Technical Report of Evaluation

Queries

Regular Templates

We have the following regular template:

```
Q1: "a*",
Q2: "ab*",
Q3: "a?b*",
Q4: "ab",
Q5: "a*/b*",
Q6: "a*/b*/c",
Q7: "a/b/c",
Q8: "a/b*/c",
Q9: "(a \| b \| c)*",
Q10: "(a \| b \| c)/b*"
```

Data Constraint:

We instantiate five data constraints into the regular expressions

- D1. Assert that the distance between an attribute and a constant should be within a threshold, i.e. exists a parameter p, such that

  ```
  |?p - attr| < constant
  ```

  And the equal data constraint is:

  ```
  ?p - attr < c and attr - ?p < c
  ```

- D2. Query the upper and lower bound of an attribute `attr', i.e.

  ```
  ?p >= attr and ?q <= attr
  ```

- D3. Query a path, such that the upper and lower bound of an attribute should be within a threshold. i.e.

  ```
  ?p >= attr and ?q <= attr and ?p - ?q < c
  ```

- D4. This query is inspired by the dating query in Geri's thesis. Query a path, such that:

  Let the start point satisfy the following condition:

  ```
  |?p - attr1| < c and ?q == attr2
  ```

  And the successor nodes should satisfy the following condition:

  ```
  |?p - attr1| < c and attr2 * 0.5 + 10 <= ?q
  ```

  And we can also expand the absolute value by:

  ```
      attr1 - ?p < c and ?p - attr < c
  ```

- D5. The 2-D manhattan distance between (a_1, b_1) and (a_2, b_2) is defined as following: $$MH = |a_1 - a_2 | + |b_1 - b_2|$$

We want to assert the manattan distance from the start point of a path to all nodes along the path within a threshold. $MH < c$.

We can use two global parameters to encode the mahattan distance between the start point and the successor along a path: For the start point, let ?p and ?q as two parameters, we have

```
  ?p == attr1 and ?q = attr2
```

For the successor nodes along the path, we have

```
?p - attr1 + ?q - attr2 < c and
?p - attr1 + attr2 - ?q < c and
attr1 - ?p + ?q - attr2 < c and
attr1 - ?p + attr2 - ?q < c
```

Experiment Setup

The experiment was run on a ubuntu subsystem on a windows 11 Laptop with i7-13700H Core CPU and 16 GB assigned memories. The following table shows the stat of each dataset.

Dataset	Node Number	Edge Number	Queries for Each Template	Queries in Total
ICIJ-Leak	1908466	3193390	10000	500000
Pokec	1632803	30622564	10000	500000

Dataset	Node Number	Edge Number	Queries for Each Template	Queries in Total
LDBC10	29987850	178101408	10000	500000
ICIJ-Paradise	163414	364456	1000	50000
Telecom	170k	50M	1000	50000
LDBC01	184329	767894	1000	50000

Evaluation Result:

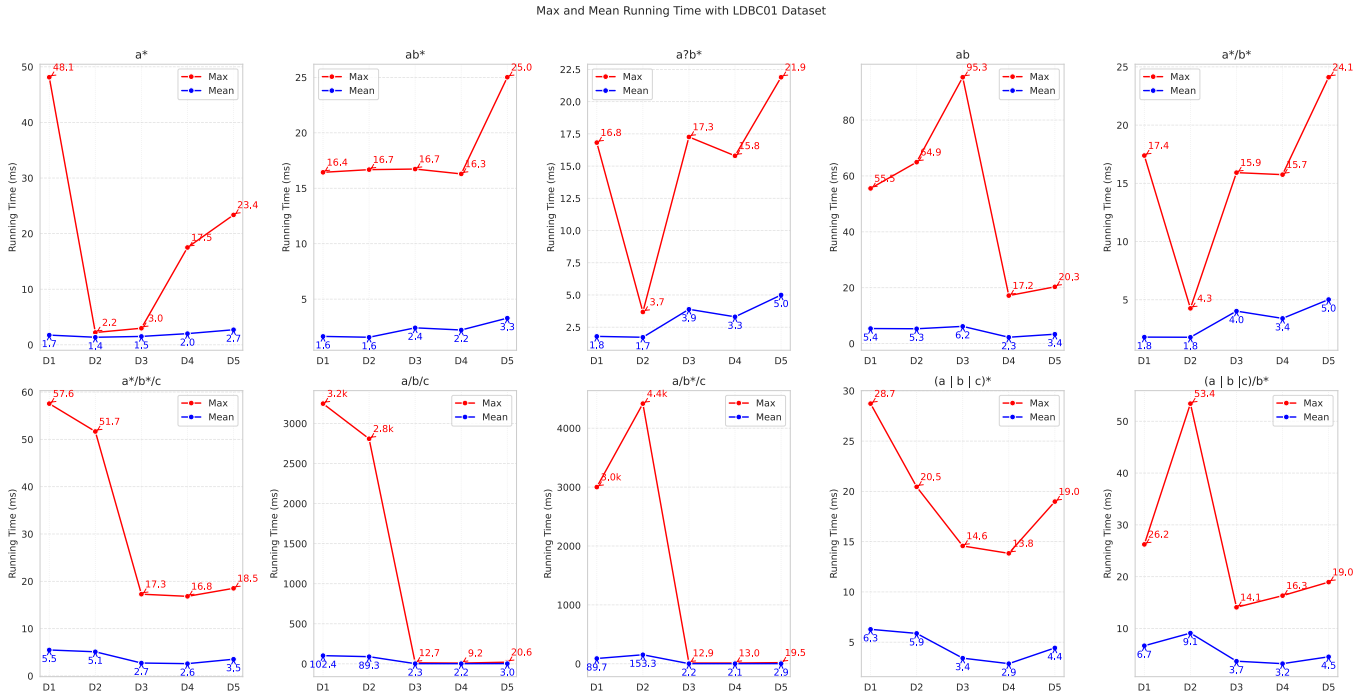
LDBC01 Dataset

Time Performace

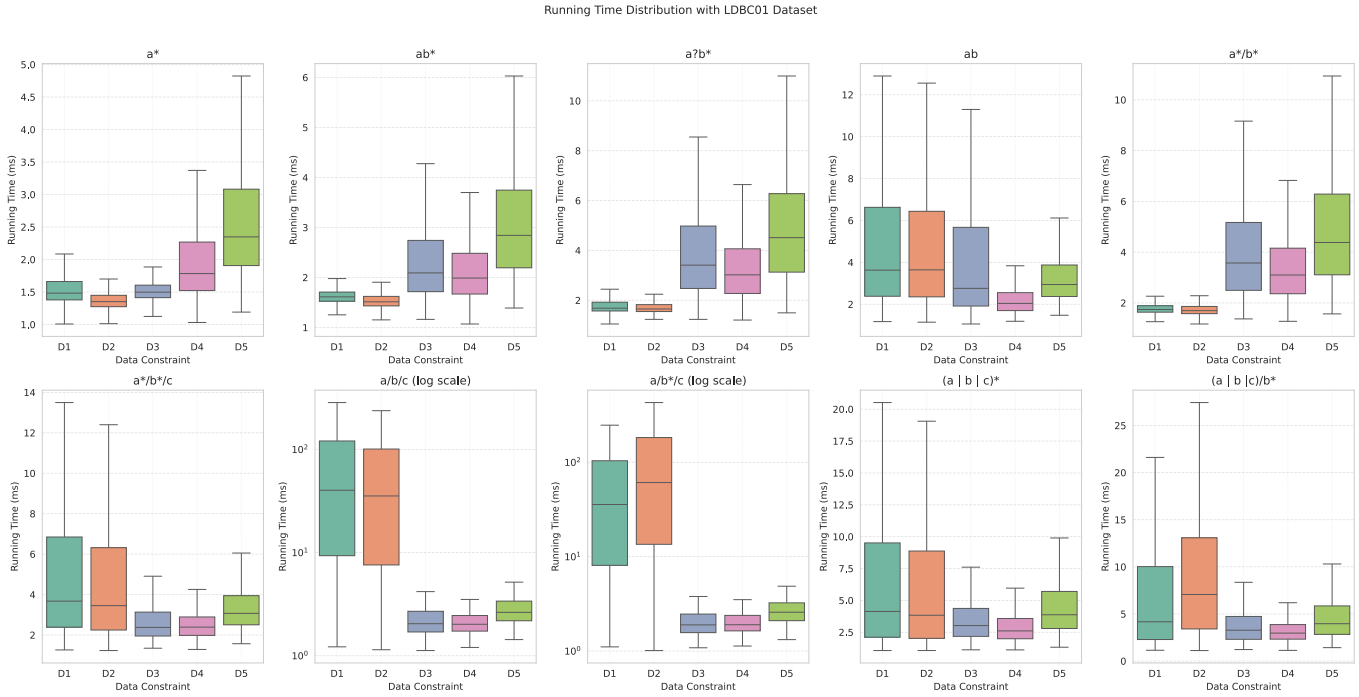
The following table shows statistics of regular path querie for each regular template.

Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.32	10.6
ab*	0.32	15.63
a?b*	0.35	0.9
ab	0.41	0.87
a*/b*	0.36	0.84
a*/b*/c	0.32	21.99
a/b/c	0.32	0.72
a/b*/c	0.25	0.72
(a | b | c)*	0.26	0.63
(a | b |c)/b*	0.26	0.57

The following figure shows the maximal and mean running time of each RDPQ query.

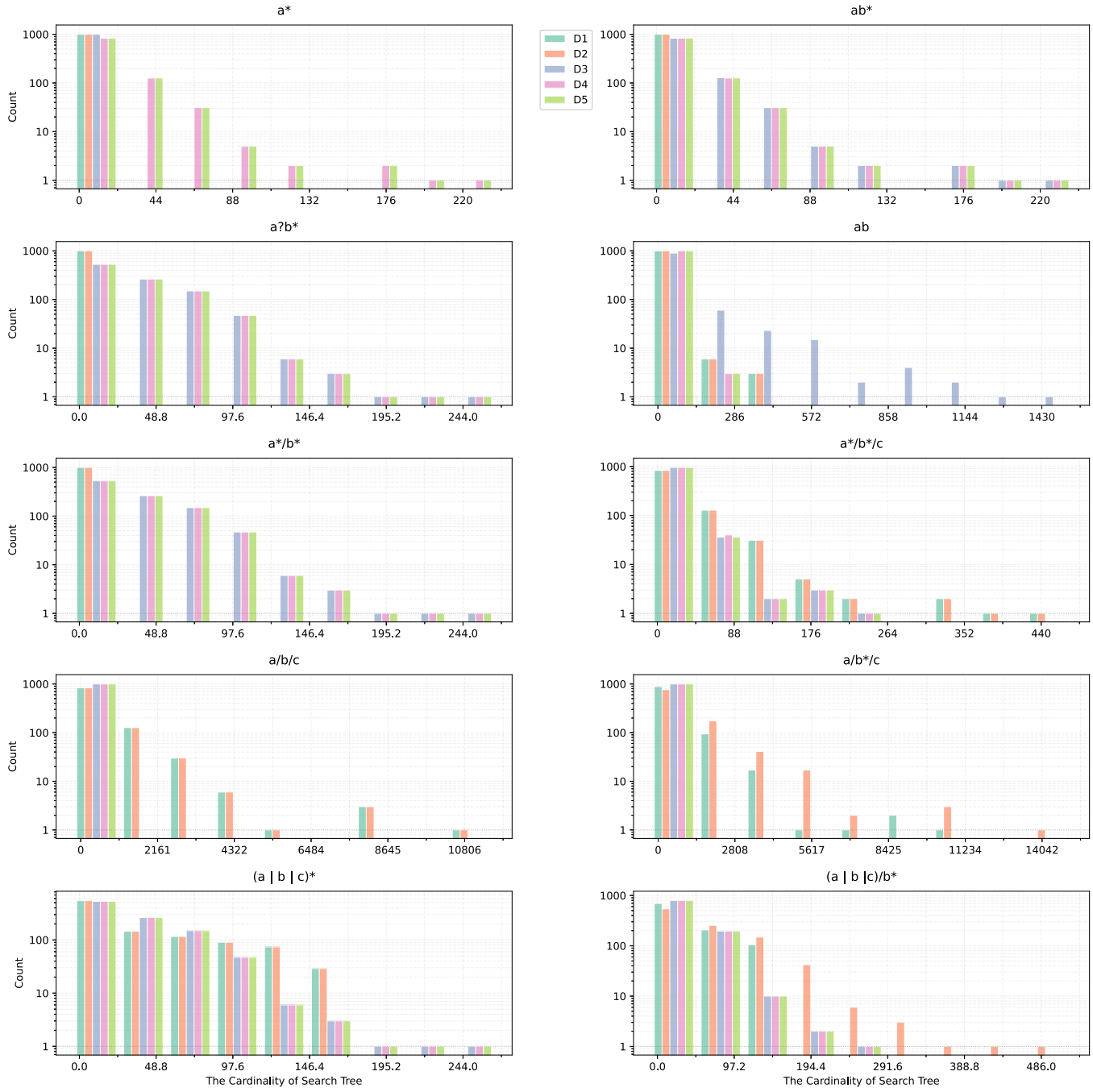


The following box figure show the distribution of running time on LDBC01 dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.



ICIJ-Paradise Dataset

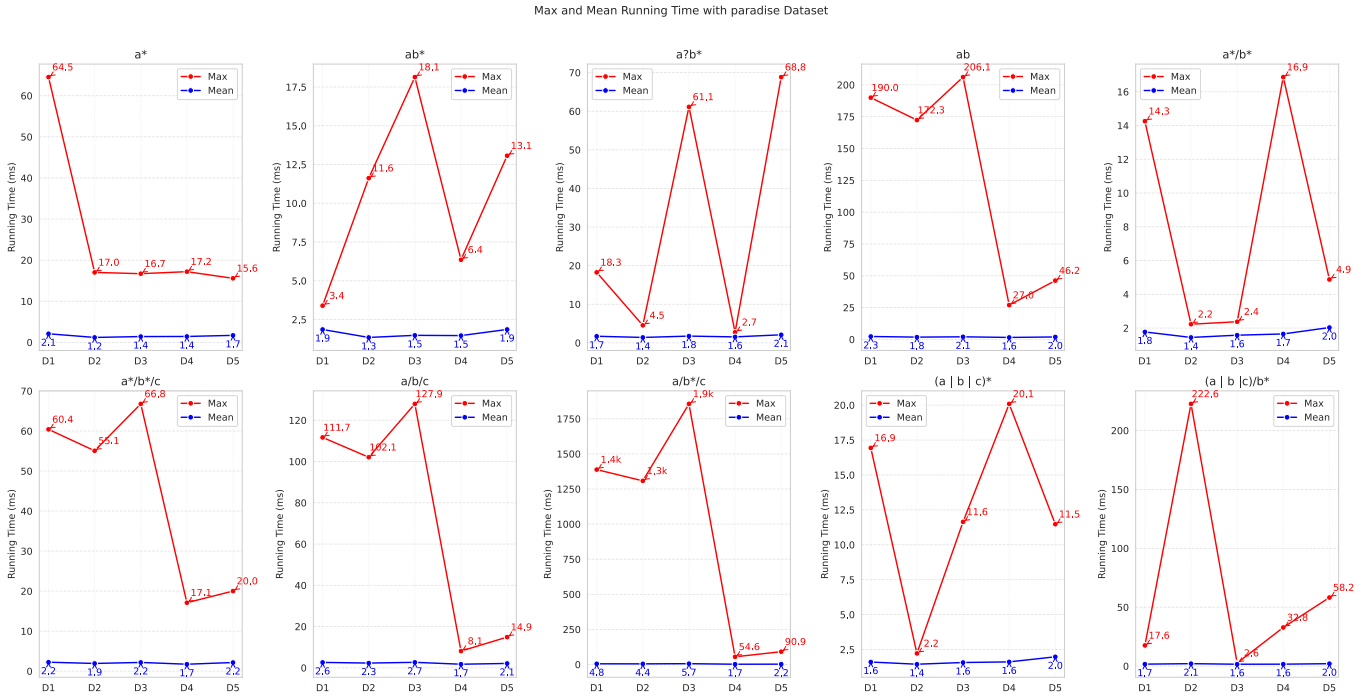
Time Performance

The following table shows statistics of regular path querie for each regular template on icij-paradise dataset.

