)	122	122	120
6-wise (k=50)	96	22	10
6-wise (k=100)	123	58	19
6-wise (k=500)	123	122	120

Table 4. Results of the number of winning benchmarks where *LS-Sampling-Plus* achieves significantly higher t-wise coverage than each of *LS-Sampling-Plus-alt1*, *LS-Sampling-Plus-alt2* and *LS-Sampling-Plus-alt3* (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over 123 public benchmarks.

	LS-Sampling-Plus vs. *LS-Sampling-Plus-alt1*	*LS-Sampling-Plus* vs. *LS-Sampling-Plus-alt2*	*LS-Sampling-Plus* vs. *LS-Sampling-Plus-alt3*
	#win	#win	#win
2-wise (k=50)	119	122	121
2-wise (k=100)	120	121	121
2-wise (k=500)	120	121	119
3-wise (k=50)	122	123	123
3-wise (k=100)	123	123	123
3-wise (k=500)	123	123	123
4-wise (k=50)	123	123	123
4-wise (k=100)	123	123	123
4-wise (k=500)	121	122	122
5-wise (k=50)	123	123	123
5-wise (k=100)	123	123	123
5-wise (k=500)	122	122	123
6-wise (k=50)	123	123	123
6-wise (k=100)	123	123	123
6-wise (k=500)	123	122	123

Table 5. Results of the number of winning benchmarks where *LS-Sampling-Plus* achieves significantly higher t-option fault detection rate than each of *NS*, *PLEDGE* and *LS-Sampling* (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over the entire benchmark collection of 31 public benchmarks for assessing the fault detection capability.

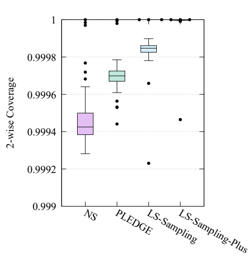
	LS-Sampling-Plus vs. *NS*	*LS-Sampling-Plus* vs. *PLEDGE*	*LS-Sampling-Plus* vs. *LS-Sampling*
	#win	#win	#win
2-option (k=50)	25	25	14
2-option (k=100)	21	22	12
2-option (k=500)	17	16	10
3-option (k=50)	30	30	24
3-option (k=100)	27	27	21
3-option (k=500)	18	19	17
4-option (k=50)	31	31	27
4-option (k=100)	30	29	25
4-option (k=500)	23	21	22
5-option (k=50)	31	30	27
5-option (k=100)	30	30	26
5-option (k=500)	23	23	23
6-option (k=50)	30	30	26
6-option (k=100)	31	31	27
6-option (k=500)	25	26	27

Table 6. Results of the number of winning benchmarks where *LS-Sampling-Plus* achieves significantly higher t-wise coverage than each of *Baital*, *NS*, *PLEDGE* and *LS-Sampling* (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over 5 non-binary benchmarks of `Healthcare4`, `Insurance`, `ProcessorComm2`, `Strorage4` and `Storage5`.

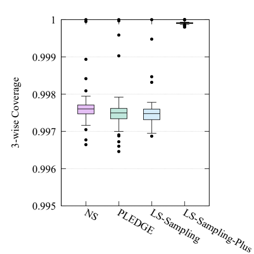
	LS-Sampling-Plus vs. *Uniform*	*LS-Sampling-Plus* vs. *Baital*	*LS-Sampling-Plus* vs. *NS*	*LS-Sampling-Plus* vs. *PLEDGE*	*LS-Sampling-Plus* vs. *LS-Sampling*
	#win	#win	#win	#win	#win
2-wise (k=50)	5	5	5	5	4
2-wise (k=100)	5	5	5	5	2
2-wise (k=500)	5	3	3	3	2
3-wise (k=50)	5	5	5	5	5
3-wise (k=100)	5	5	5	5	4
3-wise (k=500)	5	5	5	5	5
4-wise (k=50)	5	5	5	5	5
4-wise (k=100)	5	5	5	5	4
4-wise (k=500)	5	5	5	5	5
5-wise (k=50)	5	5	5	5	5
5-wise (k=100)	5	5	5	5	4
5-wise (k=500)	5	5	5	5	5
6-wise (k=50)	5	5	5	5	5
6-wise (k=100)	5	5	5	5	5
6-wise (k=500)	5	5	5	5	5

Table 7. Results of the number of winning benchmarks where *LS-Sampling-Plus* achieves significantly higher t-wise coverage than each of *Baital*, *NS*, *PLEDGE* and *LS-Sampling* (*i.e.,* the p-value of Wilcoxon signed-rank test for the related pairwise comparison is smaller than 0.05) under various settings of t ($2 \leq t \leq 6$) and k (k = 50, 100 and 500) over the remaining 15 benchmarks.

