SparkER: Scaling Entity Resolution in Spark

Luca Gagliardelli
University of Modena
and Reggio Emilia, Italy
luca.gagliardelli@unimore.it

Giovanni Simonini
MIT CSAIL
Cambridge, MA, USA
giovanni@csail.mit.edu

Domenico Beneventano
University of Modena
and Reggio Emilia, Italy
domenico.beneventano@unimore.it

Sonia Bergamaschi
University of Modena
and Reggio Emilia, Italy
sonia.bergamaschi@unimore.it

Entity Resolution (ER)

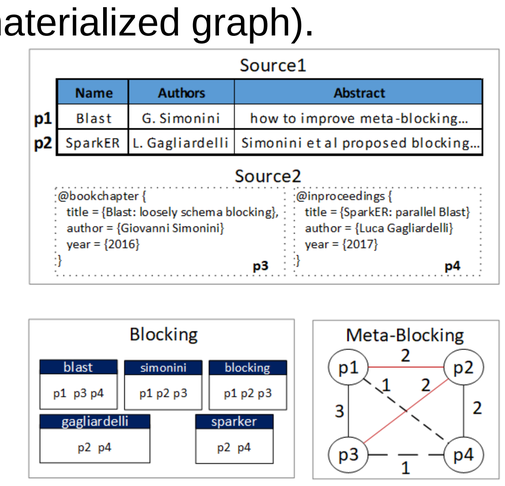
Entity Resolution is the task of identifying different representations (profiles) that pertain to the same real-world entity.

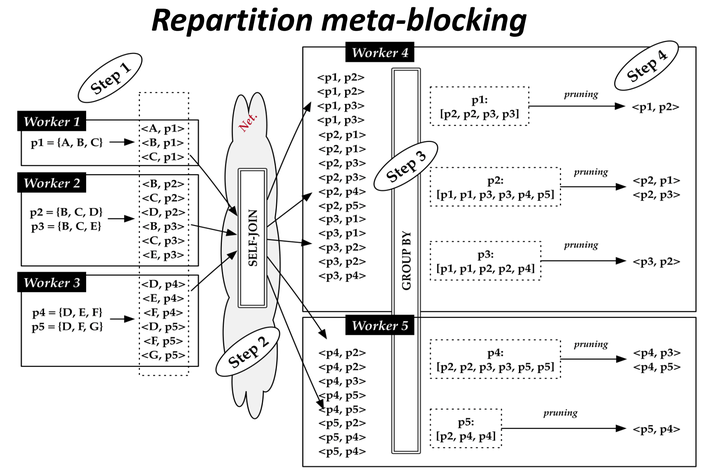
Application areas:
Data Integration, Duplicate Detection, Fraud Detection, etc.

Hard to Scale:
all pairwise comparisons among profiles has complexity $O(n^2)$

Scalable ER Techniques

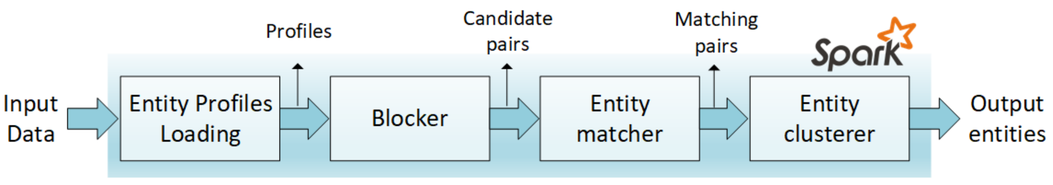
Blocking and **meta-blocking** are used to reduce the number of comparisons. **Parallel meta-blocking** ER techniques [3] to manage Big Data were proposed: all of them are based on the **repartition join** (not efficient as generates a large materialized graph).





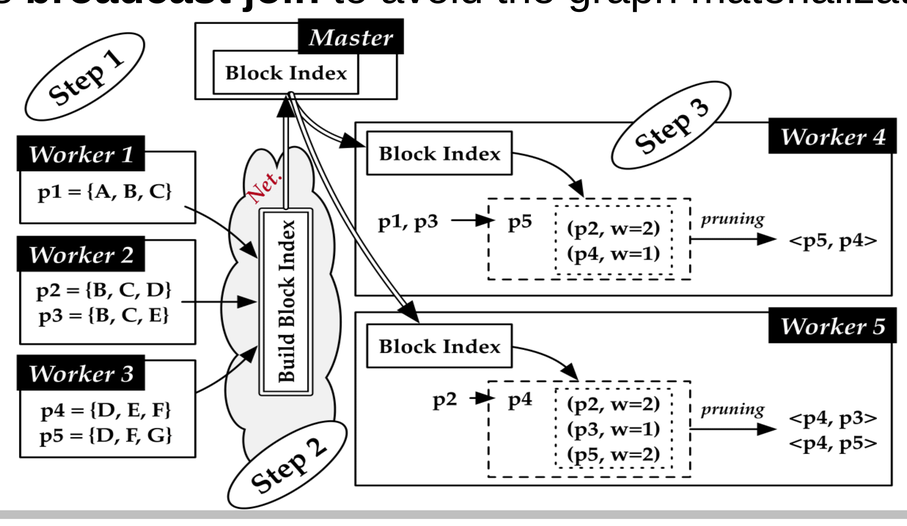
SparkER

Implements our new **broadcast meta-blocking** method by using Apache Spark.
Provides both schema-agnostic and loosely schema-aware blocking techniques [4].
Covers the main ER tasks.