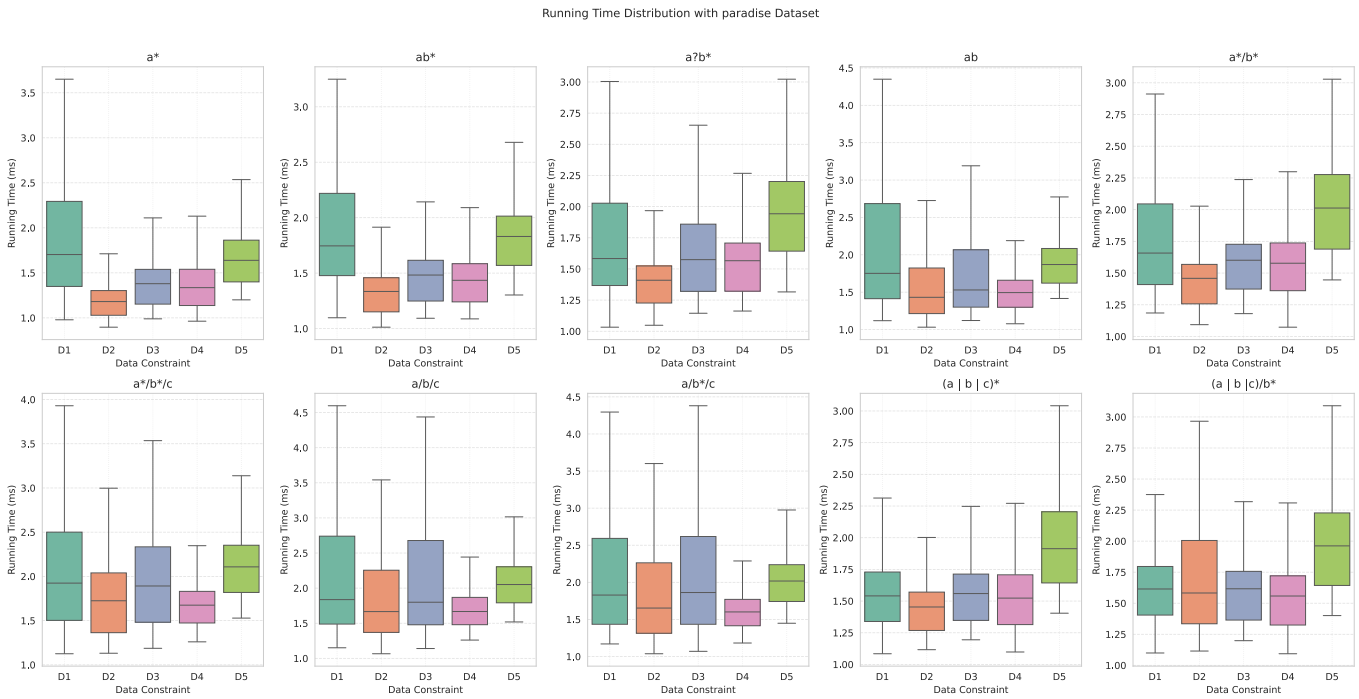
Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.31	15.57
ab*	0.32	1.26
a?b*	0.32	0.92
ab	0.32	1.27

Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*/b*	0.33	0.6
a*/b*/c	0.28	21.14
a/b/c	0.27	0.61
a/b*/c	0.28	0.56
(a | b | c)*	0.28	0.55
(a | b | c)/b*	0.27	0.77

The following figure shows the maximal and mean running time of each RDPQ query.

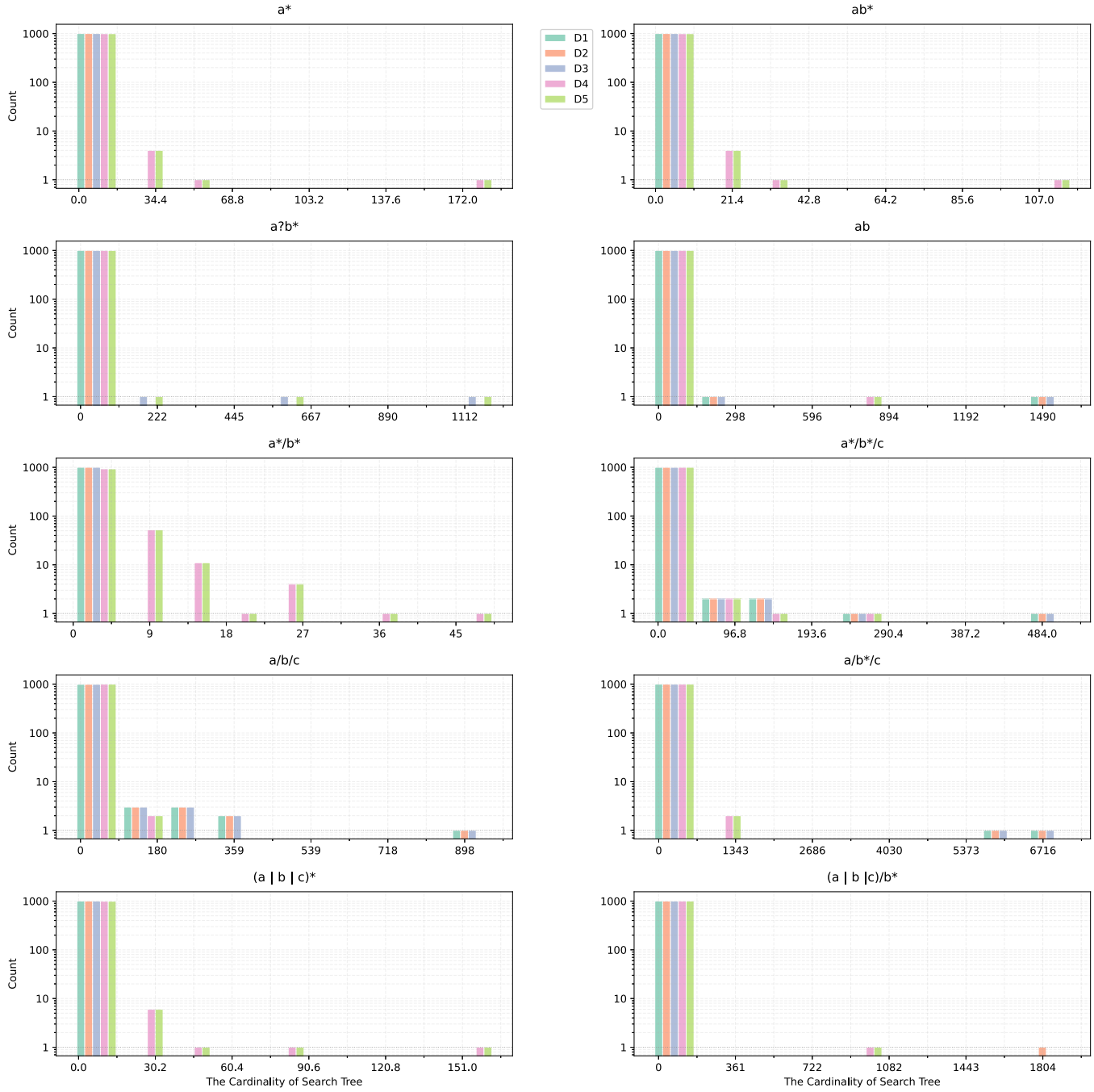


The following box figure show the distribution of running time on ICIJ-Paradise dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.



