Efficient Evidence Accumulation Clustering for large datasets

Diogo Silva[1], Helena Aidos[2] and Ana Fred[2]

[1]Portuguese Air Force Academy, Sintra, Portugal

[2]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal

dasilva@academiafa.edu.pt, {haidos, afred}@lx.it.pt

Introduction

- Evidence Accumulation Clustering (EAC) is a robust ensemble method but its computational complexity restricts its use to small datasets.
- We propose an optimized implementation of the different EAC steps for faster execution and decreased memory usage.

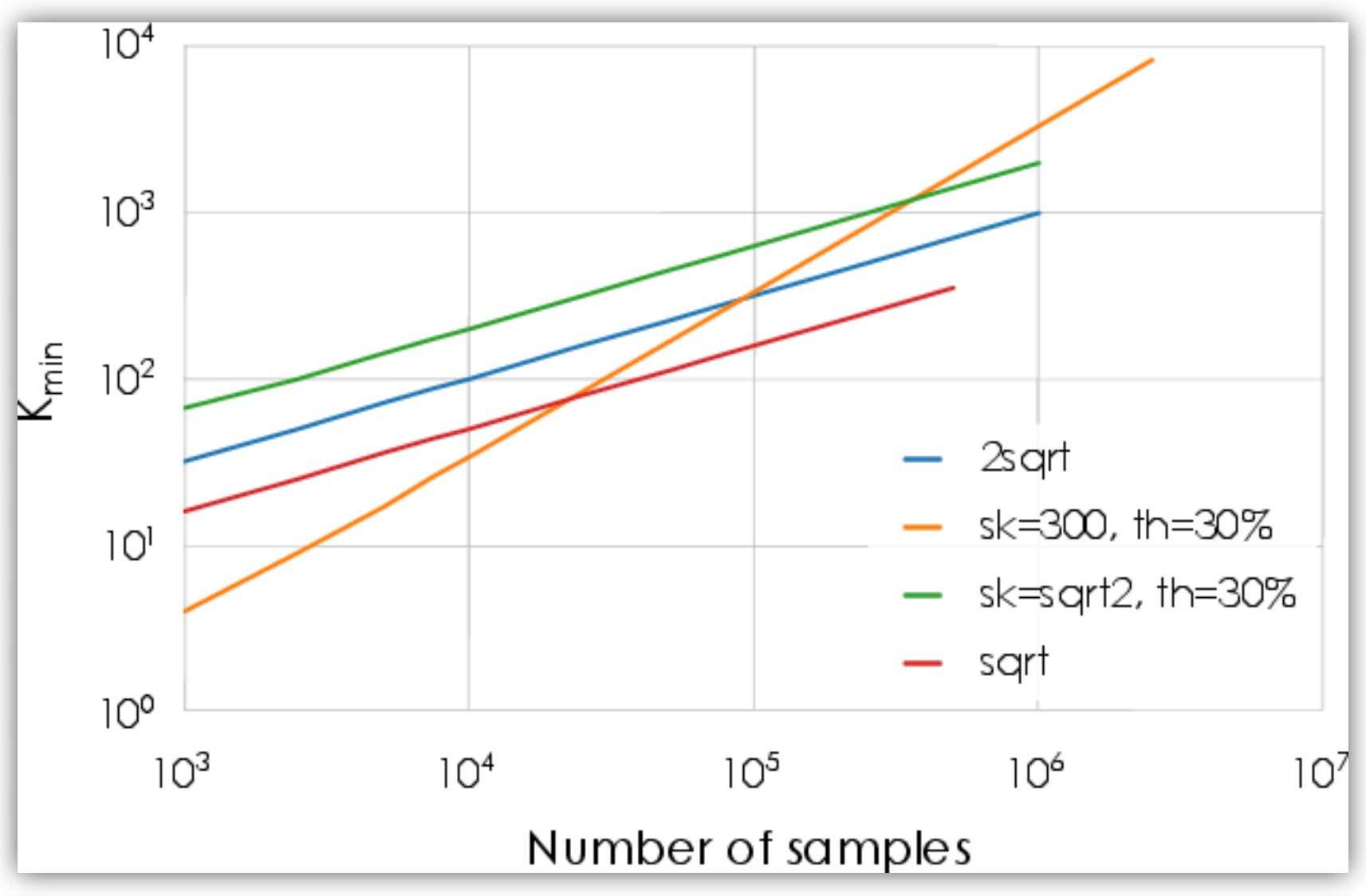
Validation and Speed-up

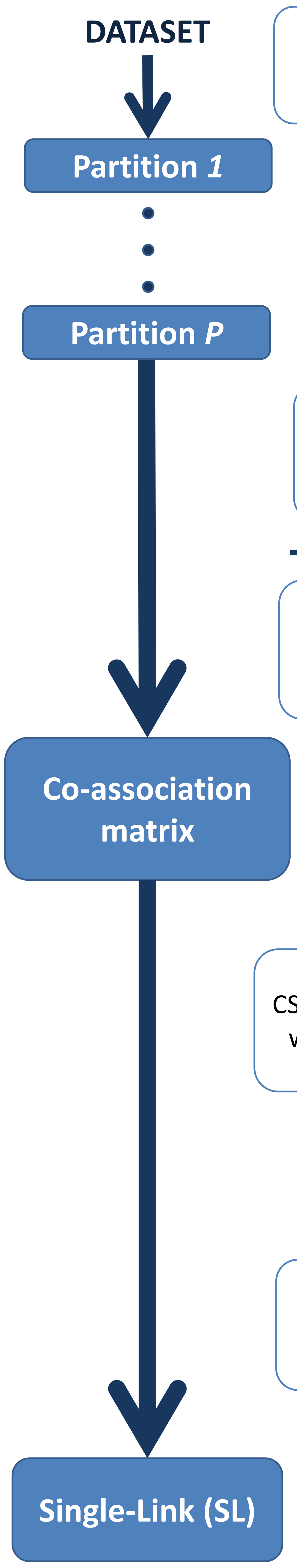
- The clustering accuracy of the optimized version relative to the original on several small benchmark datasets is negligible, validating its use on large datasets.
- Speed-up over the original version on small datasets varied between 6 and 200 on the different EAC phases.

Rules for Ensemble

The rules for the minimum and maximum number of clusters of the ensemble have a big impact on performance and memory usage. Four rules were tested.

Evolution of K_{min} with different rules





DATASET

Partition *1*

Partition *P*

Co-association matrix

Single-Link (SL)

