PreseLinked Stream Data Processing Engines: Facts and Figures

ISWC'2012

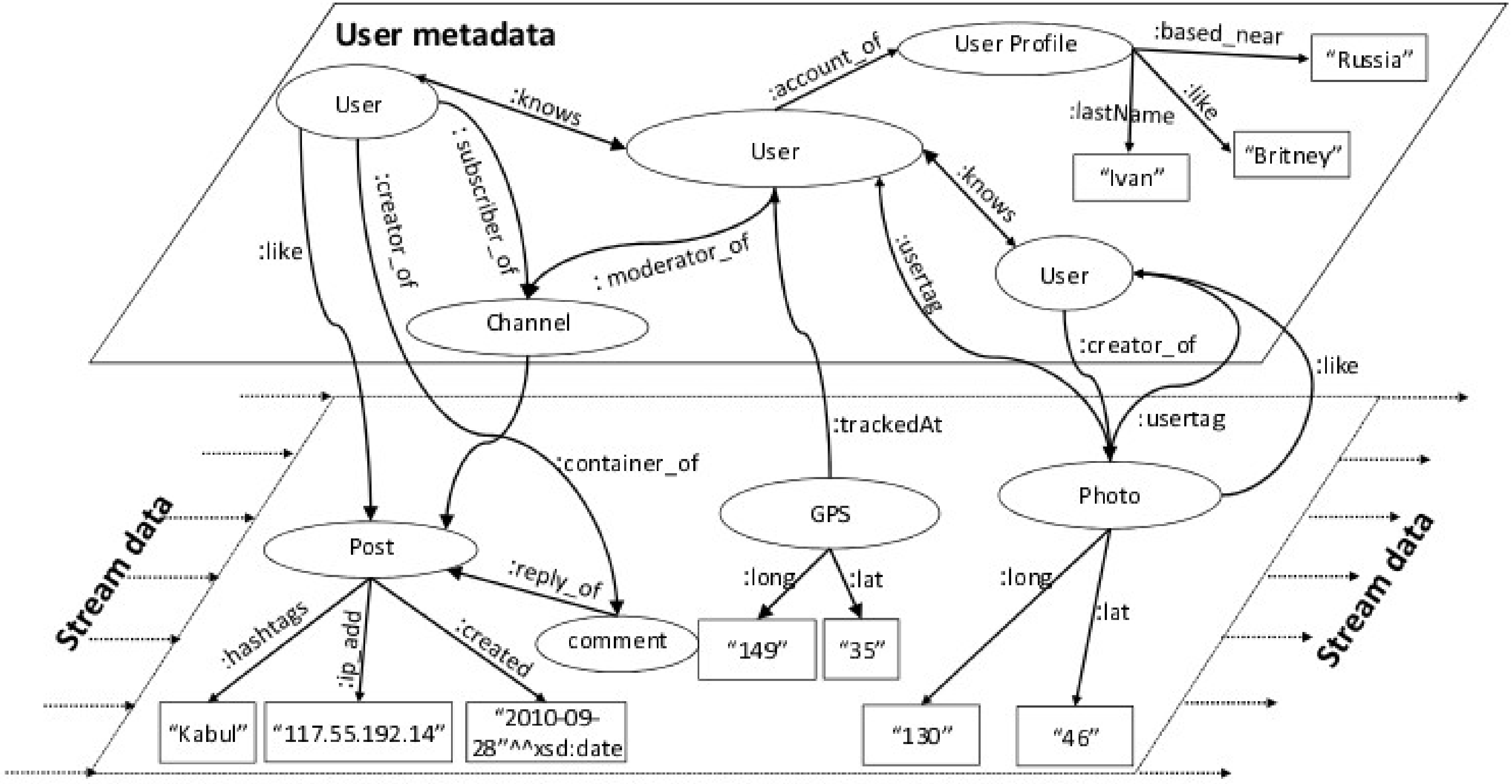
Motivation

Motivation

- Provide an open benchmarking framework to evaluate the linked stream data processing systems:
  - Functionality test
  - Correctness test
  - Performance test

Method and Result

Scenario: Social Network



Data Generator

- Stream Social network data Generator (S2Gen)
  - Add window to S3G2 (Scalable Structure Correlated Social Graph Generator): slide a window of users along all users, creates social activities for each user
  - Static data: generates the user profiles and the friendship information of all the users
- Parameters
  - Generating period: create streams with different sizes for scalability testing
  - Maximum number of activities per week: control throughput
  - Correlation probabilities: test the quary plan optimization capability