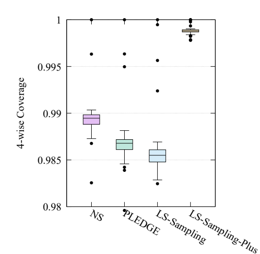
	LS-Sampling-Plus vs. *Uniform*	*LS-Sampling-Plus* vs. *Baital*	*LS-Sampling-Plus* vs. *NS*	*LS-Sampling-Plus* vs. *PLEDGE*	*LS-Sampling-Plus* vs. *LS-Sampling*
	#win	#win	#win	#win	#win
2-wise (k=50)	14	14	14	15	5
2-wise (k=100)	10	9	11	13	3
2-wise (k=500)	5	2	5	5	0
3-wise (k=50)	15	14	15	15	14
3-wise (k=100)	14	14	15	15	14
3-wise (k=500)	8	8	9	10	9
4-wise (k=50)	15	15	15	15	15
4-wise (k=100)	14	14	15	15	15
4-wise (k=500)	12	12	13	14	14
5-wise (k=50)	15	15	15	15	15
5-wise (k=100)	14	14	15	15	15
5-wise (k=500)	13	13	15	15	14
6-wise (k=50)	15	15	15	15	15
6-wise (k=100)	14	14	15	15	15
6-wise (k=500)	13	13	15	15	14



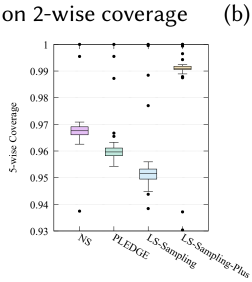
(a) Results on 2-wise coverage



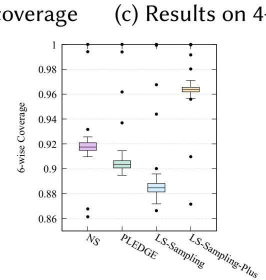
(b) Results on 3-wise coverage



(c) Results on 4-wise coverage

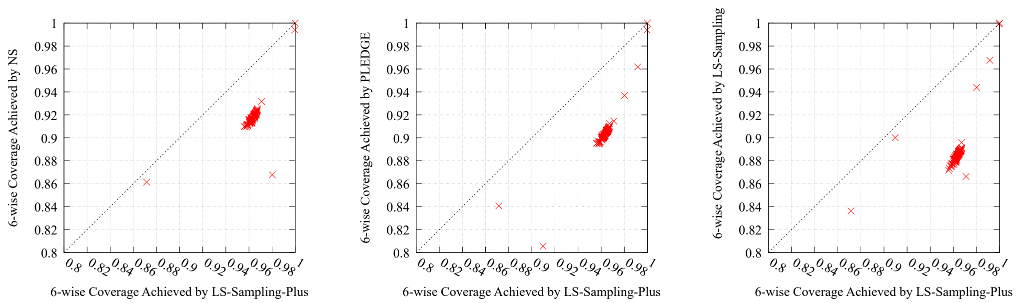


(d) Results on 5-wise coverage



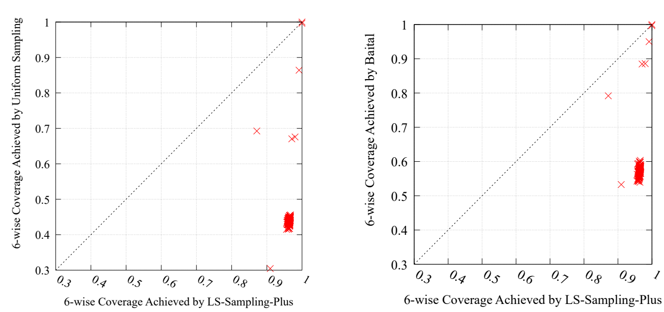
(e) Results on 6-wise coverage

Fig. 1. Box plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus*.



