Learning to partition unbounded graph streams

Vasiliki Kalavri, Boston University
vkalavri@bu.edu

Collaboration with Michal Zwolak, Zainab Abbas, Sonia Horchidan, Paris Carbone
(KTH Royal Institute of Technology)

LDBC TUC, Aug 16

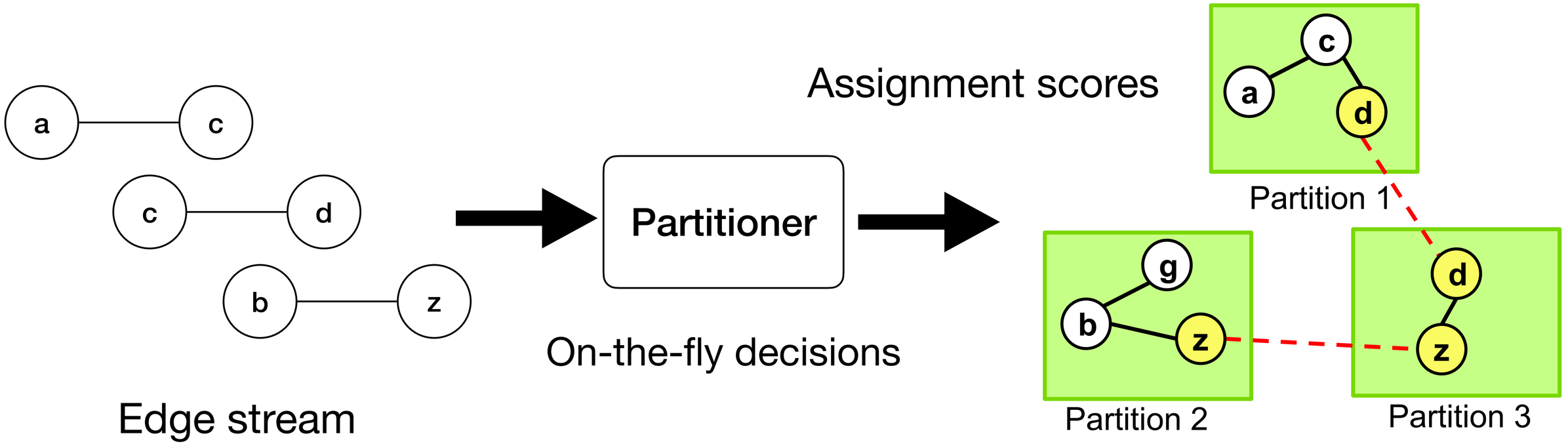
Graph streams

- Possibly unbounded sequences of timestamped relationships (edges)
- User interactions, financial transactions, driver-client locations in ridesharing services, etc.
- Continuously ingested from external, often distributed sources

Graph streams

- Possibly unbounded sequences of timestamped relationships (edges)
- User interactions, financial transactions, driver-client locations in ridesharing services, etc.
- Continuously ingested from external, often distributed sources