Functionality Test

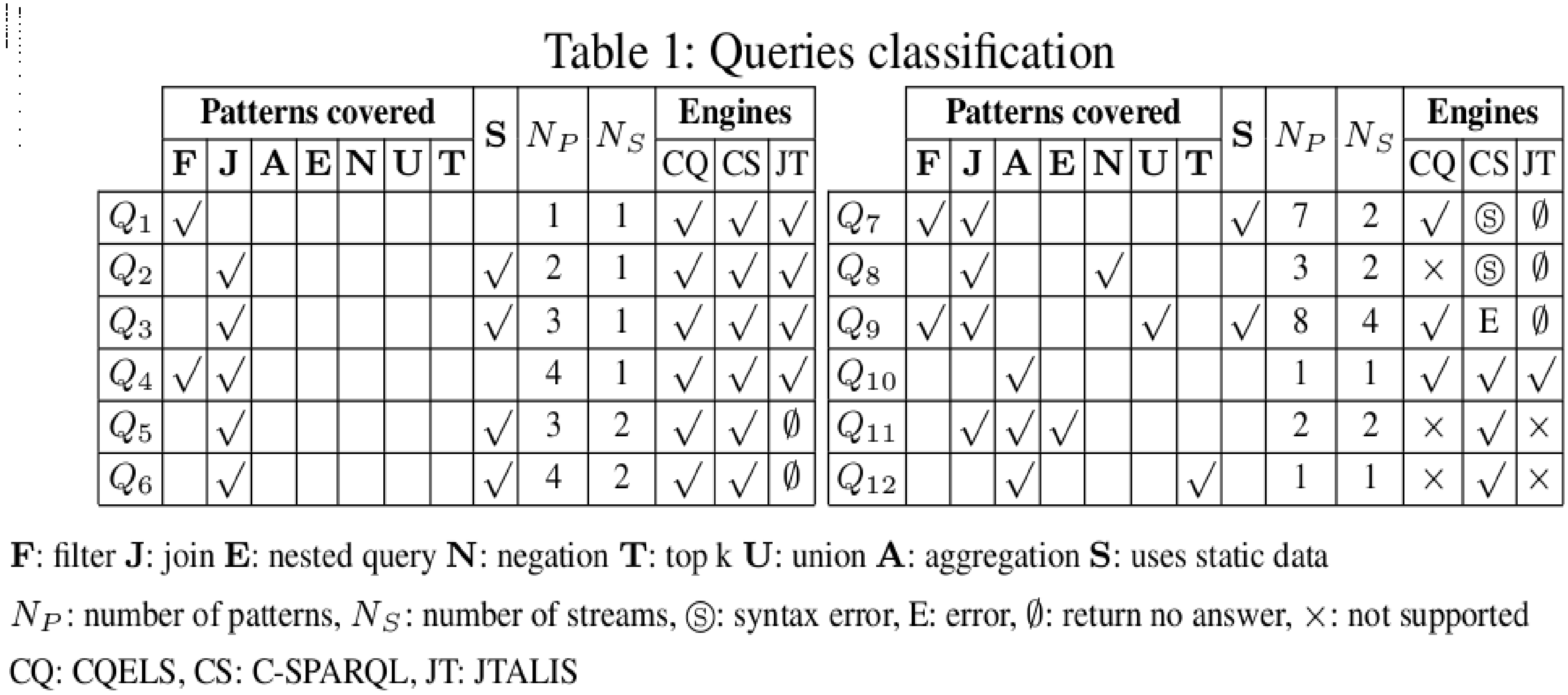


Table 1: Queries classification

	Patterns covered							S	N_P	N_S	Engines		
	F	**J**	**A**	**E**	**N**	**U**	**T**				CQ	CS	JT
Q_1	✓								1	1	✓	✓	✓
Q_2		✓						✓	2	1	✓	✓	✓
Q_3		✓						✓	3	1	✓	✓	✓
Q_4	✓	✓							4	1	✓	✓	✓
Q_5		✓						✓	3	2	✓	✓	∅
Q_6		✓						✓	4	2	✓	✓	∅
Q_7	✓	✓						✓	7	2	✓	Ⓢ	∅
Q_8		✓			✓				3	2	×	Ⓢ	∅
Q_9	✓	✓				✓		✓	8	4	✓	E	∅
Q_{10}			✓						1	1	✓	✓	✓
Q_{11}		✓	✓	✓					2	2	×	✓	×
Q_{12}			✓				✓		1	1	×	✓	×

