Rule-Based Graph Cleaning with GPUs on a Single Machine

Wenchao Bai, Wenfei Fan, Shuhao Liu, Kehan Pang, Xiaoke Zhu, Jiahui Jin*

MOTIVATION

Graph Data Quality Matters.
Ubiquitous graphs, inevitable errors.

Single Machine Systems.
Limited budget, privacy concerns.

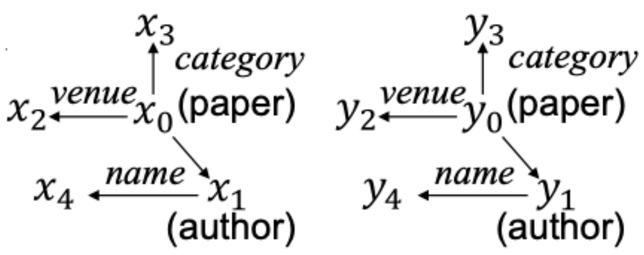
Trillema.
Generalizability, reliability, explainability.

PROBLEM DEFINITON

Graph Cleaning Rules (GCRs).

$$\mathrm{Q}[\mathrm{x}_0, y_0](X \rightarrow p_0)$$

$$p ::= x.A \oplus y.B \mid z.A \oplus c | M(x.A, y.B)$$



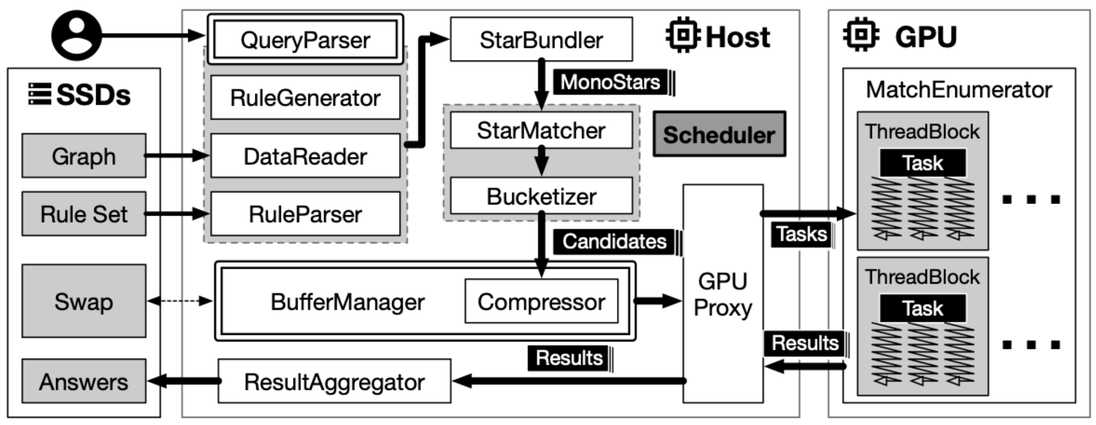
Graph Cleaning with GCRs.

- ✓ **Rule Discovery** $(G) \rightarrow \Sigma$
- ✓ **Error Detection** $(G, \Sigma) \rightarrow err$
- ✓ **Error Correction** $(G, \Sigma, err) \rightarrow fix$

ARCHITECTURE

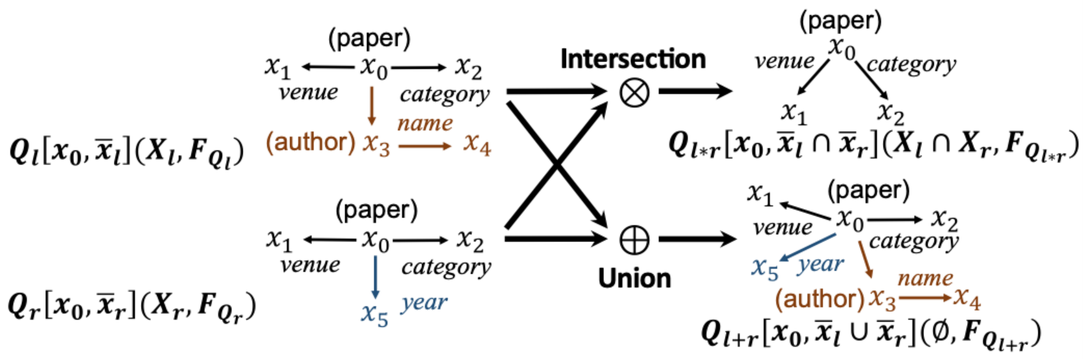
Hybrid Parallel Computing.
Multicore (CPU), SIMD (GPU).