Telecom Dataset

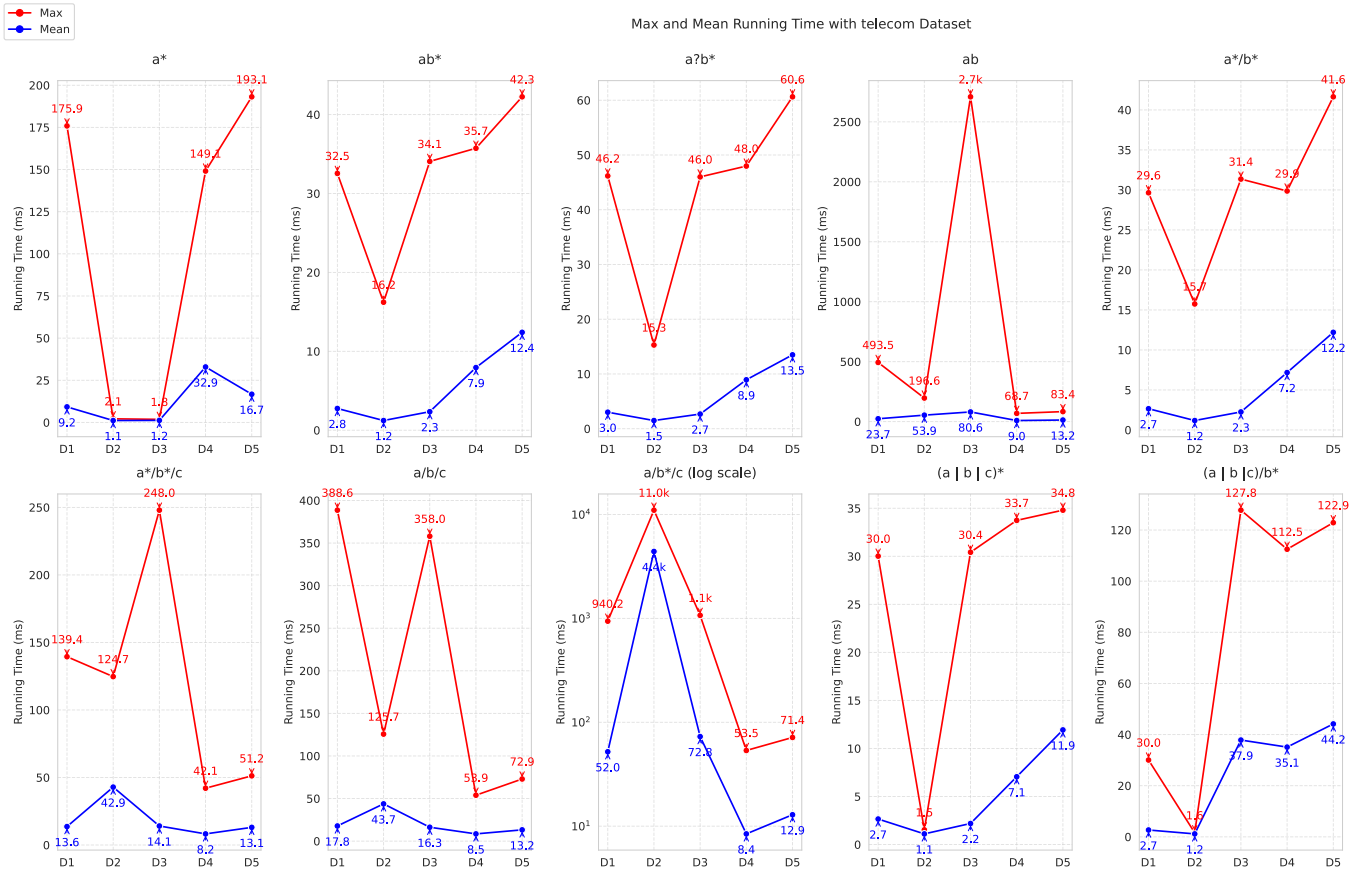
Time Performance

The following table shows statistics of regular path querie for each regular template on telecom dataset.

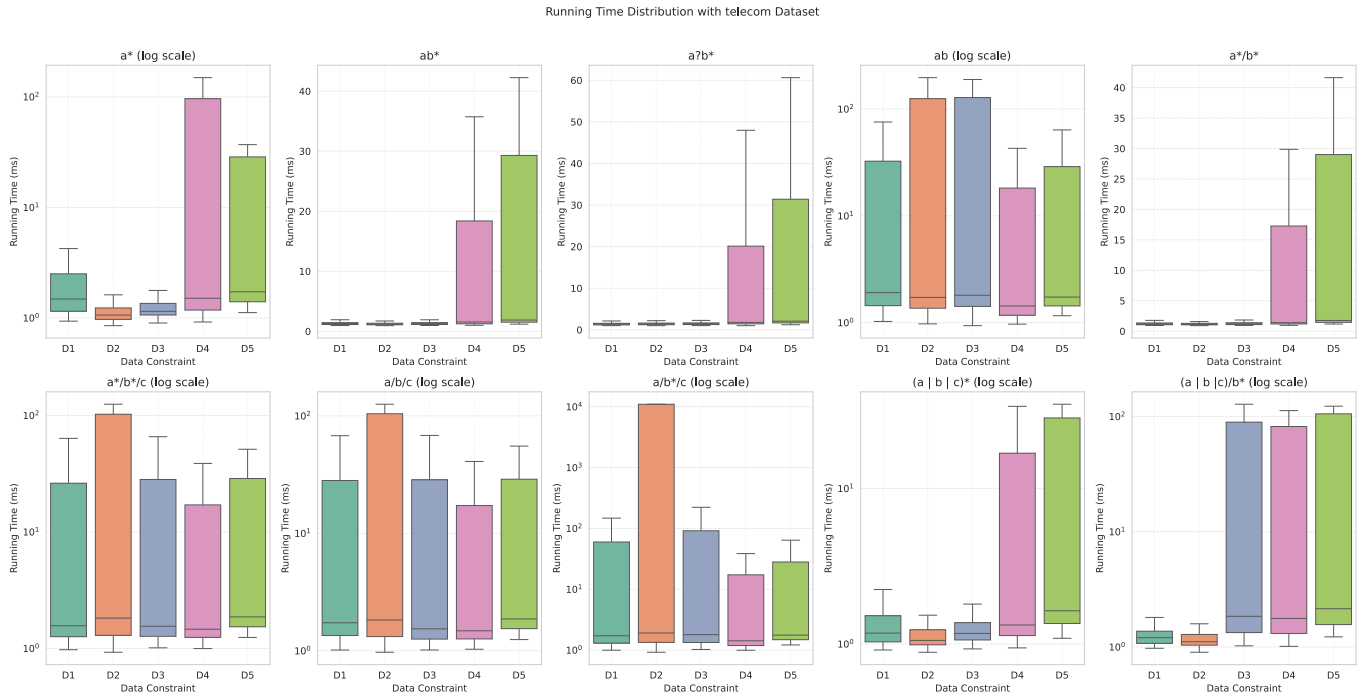
Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.26	10.54
ab*	0.3	14.97
a?b*	0.27	0.62

Regular Expression	Averge Time(ms)	Maximal Time(ms)
ab	0.35	15.03
a*/b*	0.27	1.01
a*/b*/c	0.22	16.79
a/b/c	0.2	0.42
a/b*/c	0.2	0.62
(a | b | c)*	0.2	0.53
(a | b | c)/b*	0.21	0.42

The following figure shows the maximal and mean running time of each RDPQ query.