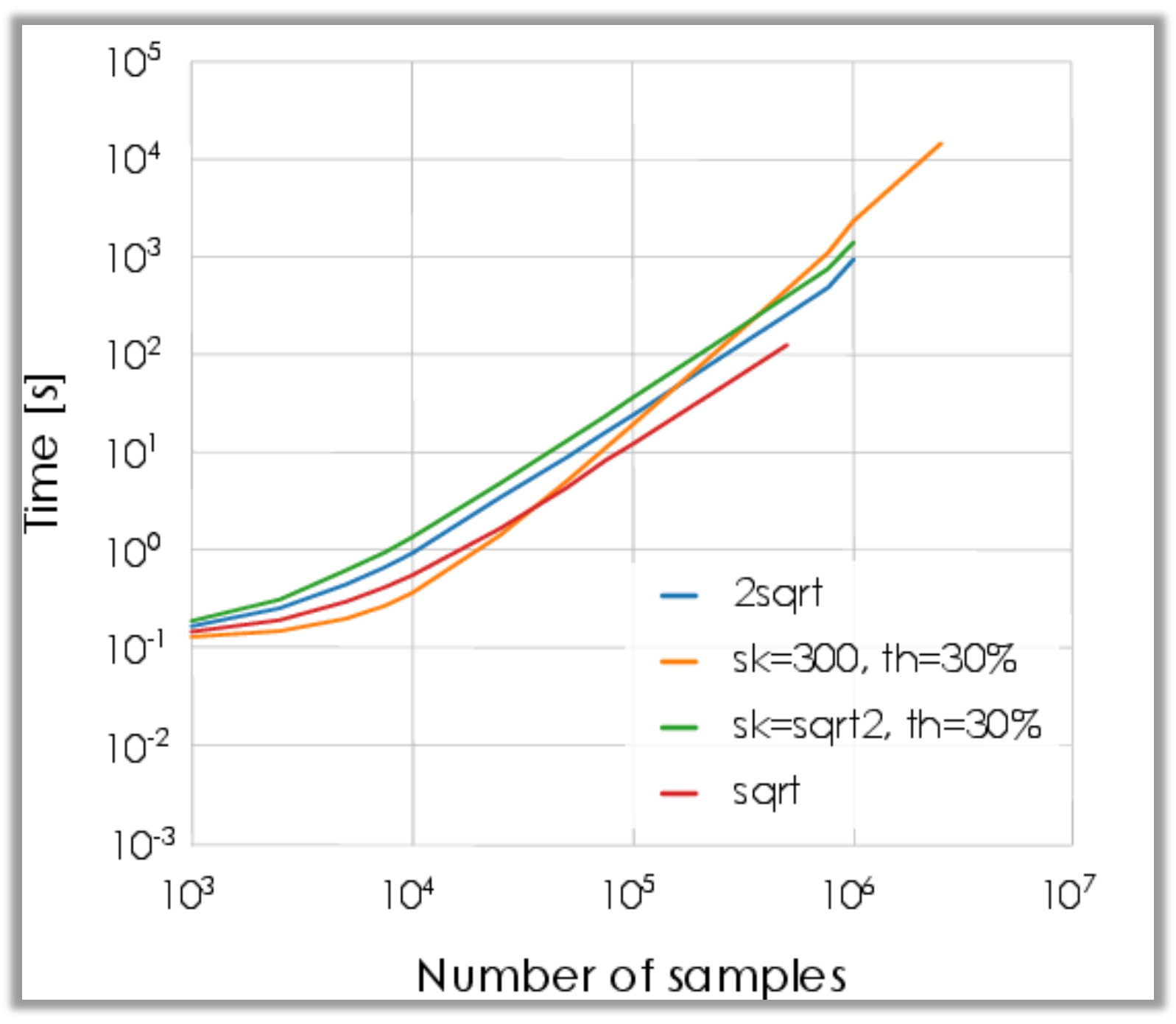
PRODUCTION OF ENSEMBLE

Challenge
Fast generation of ensemble.