Pipelined Architecture.
Eliminating data transferring overhead.



METHODOLOGIES

Computation: Bundled Matching



I/O: Conditional Succinct Table

Conditional succinct matches

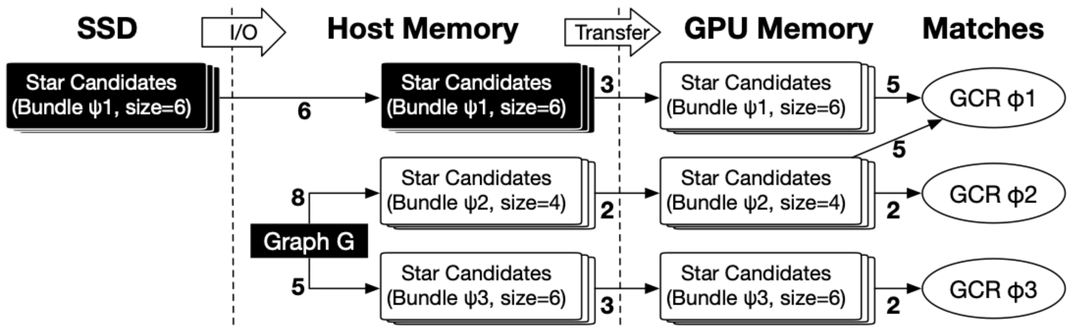
Matches			Conditions		
x_0.id	x_1.val	x_2.val	x_4.val	x_0.title	x_5.val
u_3	OSDI	CS	{J, S}	MR	null
u_4	OSDI	CS	{J, S}	GFS	null
u_1	OSDI	null	{M, P, J, S}	TF	2016
u_2	OSDI	null	{M, P, J, S}	tf	2016

Unfolded matches

x_0.id	...	x_4.val	...
u_3	...	J	...
u_3	...	S	...
...	...	...	...

Filtered out by $x_2.\mathrm{val} = \mathrm{CS}$

Utilization: Task Scheduling



EVALUATION

Table 3. Error detection efficiency with single-machine systems.

Graph & Metric	BioGRID		IMDB		SemScholar	
	Time (s)	I/O (GB)	Time (s)	I/O (GB)	Time (s)	I/O (GB)
MiniClean	259.6	18.97	325.5	23.82	2993.7	92.54
CoroGraph/Blaze	OOM	OOM	OOM	OOM	OOM	OOM
MiniGraph	65.34×	10.73×	TO	TO	TO	TO
HyperBlocker	11.29×	5.74×	OOM (GPU)	OOM (GPU)	OOM (GPU)	OOM (GPU)

◇ "OOM" denotes out-of-memory, "TO" denotes timeout after 8h.

Table 4. Efficiency and accuracy of MiniClean vs. ML models.

System	BioGRID			IMDB			SemScholar		
	Time	ER-F1(%) all (real)	CR-F1(%) synthetic	Time	ER-F1(%) all (real)	CR-F1(%) synthetic	Time	ER-F1(%) real	CR-F1(%) real
MiniClean	312.7s	98.7 (97.5)	97.2	409.1s	99.9 (94.0)	80.4	4089.1s	94.6	70.6
Ditto	7.7×	91.2 (91.0)	N/A	8.0×	95.6 (90.8)	N/A	4.3×	90.3	N/A
KGClean	11.9×	N/A	54.9	64.2×	N/A	20.0	11.9×	N/A	29.8

Findings.

- ✓ **More Efficient**. Faster than SOTA GPU-based system by over $11\times$.
- ✓ **More Accurate**. Outperform SOTA ER model by 4.3% and CR model by 40.4%