Benchmarking Filtering Techniques for Entity Resolution

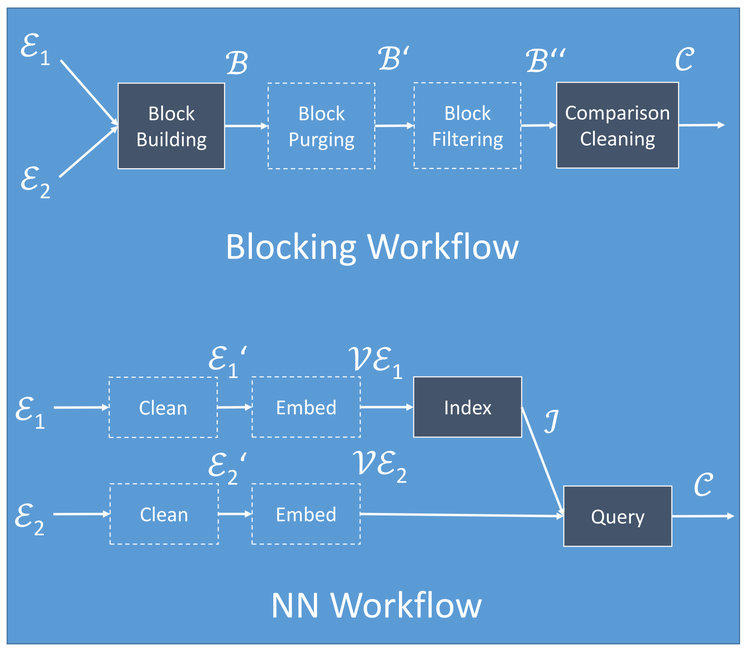
Entity resolution wants to identify pairs of entity profiles that represent the same real-world object. Checking every possible pair of entity profiles uses a lot of time and computational resource, so filtering techniques are used to limit these comparisons to only the most promising candidates. As there are many different filtering techniques, we want to compare them systematically.

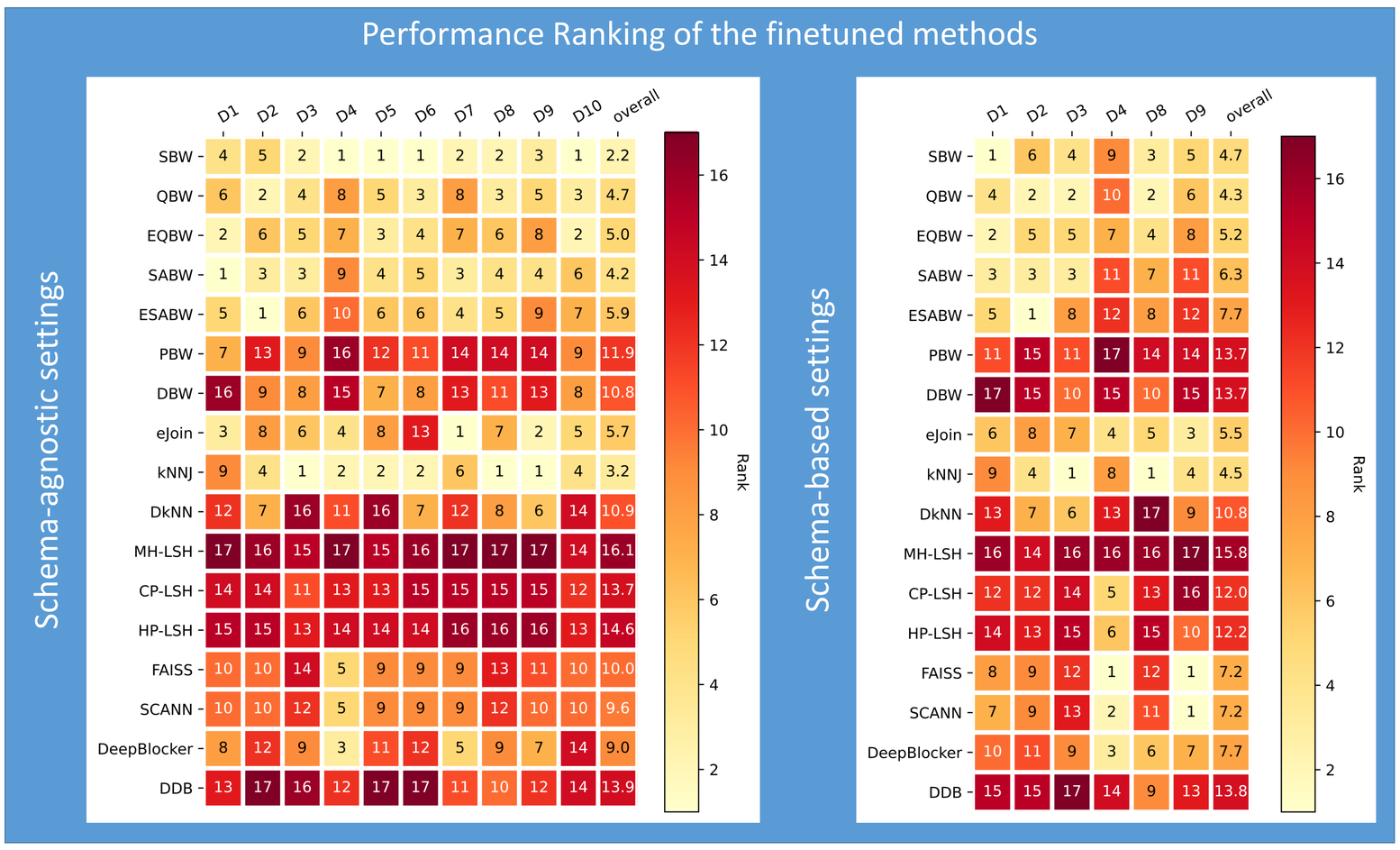
Measures of Effectiveness:

- Pair Completeness (PC)
  - $\frac{D(Candidates) \cap D(matches)}{D(matches)}$
- Pairs Quality (PQ)
  - $\frac{D(Candidates) \cap D(matches)}{D(Candidates)}$
- Runtime

Problem Statement (Configuration Optimization):
Given are two sets of entity profiles, a filter method and a threshold on PC (90%), finetune the parameters of the optimization such that the resulting PQ is maximized while the PC is above the threshold.

	Method	Number of Configurations
Blocking Methods	Standard Blocking	3,440
	Q-Grams Blocking	17,200
	Extended Q-Grams Blocking	68,800
	(Ex.) Suffix Arrays Blocking	21,285
Sparse NN Methods	ε-Join	6,000
	kNN-Join	12,000
Dense NN Methods	MH-LSH	168
	HP-LSH	400
	CP-LSH	2,000
	FAISS	2,720
	SCANN	10,880
	DeepBlocker	2,720





Datasets:

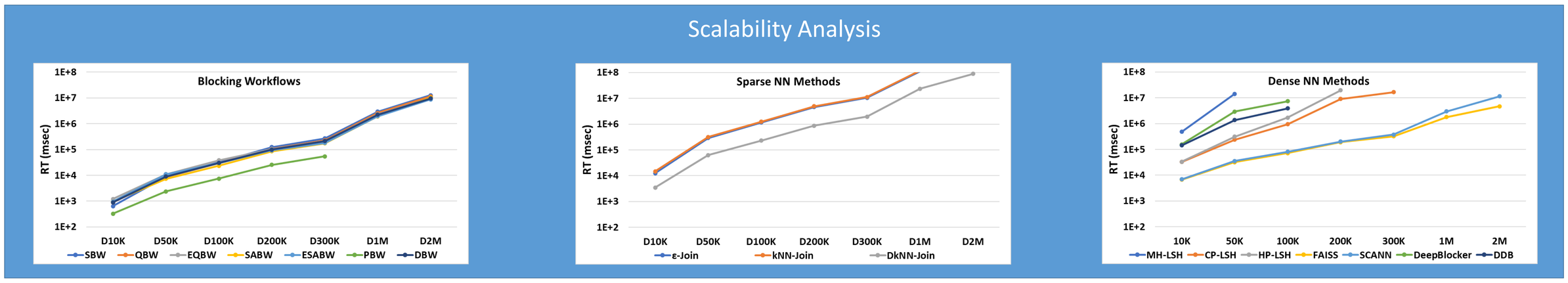
- 10 real world Clean-Clean ER datasets
  - between a few hundred and over 60,000 entries per file
  - Used for performance comparison
- 7 synthetic Dirty ER datasets
  - between 10,000 and 2 Million entries
  - Used for scalability analysis

Taxonomies:

- Scope:
  - Syntactic vs. Semantic representations
  - Schema-based vs. Schema-agnostic settings
- Internal functionality:
  - Lazy vs. Proactive
  - Deterministic vs. Stochastic
  - Similarity based vs. Cardinality based threshold

Take-away Messages:

- Performance heavily depends on dataset characteristics
- Parameter finetuning increases the performance significantly
- Schema-agnostic settings offer a more stable performance
- Cardinality based thresholds are preferable
- Syntactic representations are preferable



George Papadakis, Marco Fisichella, **Franziska Schoger**, George Mandilaras, Nikolaus Augsten, Wolfgang Nejdl

Forschungszentrum · Research Center L3S

https://github.com/gpapadis/ContinuousFilteringBenchmark