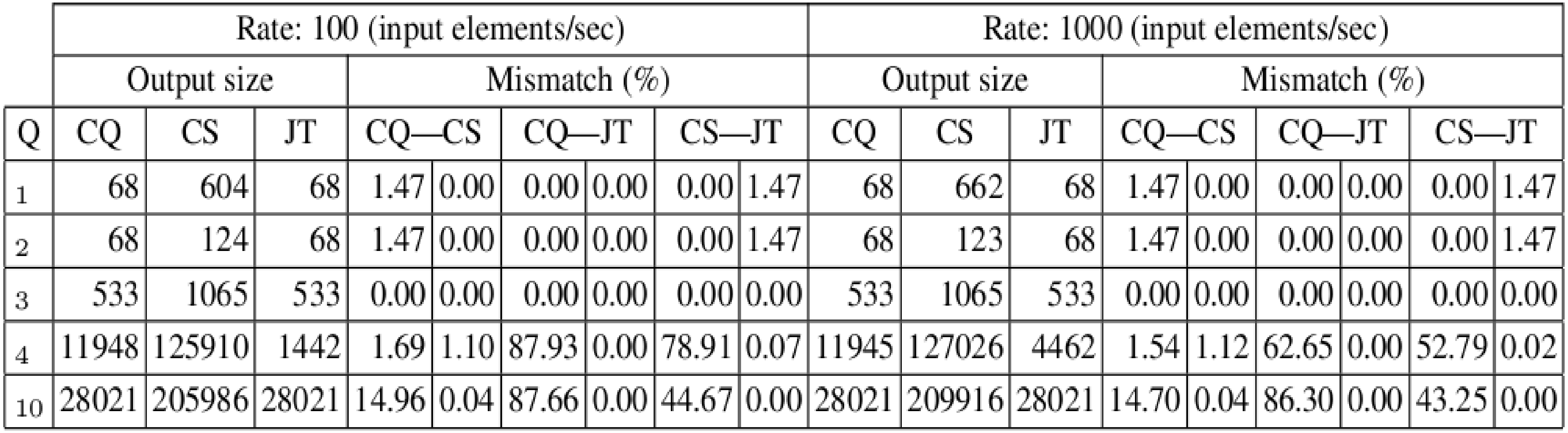
F: filter **J**: join **E**: nested query **N**: negation **T**: top k **U**: union **A**: aggregation **S**: uses static data

N_P: number of patterns, N_S: number of streams, Ⓢ: syntax error, E: error, ∅: return no answer, ×: not supported

CQ: CQELS, CS: C-SPARQL, JT: JTALIS

Correctness Test

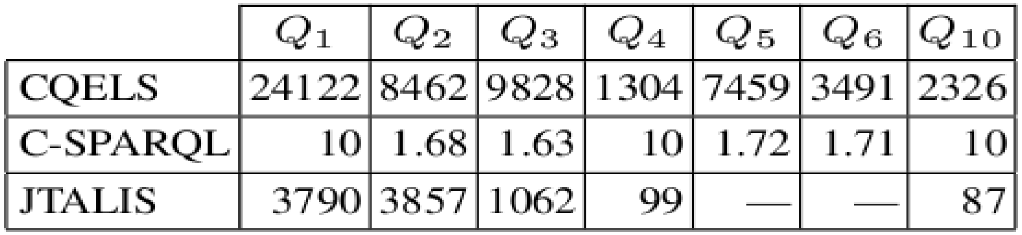
Table 2: Output Mismatch, $|U_{data}| = 219825$, $|S_{pc}| = 102955$



	Rate: 100 (input elements/sec)									Rate: 1000 (input elements/sec)								
	Output size			Mismatch (%)						Output size			Mismatch (%)					
Q	CQ	CS	JT	CQ—CS		CQ—JT		CS—JT		CQ	CS	JT	CQ—CS		CQ—JT		CS—JT	
1	68	604	68	1.47	0.00	0.00	0.00	0.00	1.47	68	662	68	1.47	0.00	0.00	0.00	0.00	1.47
2	68	124	68	1.47	0.00	0.00	0.00	0.00	1.47	68	123	68	1.47	0.00	0.00	0.00	0.00	1.47
3	533	1065	533	0.00	0.00	0.00	0.00	0.00	0.00	533	1065	533	0.00	0.00	0.00	0.00	0.00	0.00
4	11948	125910	1442	1.69	1.10	87.93	0.00	78.91	0.07	11945	127026	4462	1.54	1.12	62.65	0.00	52.79	0.02
10	28021	205986	28021	14.96	0.04	87.66	0.00	44.67	0.00	28021	209916	28021	14.70	0.04	86.30	0.00	43.25	0.00