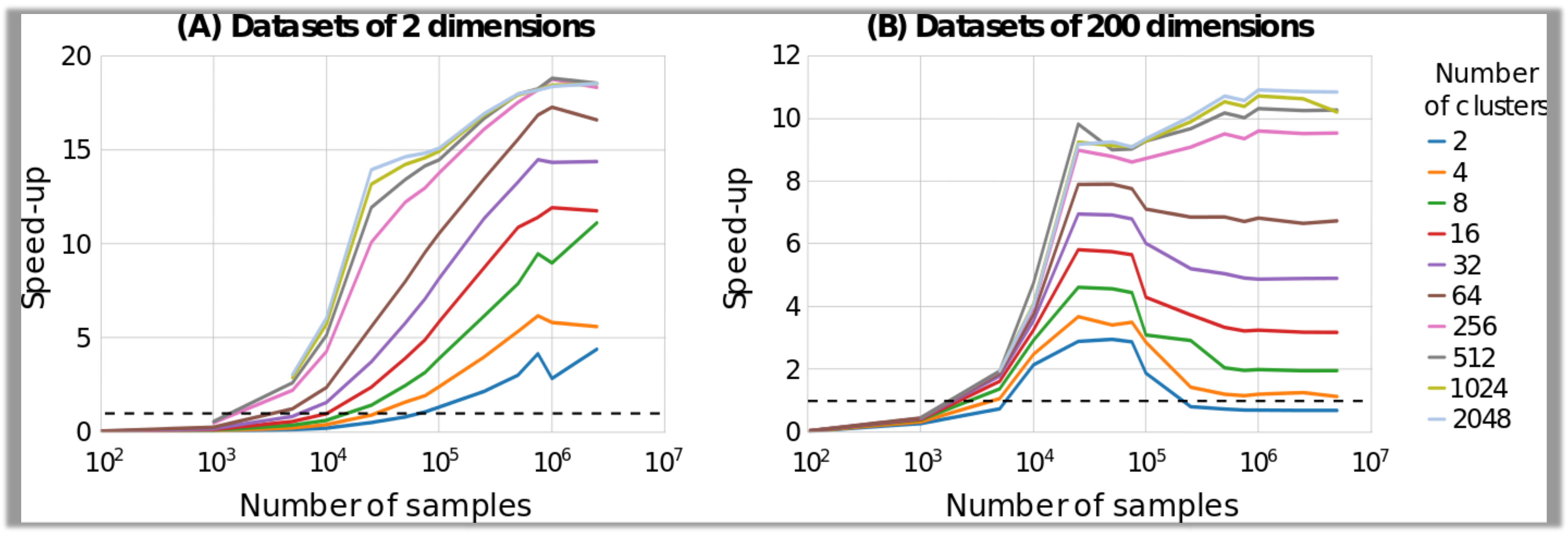
We used a 2-dimensional mixture of 6 Gaussians for most tests.

Solution
Parallel GPU K-Means

Production of the ensemble



GPU K-Means speed-up



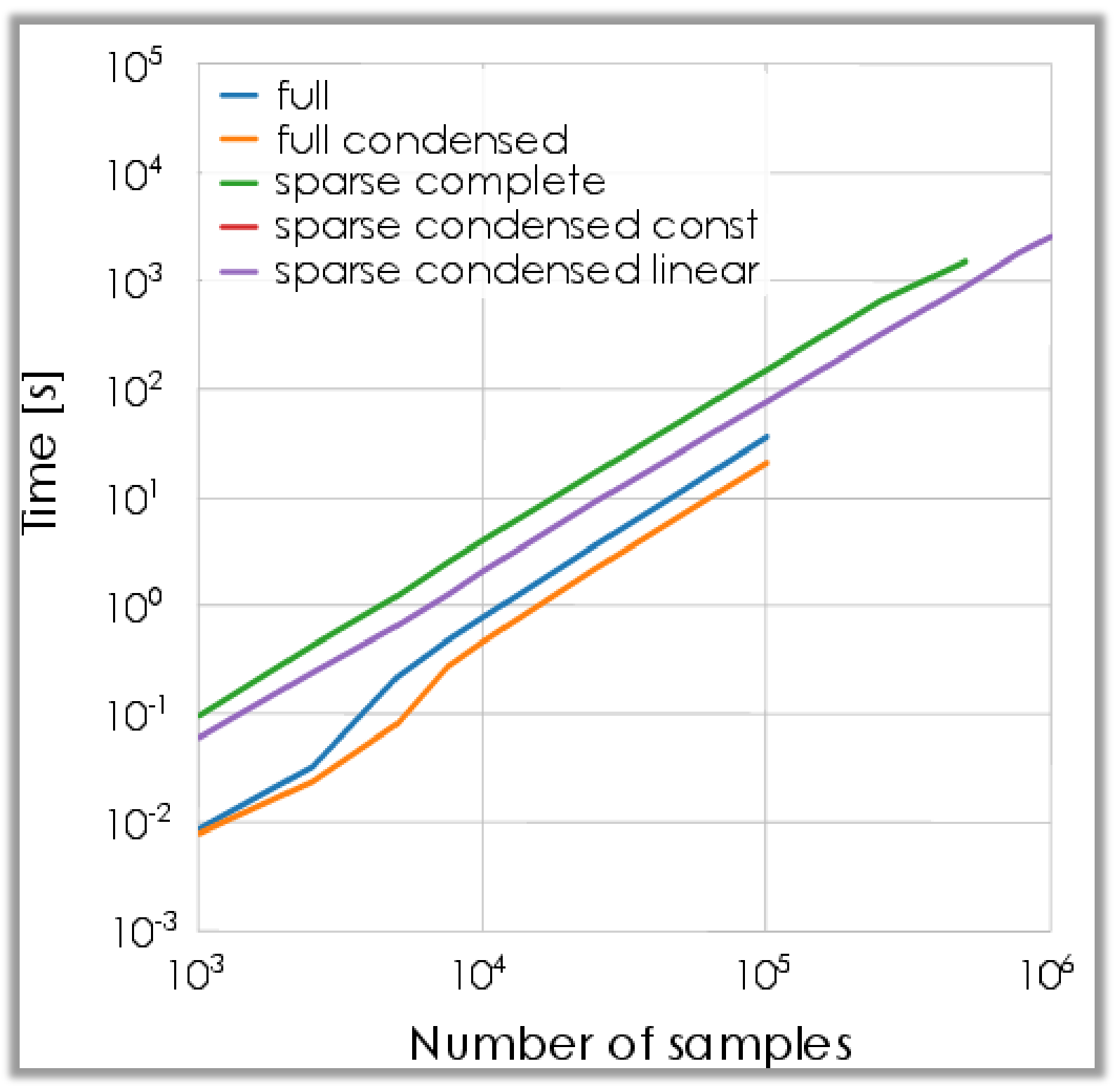
COMBINATION OF PARTITIONS