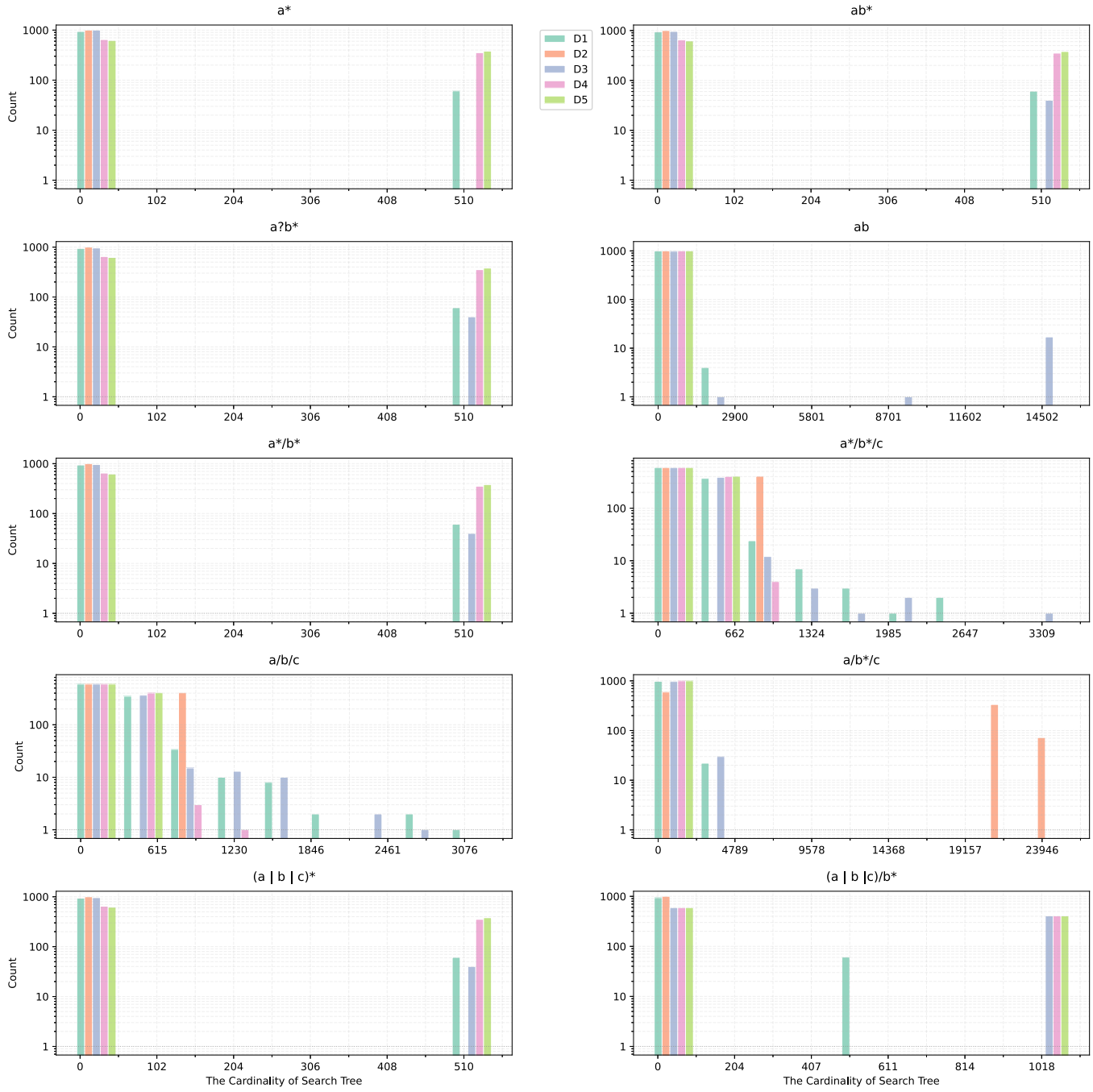


The following box figure show the distribution of running time on Telecom dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.



Pokec Dataset

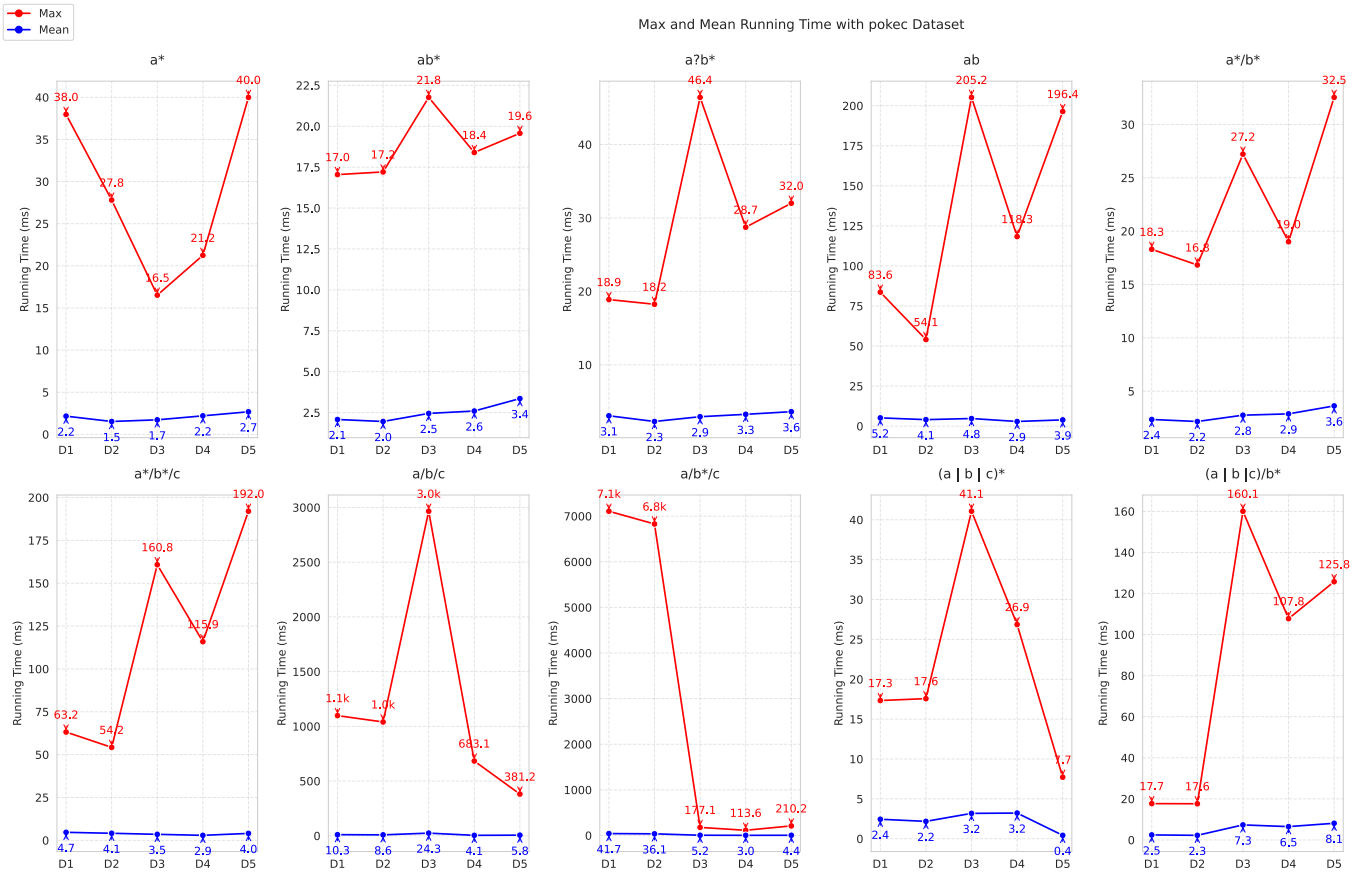
Time Performance

The following table shows statistics of regular path querie for each regular template on pokec dataset.