Performance Test

Table 3: (Comparable) Maximum Execution Throughput



	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_{10}
CQELS	24122	8462	9828	1304	7459	3491	2326
C-SPARQL	10	1.68	1.63	10	1.72	1.71	10
JTALIS	3790	3857	1062	99	—	—	87

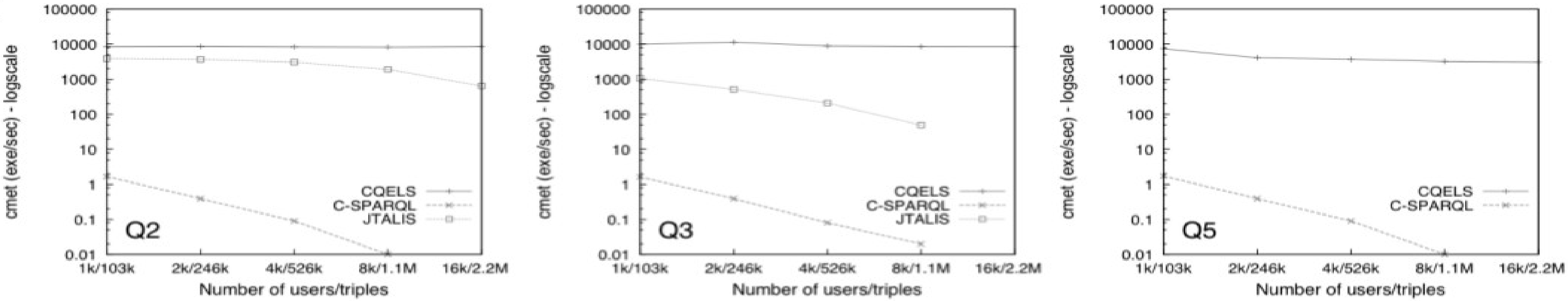


Fig. 2: Comparable max. execution throughput for varying size of static data.

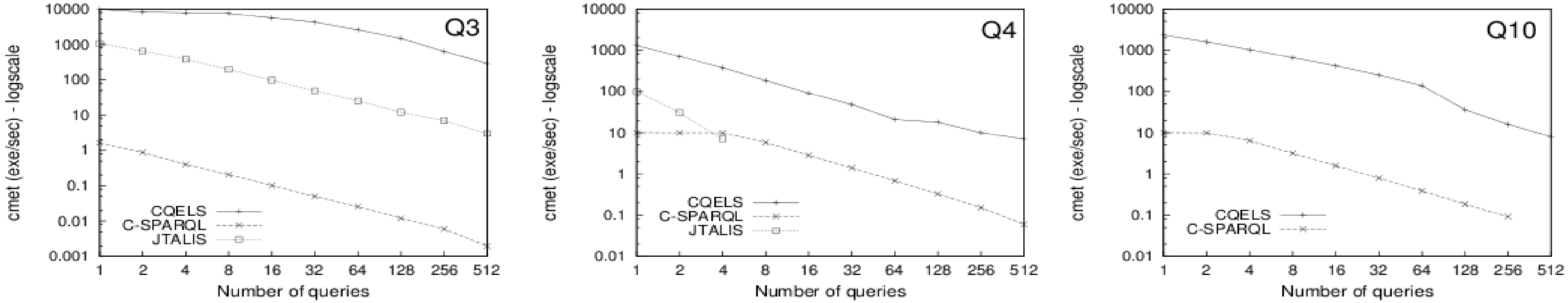


Fig. 3: Comparable max. execution throughput running multiple query instances.

Conclusion and Lesson

Conclusion

- CSPARQL: low scalability(static data), high “throughput”, low correctness
- CQELS: high scalability(static data), low throughput, high correctness
- JTALIS: low correctness, low functionality

Lesson

- We can provide big LSD evaluation framework or big LSB benchmark for current efficient/scalable LSD processing systems – CQELS-Cloud, TrOWL, SR on s4, etc
- The paper provides some criterias for our own stream reasoning system – throughput, static data size, complexity of linkage (correlation probabilities), number of streams, functionality, correctness, etc.