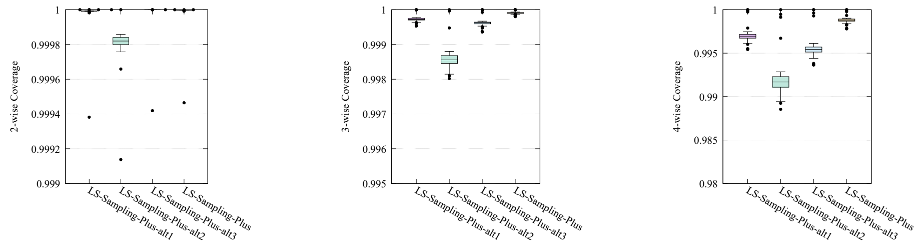
(a) *LS-Sampling-Plus* vs. *NS* (b) *LS-Sampling-Plus* vs. *PLEDGE* (c) *LS-Sampling-Plus* vs. *LS-Sampling*

Fig. 2. Scatter plots demonstrating the 6-wise coverage achieved by *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus*.

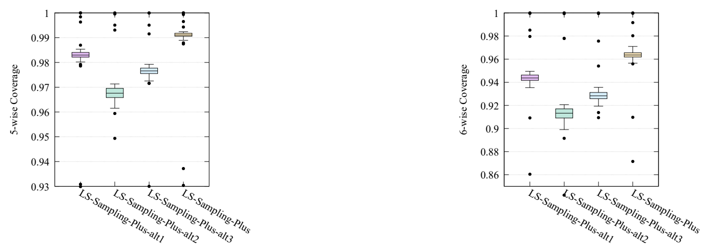


(a) *LS-Sampling-Plus* vs. Uniform Sampling (b) *LS-Sampling-Plus* vs. *Baital*

Fig. 3. Scatter plots demonstrating the 6-wise coverage achieved by Uniform Sampling, *Baital* and *LS-Sampling-Plus*.

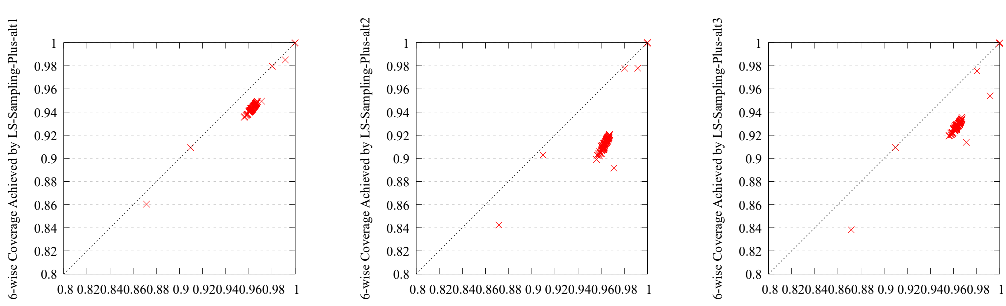


(a) Results on 2-wise coverage (b) Results on 3-wise coverage (c) Results on 4-wise coverage



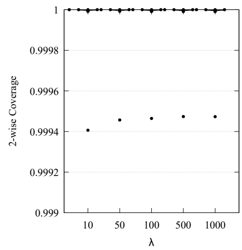
(d) Results on 5-wise coverage (e) Results on 6-wise coverage

Fig. 4. Box plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by *LS-Sampling-Plus-alt1*, *LS-Sampling-Plus-alt2*, *LS-Sampling-Plus-alt3* and *LS-Sampling-Plus*.