Streaming edge partitioning



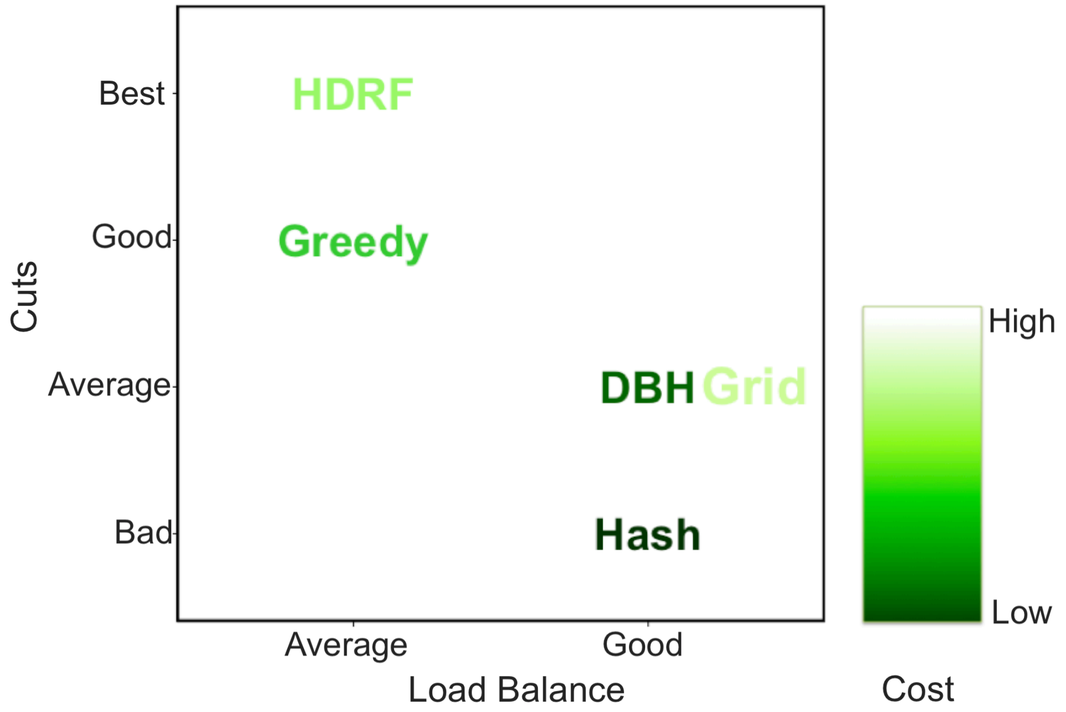
Balance number of edges per partition

Minimize number of replicated vertices

Stateful edge partitioning has better performance

But state can grow indefinitely for unbounded streams

- Current assignment of vertices to partitions needs to be stored
- The state needs to be queried and updated for every edge in the stream
- Difficult to support high-throughput streams with **global mutable state**

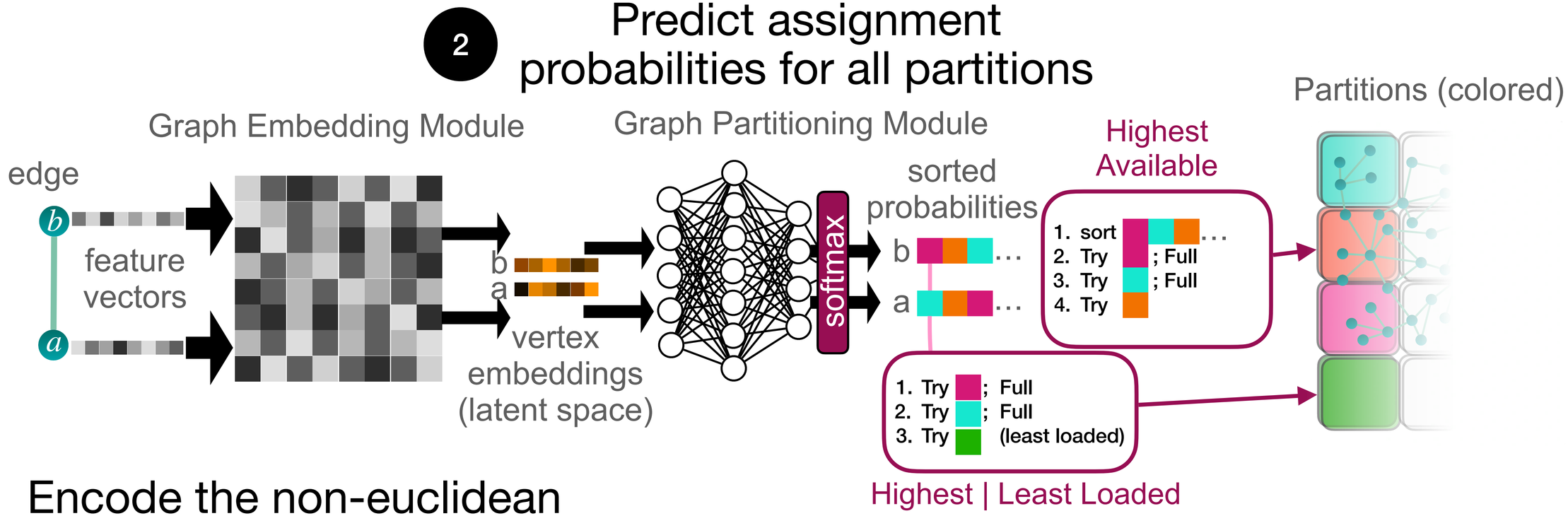


Abbas, Kalavri, Carbone, Vlassov. *Streaming graph partitioning: An experimental study*. (VLDB'18).

Can we partition unbounded graph streams with high quality and bounded state?