Challenge
O(n²) space complexity

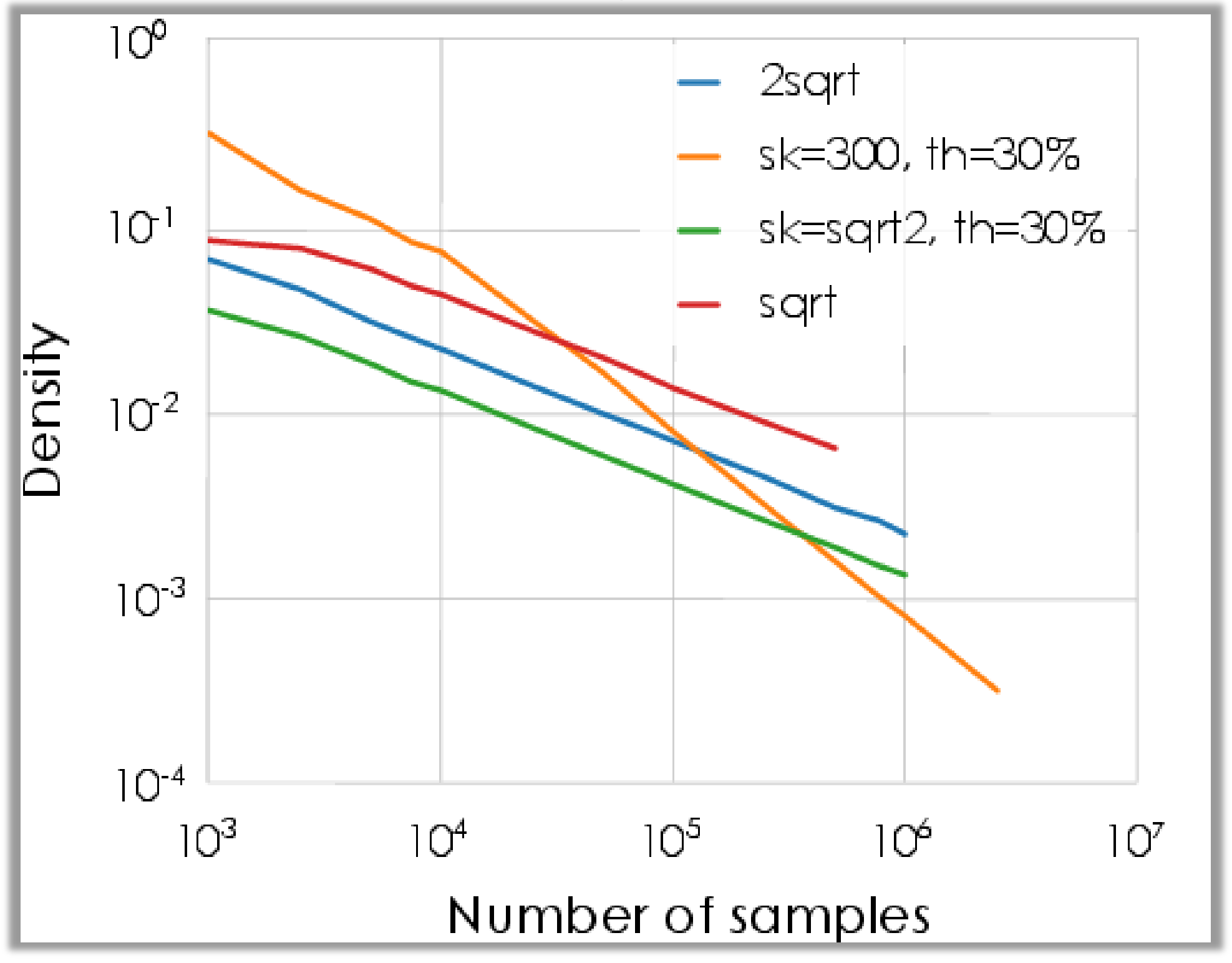
Upper triangular matrices are referred to as condensed.