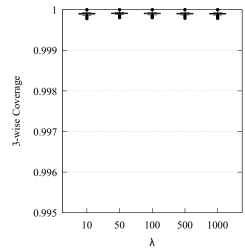


(a) *LS-Sampling-Plus* vs. *LS-Sampling-Plus-alt1* (b) *LS-Sampling-Plus* vs. *LS-Sampling-Plus-alt2* (c) *LS-Sampling-Plus* vs. *LS-Sampling-Plus-alt3*

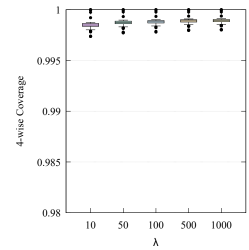
Fig. 5. Scatter plots demonstrating the 6-wise coverage achieved by *LS-Sampling-Plus-alt1*, *LS-Sampling-Plus-alt2*, *LS-Sampling-Plus-alt3* and *LS-Sampling-Plus*.



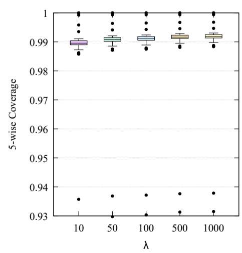
(a) Results on 2-wise coverage



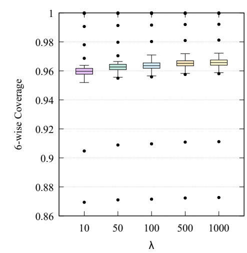
(b) Results on 3-wise coverage



(c) Results on 4-wise coverage

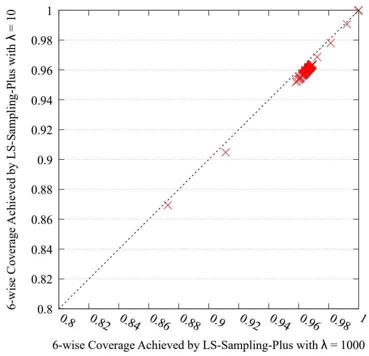


(d) Results on 5-wise coverage

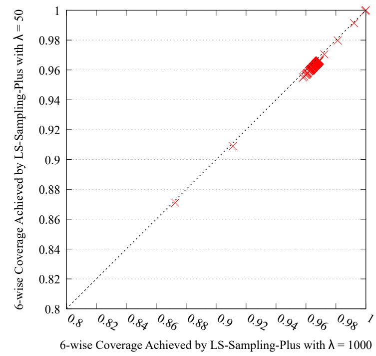


(e) Results on 6-wise coverage

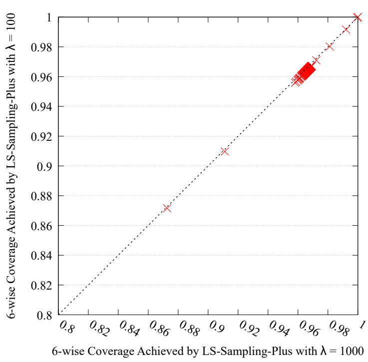
Fig. 6. Box plots demonstrating the t-wise coverage ($2 \le t \le 6$) achieved by *LS-Sampling-Plus* with different hyper-parameter settings of λ.