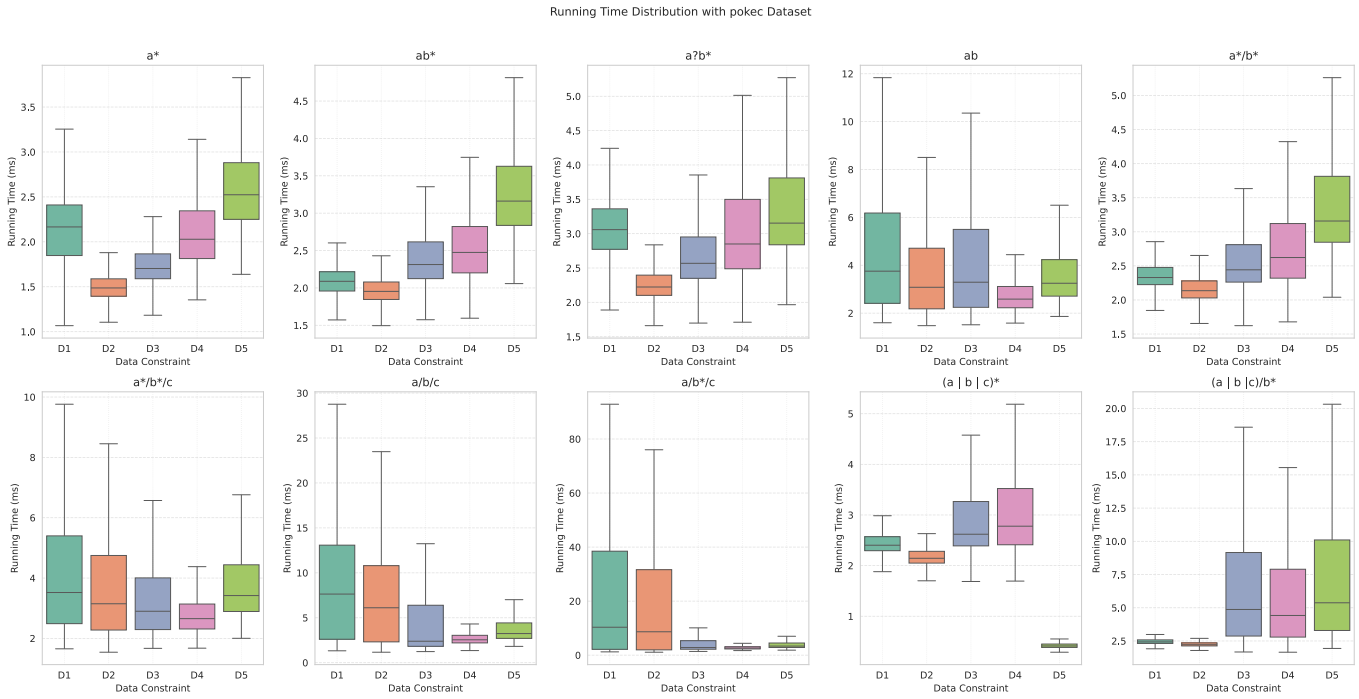
Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.33	12.75
ab*	0.39	1.04
a?b*	0.51	1.74
ab	0.56	14.86

Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*/b*	0.44	1.19
a*/b*/c	0.46	1.09
a/b/c	0.73	16.5
a/b*/c	0.46	1.24
(a | b | c)*	0.45	1.33
(a | b | c)/b*	0.47	1.16

The following figure shows the maximal and mean running time of each RDPQ query.



The following box figure show the distribution of running time on Telecom dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.

ICIJ-LEAK Dataset

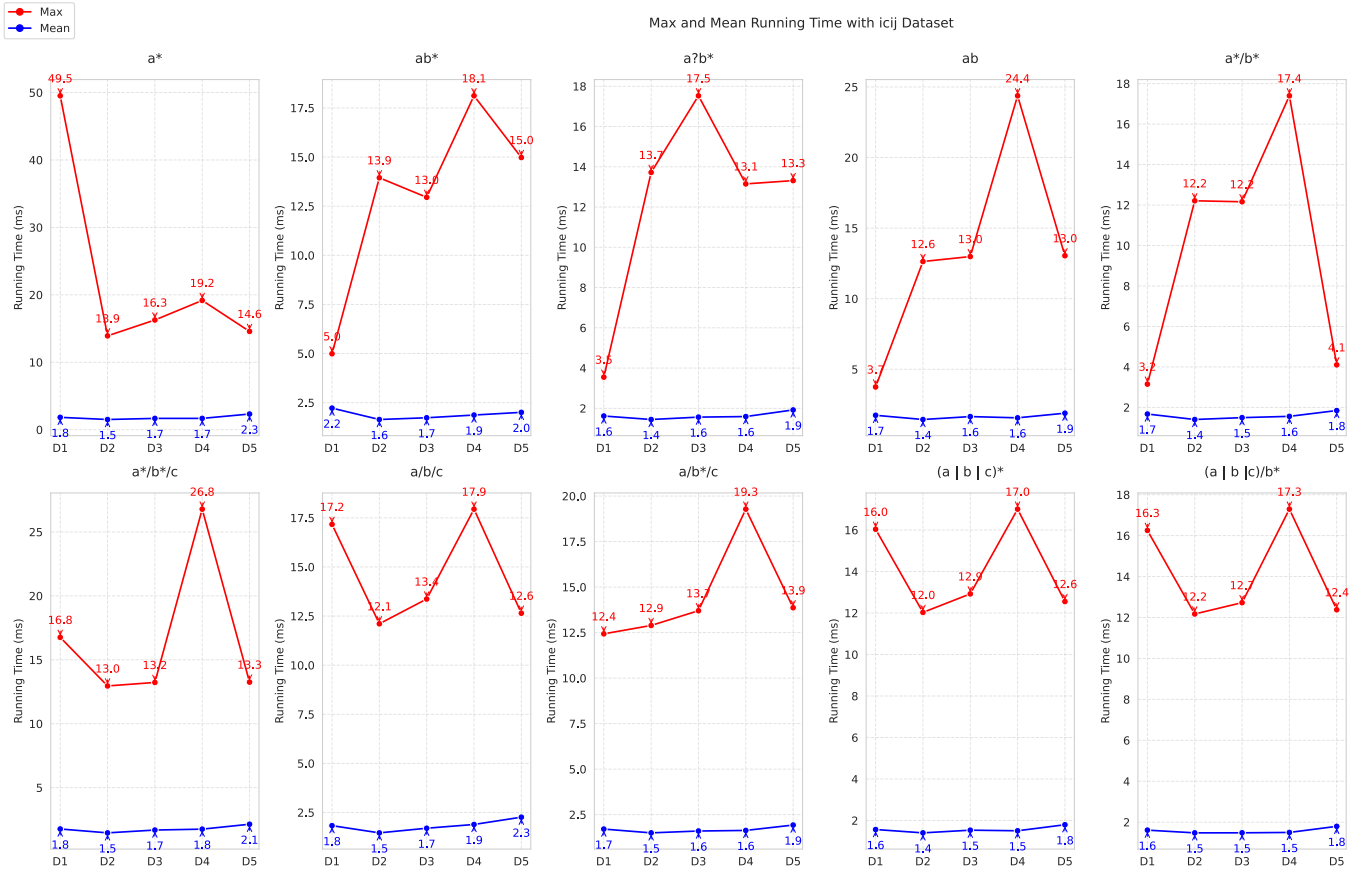
Time Performance

The following table shows statistics of regular path querie for each regular template on ICIJ-LEAK dataset.