ML-added graph partitioning

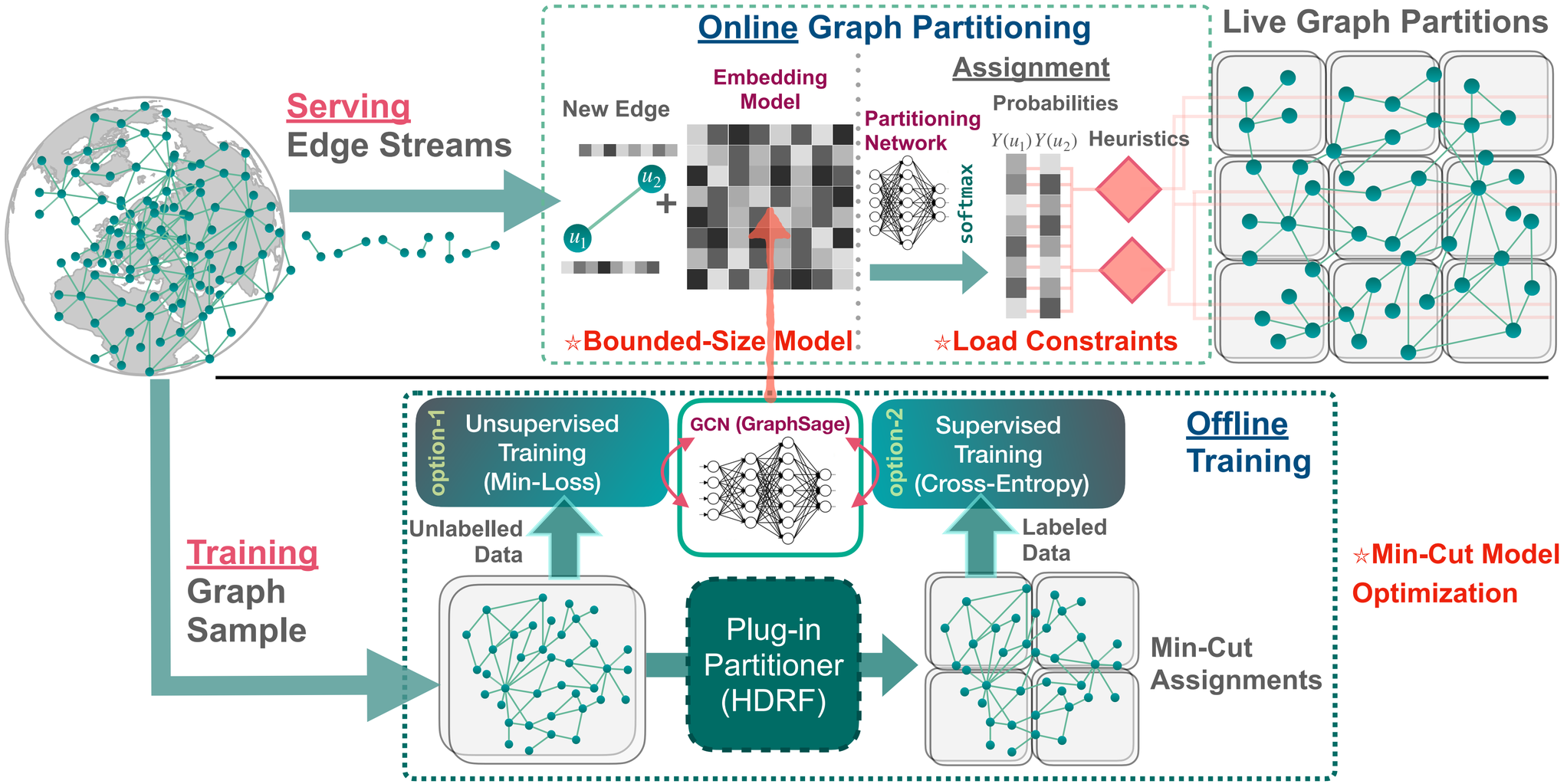
With graph representation learning



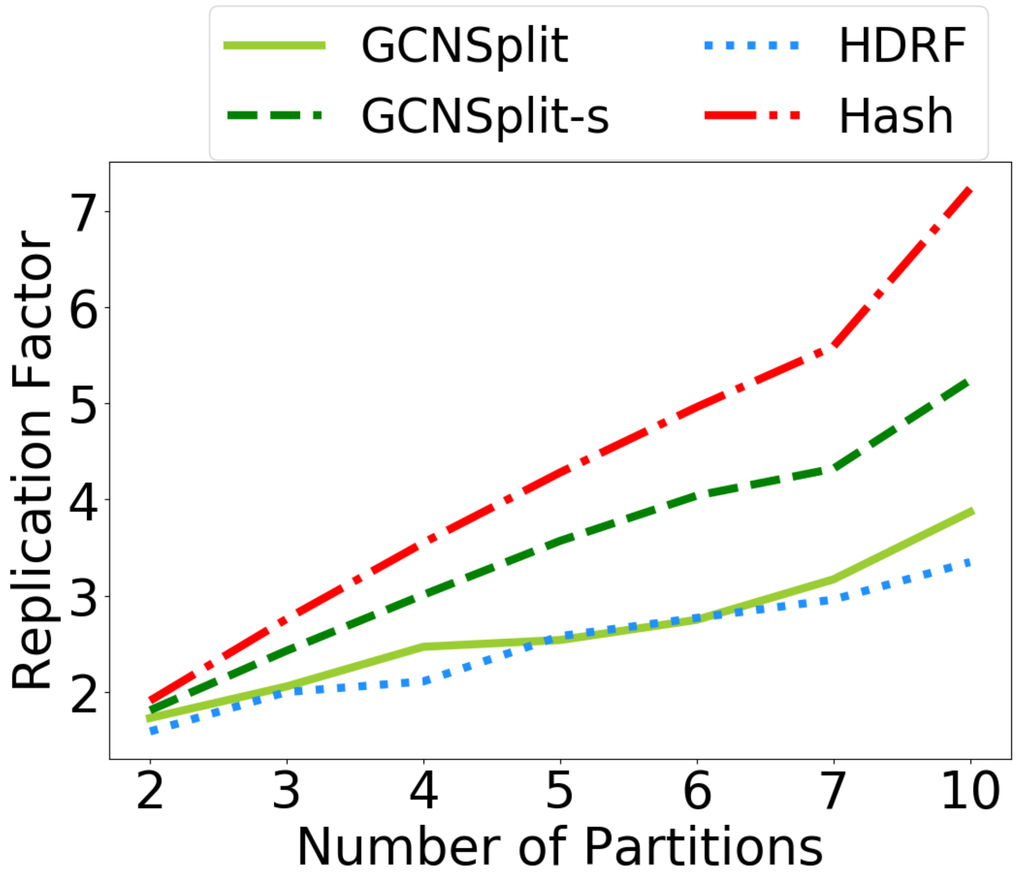
1 Encode the non-euclidean input graph stream into vectors of defined size in latent space