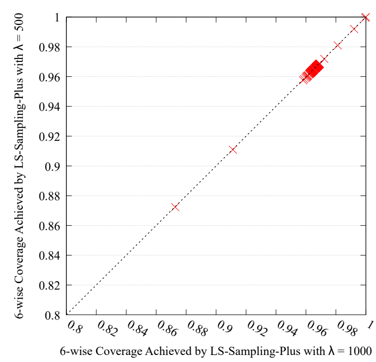
(a) λ = 1000 vs. λ = 10



(b) λ = 1000 vs. λ = 50

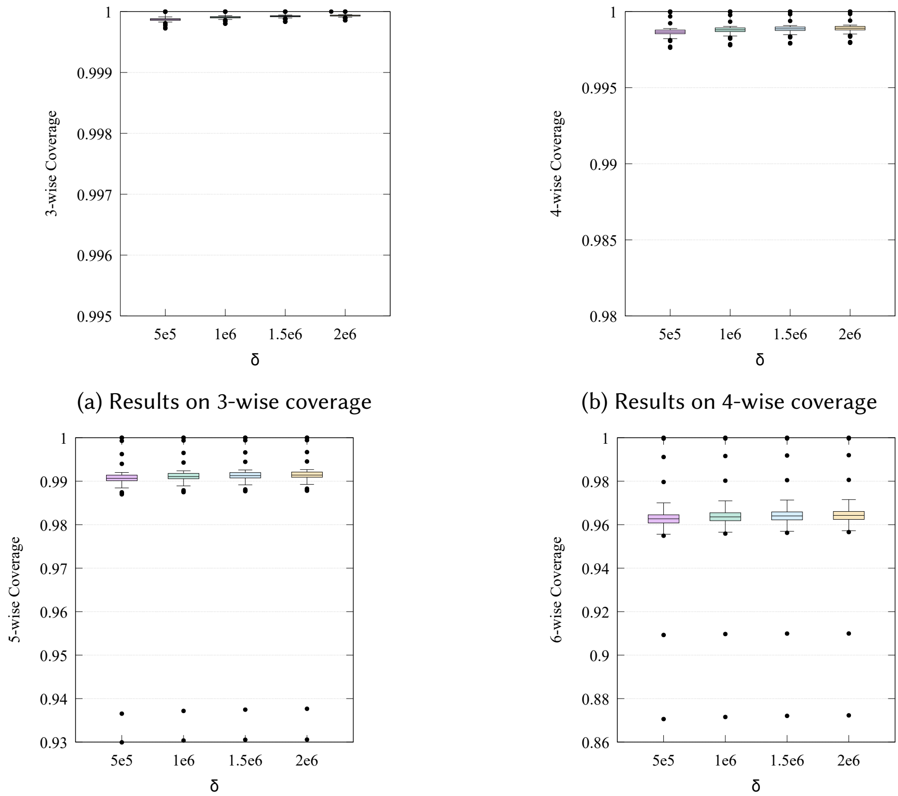


(c) λ = 1000 vs. λ = 10



(d) λ = 1000 vs. λ = 500

Fig. 7. Scatter plots demonstrating the 6-wise coverage achieved *LS-Sampling-Plus* with different hyper-parameter settings of λ.



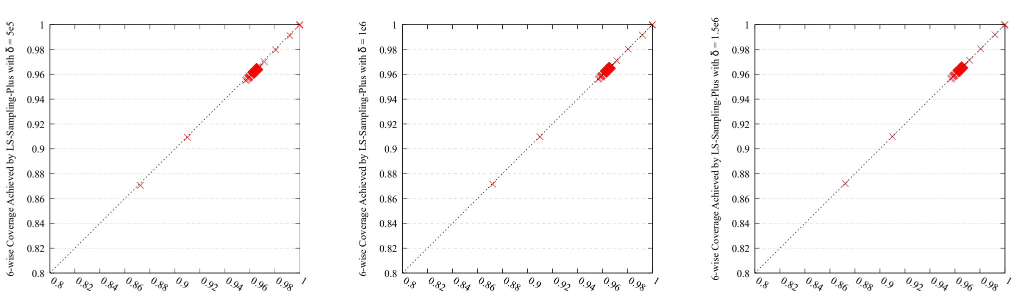
(a) Results on 3-wise coverage

(b) Results on 4-wise coverage

(c) Results on 5-wise coverage

(d) Results on 6-wise coverage

Fig. 8. Box plots demonstrating the t-wise coverage ($3 \leq t \leq 6$) achieved by *LS-Sampling-Plus* with different hyper-parameter settings of δ.

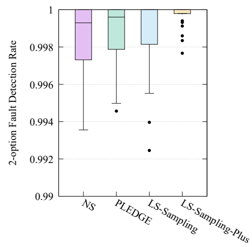


(a) $\delta = 2e6$ vs. $\delta = 5e5$

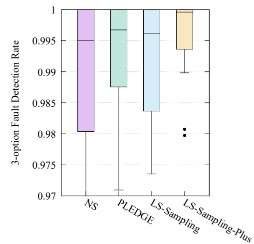
(b) $\delta = 2e6$ vs. $\delta = 1e6$

(c) $\delta = 2e6$ vs. $\delta = 1.5e6$

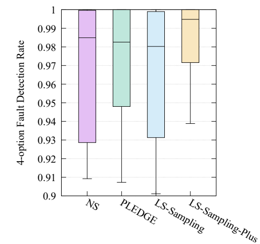
Fig. 9. Scatter plots demonstrating the 6-wise coverage achieved by *LS-Sampling-Plus* with different hyper-parameter settings of δ.