Broadcast meta-blocking

A novel optimized meta-blocking parallelization strategy that exploits **broadcast join** to avoid the graph materialization.



Experimental Results

Datasets characteristics

	$\|P_1\|$	$\|P_2\|$	$\|D_P\|$
Citations	1.8M	2.5M	0.6M
DBpedia	1.2M	2.2M	0.9M
Freebase	4.2M	3.7M	1.5M

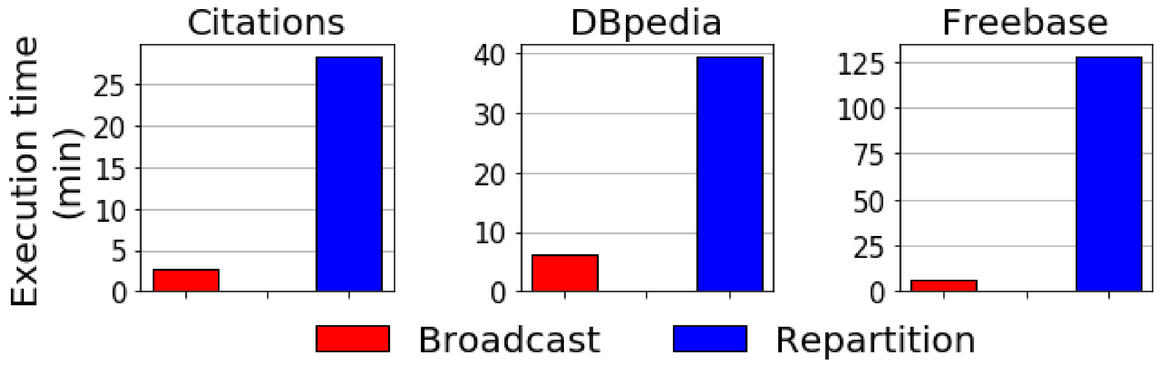
Node characteristics

CPU	2 x Intel Xeon E5-2670v2 2.5 GHz (20 cores)
RAM	128 GB
OS	Ubuntu 14.04

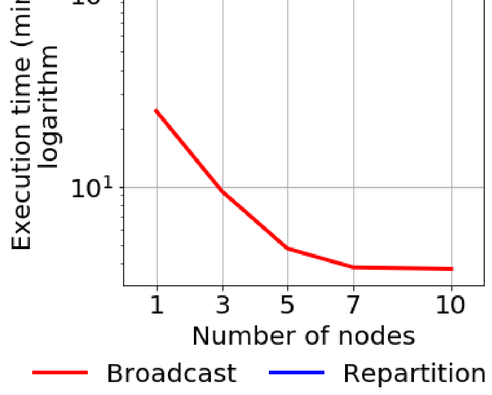
Repartition meta-blocking is not able to run on less than 7 nodes.
Broadcast meta-blocking can run also on a single node.

Broadcast meta-blocking always outperforms repartition meta-blocking: on 10 nodes it is 5 to 12 times faster.

Meta-blocking execution time on 10 nodes



Meta-blocking scalability (on Freebase)



Conclusions

Broadcast meta-blocking
- A novel meta-blocking parallelization strategy;
- Outperforms the state-of-the-art parallel implementations in term of execution time and resources consumption.

SparkER
- A highly scalable tool for Entity Resolution with Apache Spark;
- Implements both schema-agnostic and loosely schema-aware blocking techniques;
- Implements the main ER tasks.

References

1. Simonini, G., Gagliardelli L., Bergamaschi, S., & Jagadish, H. V. (2019). Scaling Entity Resolution: A Loosely Schema-aware Approach. *Information Systems*.
2. Gagliardelli, L., Zhu, S., Simonini, G., & Bergamaschi, S. (2018). Bigdedup: a Big Data integration toolkit for duplicate detection in industrial scenarios. In *25th International Conference on Transdisciplinary Engineering (TE2018)* (Vol. 7, pp. 1015-1023).
3. Efthymiou, V., Papadakis, G., Papastefanatos, G., Stefanidis, K., & Palpanas, T. (2017). Parallel meta-blocking for scaling entity resolution over big heterogeneous data. *Information Systems*, 65, 137-157.
4. Simonini, G., Bergamaschi, S., & Jagadish, H. V. (2016). BLAST: a loosely schema-aware meta-blocking approach for entity resolution. *PVLDB* 9(12): 1173-1184 (2016).

MIT CSAIL

SparkER

DBGROUP

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA