# Unveiling Dis-Integration

George Papadakis
*Public Power Corporation*
g_a.papadakis@ppcgroup.com

Ekaterini Ioannou
*Tilburg University*
ekaterini.ioannou@uvt.edu

Yannis Velegrakis
*University of Trento & Utrecht University*
i.velegrakis@uu.nl

## I. Background

Entity Resolution (ER) has been extensively studied over the last decade, with a plethora of algorithmic solutions, techniques, and methodologies having been proposed [1]. The individual state-of-the-art ER algorithms are offered through open-source systems, such as Magellan [2] and JedAI [3], which typically implement end-to-end solutions through a sequence of workflow steps. Each workflow step requires its own special configuration and fine tuning, thus turning the creation of complete ER solutions into a non-trivial, time-consuming process that requires adapting, among others, to the characteristics of the data to be resolved (e.g., relational, semi-structured, etc.), to its intrinsic noise (e.g., misspellings, abbreviations, etc.) as well as to application constraints (e.g., execution time).

A core assumption of existing works is that the same end-to-end pipeline should be apply to all entities of a particular dataset. In contrast, we argue that different algorithms are more suitable for different parts of a dataset depending on their entity type. Our approach *addresses Entity Resolution through Dis-Integration*, dividing the given dataset into fragments and then detecting the best ER pipeline per fragment. As such, each fragment might be processed by different pipelines, or by pipelines with the same methods but with different configurations, or by the same pipeline and configuration.

## II. Results

Our dis-integration approach includes various novel mechanisms focusing on: (i) powerful data profiling using features with simple metadata, aggregated statistics, etc; (ii) partition algorithms that generate dataset fragments; (iii) association of the data fragments to one of several fine-tuned ER pipelines that is expected to yield the highest performance; (iv) combing all mechanisms together into a complete approach.

The introduced approach was investigated with respect to both the motivation as well as the overall improvement in resolution. For the former, we measured the performance of various state-of-the-art ER pipelines over 10 established datasets, with the results validating the significant differences in their performance when applied to the same data, as envisaged by the motivation behind this work. For the latter, we measured the overall performance and efficiency of our system. This includes comparing against a single, established ER approach as well as assessing the effectiveness of our novel algorithm recommendation and data partitioning mechanisms with 3.3 million entities. The results illustrate that combining evidence from the content of entity profiles with schema information yields $\sim$700 entity types, whose independent resolution reduces recall to a low extent ($<$5%) for a significantly lower number of pairs that are processed ($>$50%), thus enhancing time efficiency and scalability at a limited cost in effectiveness.

## Acknowledgments

This research was partially funded by the Horizon Europe project STELAR (GA No. 101070122).

## References

[1] G. Papadakis, E. Ioannou, E. Thanos, and T. Palpanas, *The Four Generations of Entity Resolution*, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2021.

[2] A. Doan, P. Konda, P. Suganthan GC, Y. Govind, D. Paulsen, K. Chandrasekhar, P. Martinkus, and M. Christie, "Magellan: toward building ecosystems of entity matching solutions," *Communications of the ACM*, vol. 63, no. 8, pp. 83–91, 2020.

[3] G. Papadakis, L. Tsekouras, E. Thanos, N. Pittaras, G. Simonini, D. Skoutas, P. Isaris, G. Giannakopoulos, T. Palpanas, and M. Koubarakis, "Jedai3: beyond batch, blocking-based entity resolution," in *EDBT*, 2020, pp. 603–606.