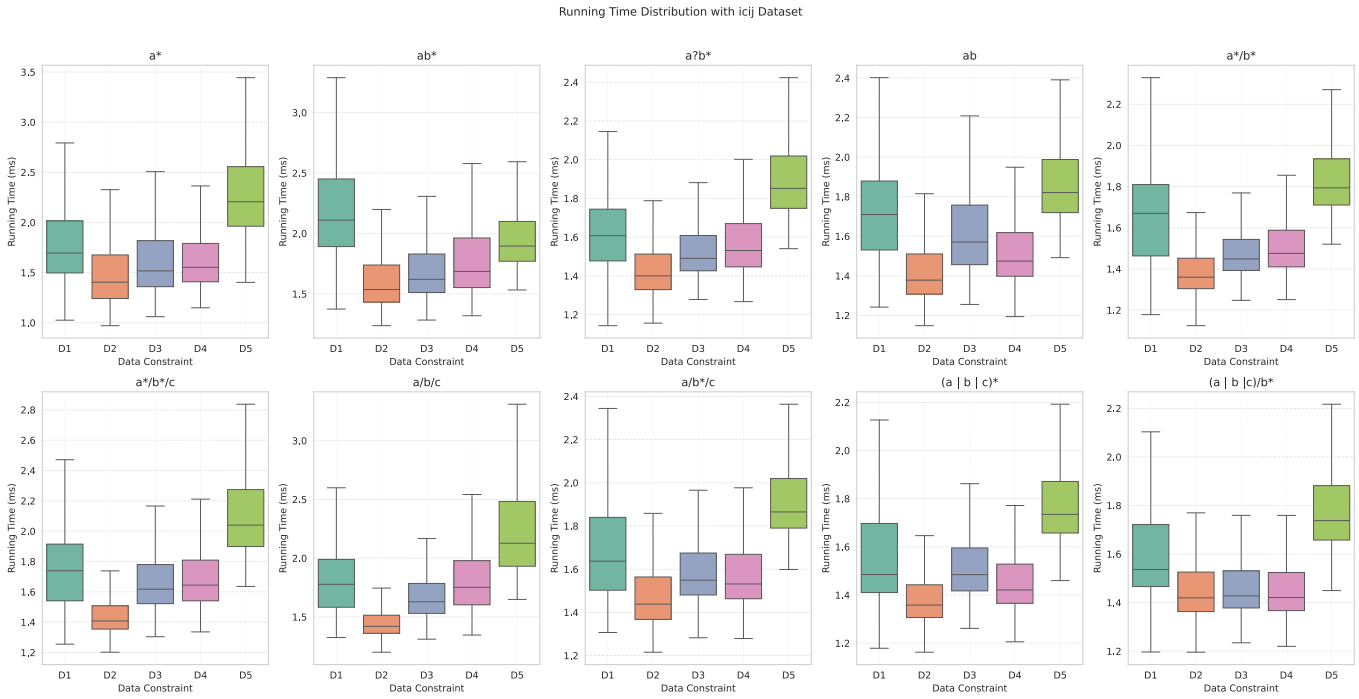
Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.4	20.01
ab*	0.51	15.95
a?b*	0.41	15.14
ab	0.39	15.3

Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*/b*	0.38	13.4
a*/b*/c	0.27	21.85
a/b/c	0.27	0.81
a/b*/c	0.36	1.16
(a | b | c)*	0.28	0.67
(a | b | c)/b*	0.28	0.77

The following figure shows the maximal and mean running time of each RDPQ query.

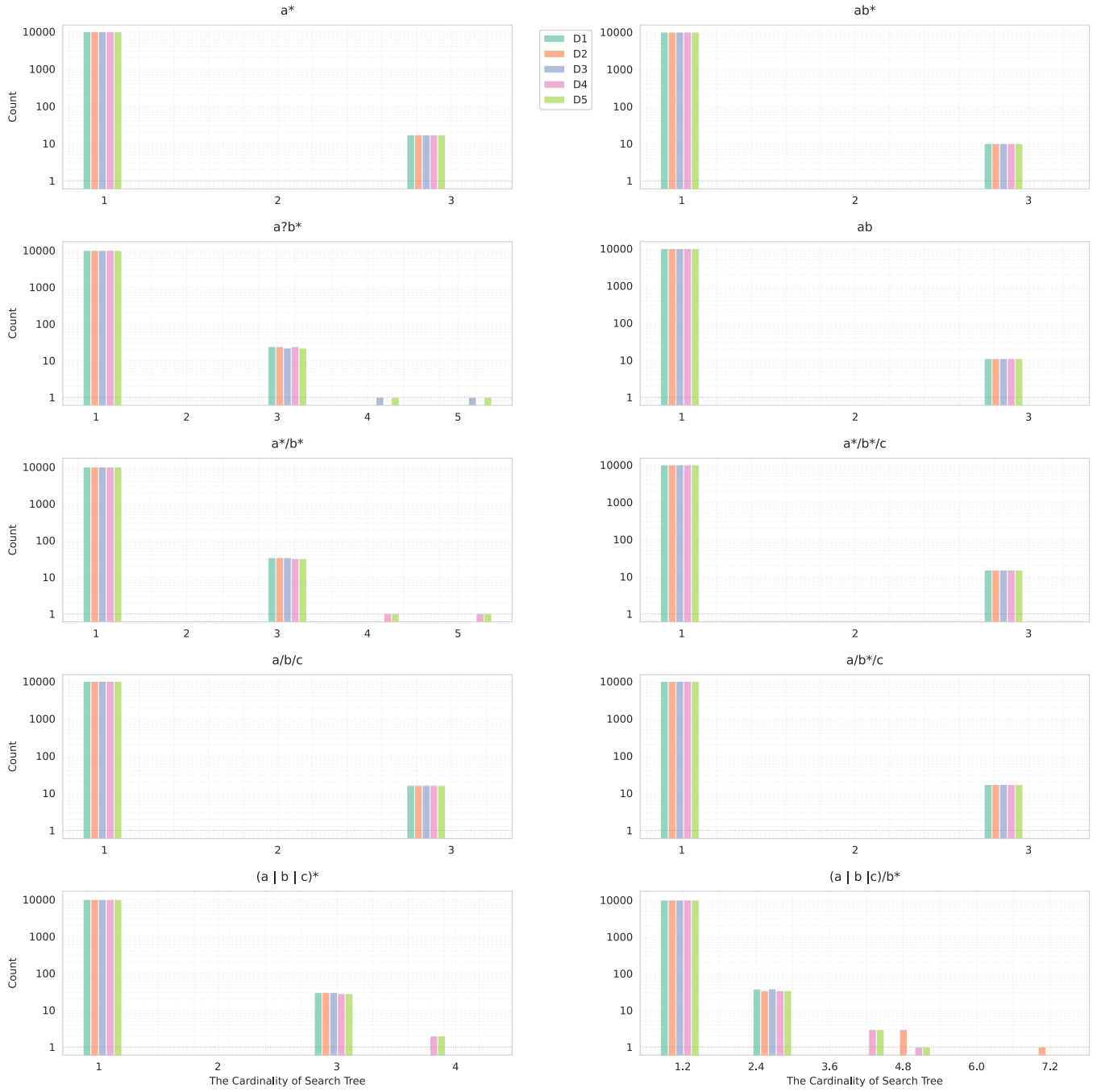


The following box figure show the distribution of running time on Telecom dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.