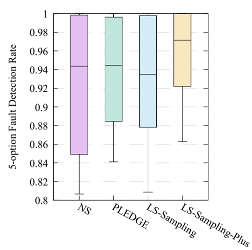
(a) Results on 2-option fault detection rate



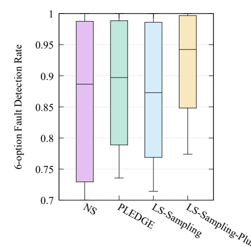
(b) Results on 3-option fault detection rate



(c) Results on 4-option fault detection rate

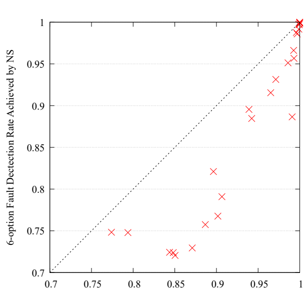


(d) Results on 5-option fault detection rate

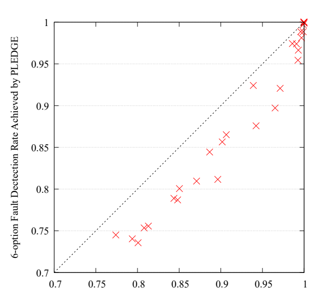


(e) Results on 6-option fault detection rate

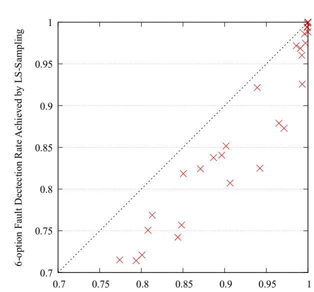
Fig. 10. Box plots demonstrating the t-option fault detection rate ($2 \leq t \leq 6$) achieved by *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus*.



(a) *LS-Sampling-Plus* vs. *NS*

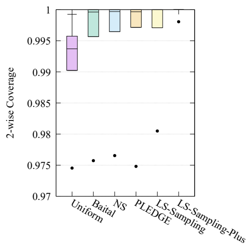


(b) *LS-Sampling-Plus* vs. *PLEDGE*

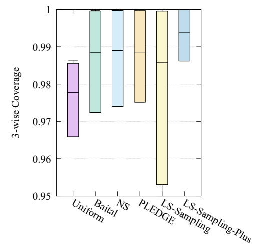


(c) *LS-Sampling-Plus* vs. *LS-Sampling*

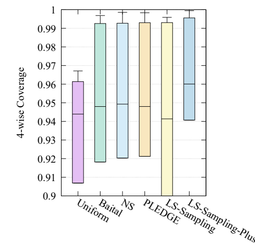
Fig. 11. Scatter plots demonstrating the 6-option fault detection rate achieved by *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus*.



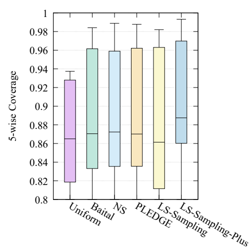
(a) Results on 2-wise coverage



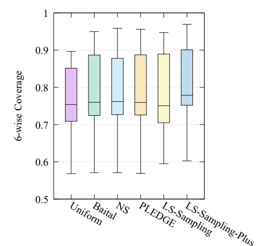
(b) Results on 3-wise coverage



(c) Results on 4-wise coverage

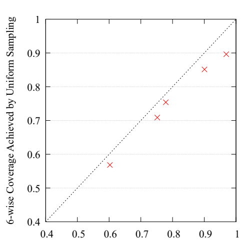


(d) Results on 5-wise coverage

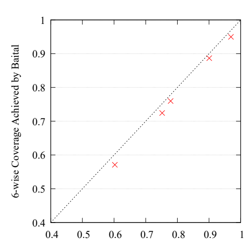


(e) Results on 6-wise coverage

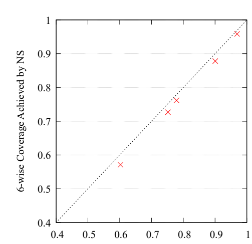
Fig. 12. Box plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by uniform sampling, *Baital*, *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus* over 5 non-binary benchmarks of `Healthcare4`, `Insurance`, `ProcessorComm2`, `Strorage4` and `Storage5`.