Solution
CSR sparse matrix with optimized building

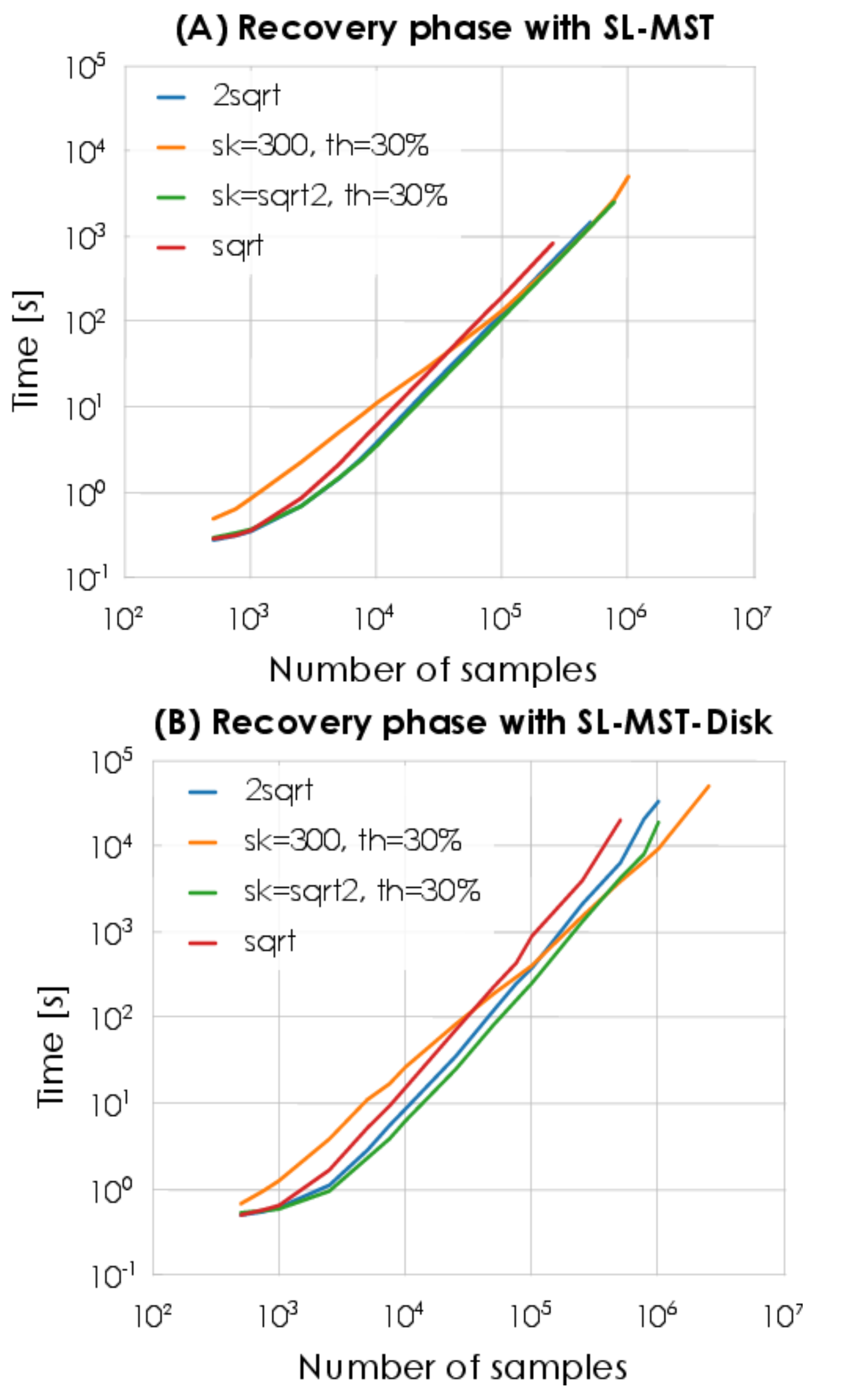
Building with different matrix formats



Density of associations relative to the full square matrix



Total time



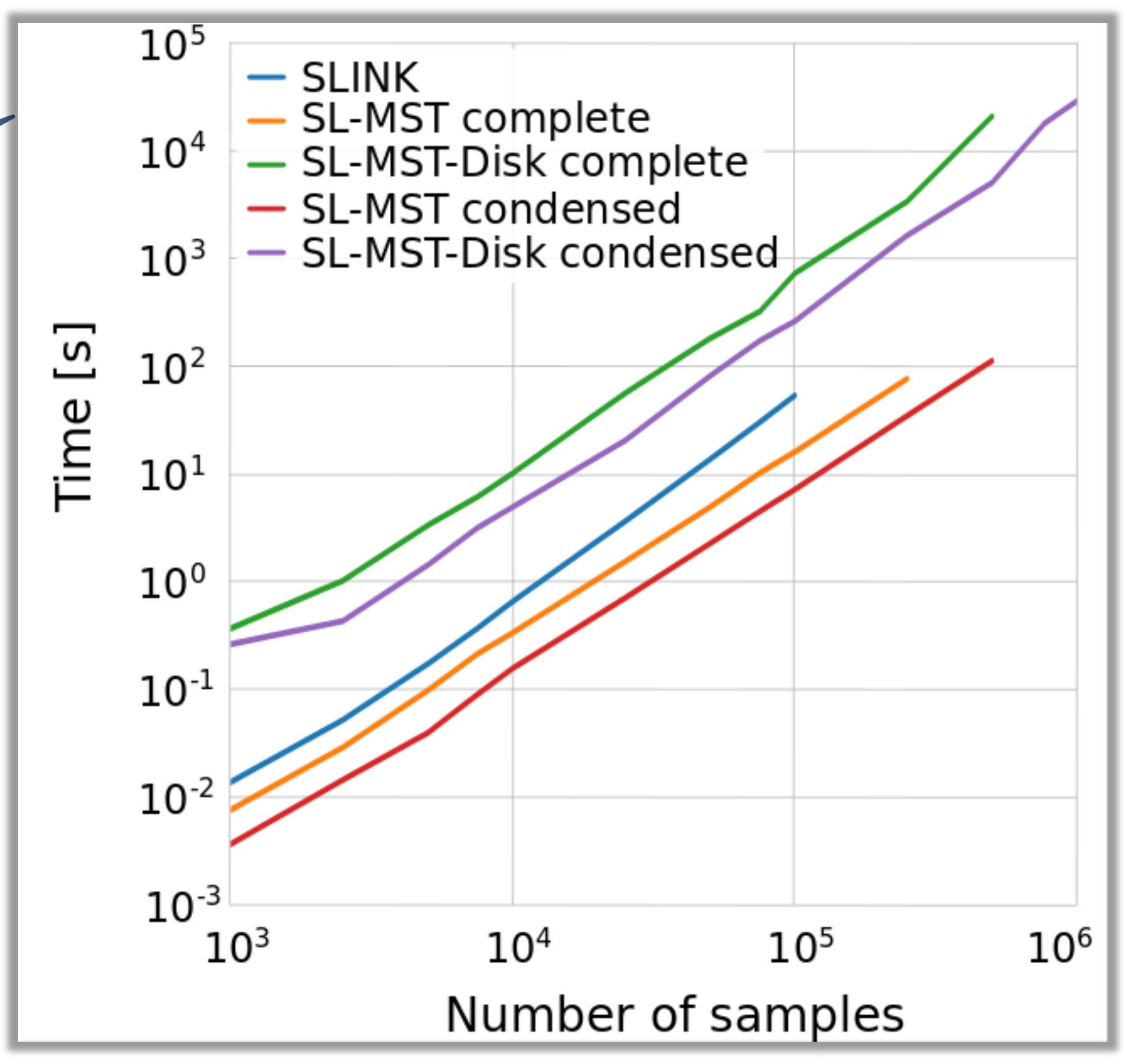
RECOVERY OF FINAL PARTITION

Challenge
O(n²) space complexity

SLINK is a fast implementation of SL that works over non-sparse matrices.

Solution
MST based SL
MST disk-based SL

Comparison of three methods for extraction



Conclusions

- EAC is now applicable to a wider spectrum of datasets – we clustered datasets of up to 10 times bigger what was before possible and the implementation supports bigger.
- Speed-up from 6 to 200 compared to original implementation on the different phases for small datasets.
- Better understanding of how ensemble rules affect the performance of the overall algorithm.