LDBC10 Dataset

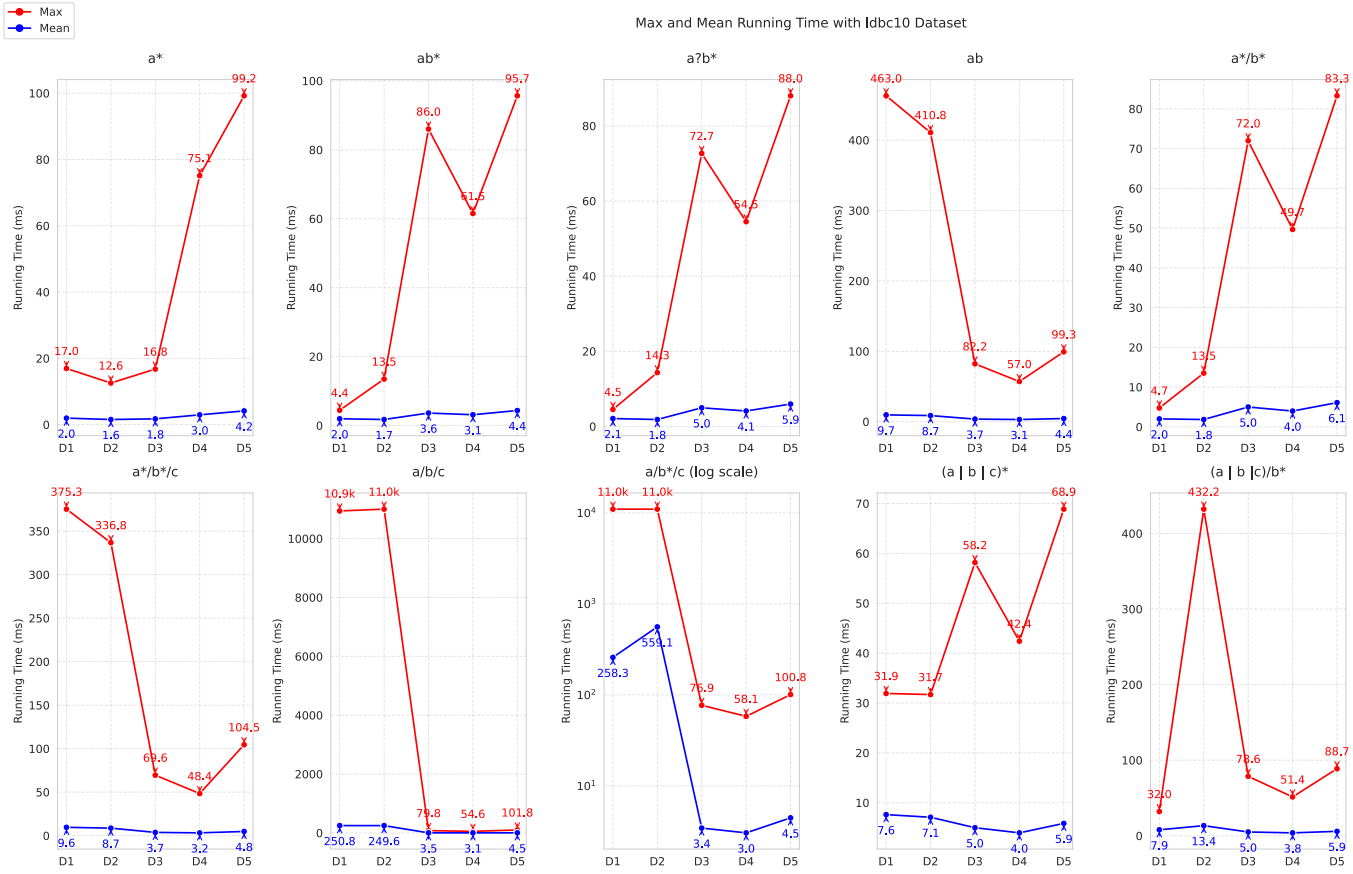
Time Performance

The following table shows statistics of regular path querie for each regular template on LDBC10 dataset.

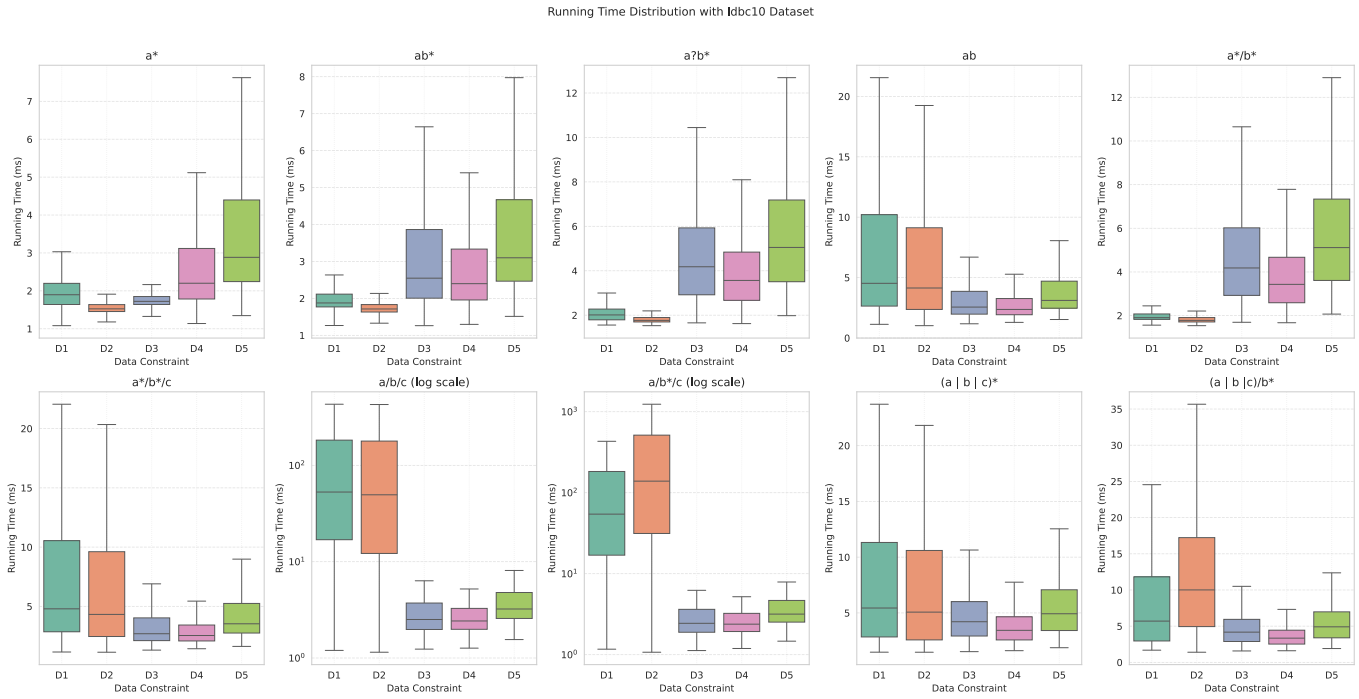
Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*	0.5	14.7
ab*	0.52	16.39
a?b*	0.51	15.46
ab	0.47	15.53

Regular Expression	Averge Time(ms)	Maximal Time(ms)
a*/b*	0.46	15.53
a*/b*/c	0.29	16.18
a/b/c	0.3	0.98
a/b*/c	0.27	1.22
(a | b | c)*	0.27	1.13
(a | b | c)/b*	0.29	1.27

The following figure shows the maximal and mean running time of each RDPQ query.



The following box figure show the distribution of running time on Telecom dataset



Search Tree Cardinality

The following figure shows the distribution of search tree cardinalities for each query.