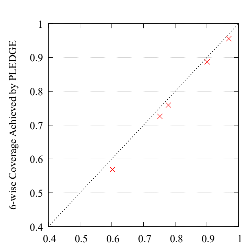
(a) *LS-Sampling-Plus* vs. Uniform Sampling



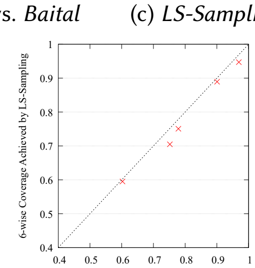
(b) *LS-Sampling-Plus* vs. *Baital*



(c) *LS-Sampling-Plus* vs. *NS*

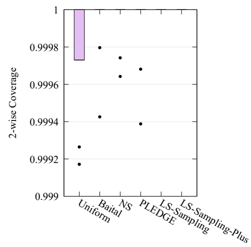


(d) *LS-Sampling-Plus* vs. *PLEDGE*

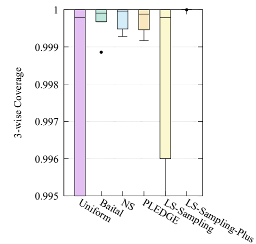


(e) *LS-Sampling-Plus* vs. *LS-Sampling*

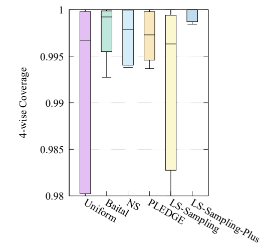
Fig. 13. Scatter plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by uniform sampling, *Baital*, *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus* over 5 non-binary benchmarks of `Healthcare4`, `Insurance`, `ProcessorComm2`, `Strorage4` and `Storage5`.



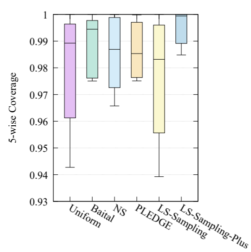
(a) Results on 2-wise coverage



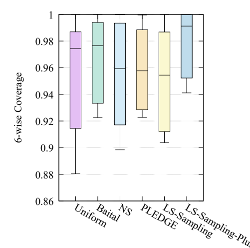
(b) Results on 3-wise coverage



(c) Results on 4-wise coverage

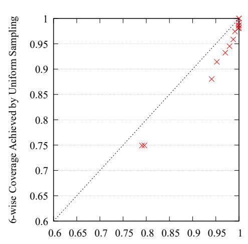


(d) Results on 5-wise coverage

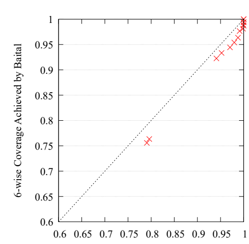


(e) Results on 6-wise coverage

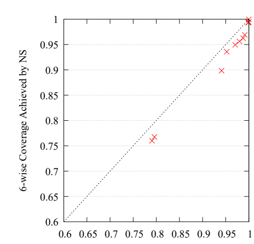
Fig. 14. Box plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by uniform sampling, *Baital*, *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus* over the remaining 15 non-binary benchmarks.



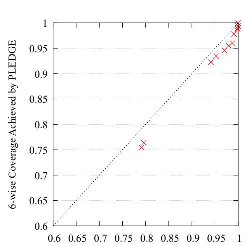
(a) *LS-Sampling-Plus* vs. Uniform Sampling



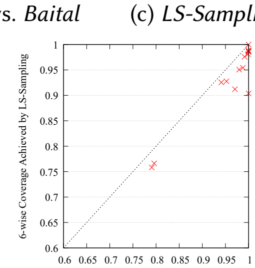
(b) *LS-Sampling-Plus* vs. *Baital*



(c) *LS-Sampling-Plus* vs. *NS*



(d) *LS-Sampling-Plus* vs. *PLEDGE*



(e) *LS-Sampling-Plus* vs. *LS-Sampling*

Fig. 15. Scatter plots demonstrating the t-wise coverage ($2 \leq t \leq 6$) achieved by uniform sampling, *Baital*, *NS*, *PLEDGE*, *LS-Sampling* and *LS-Sampling-Plus* over the remaining 15 non-binary benchmarks.