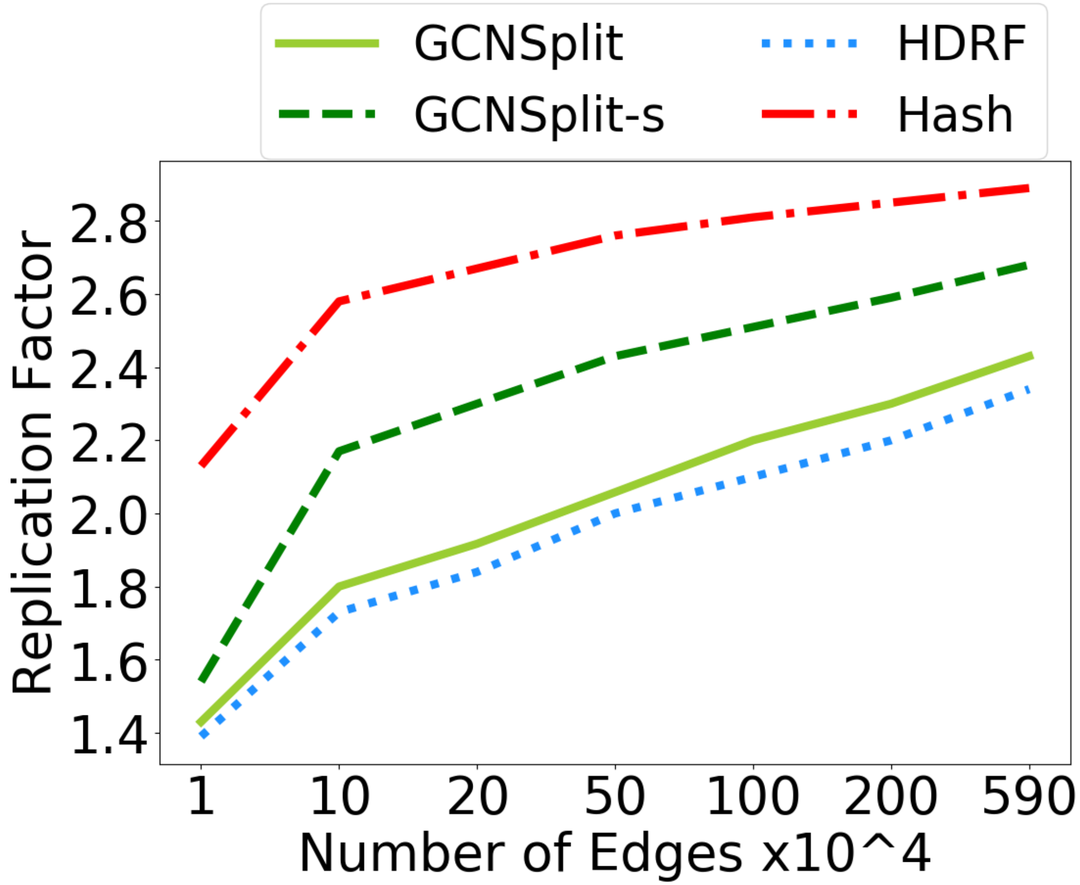
3 Apply assignment heuristics

Overview of GCNSplit



GCNSplit offers partitioning quality on par with stateful partitioning

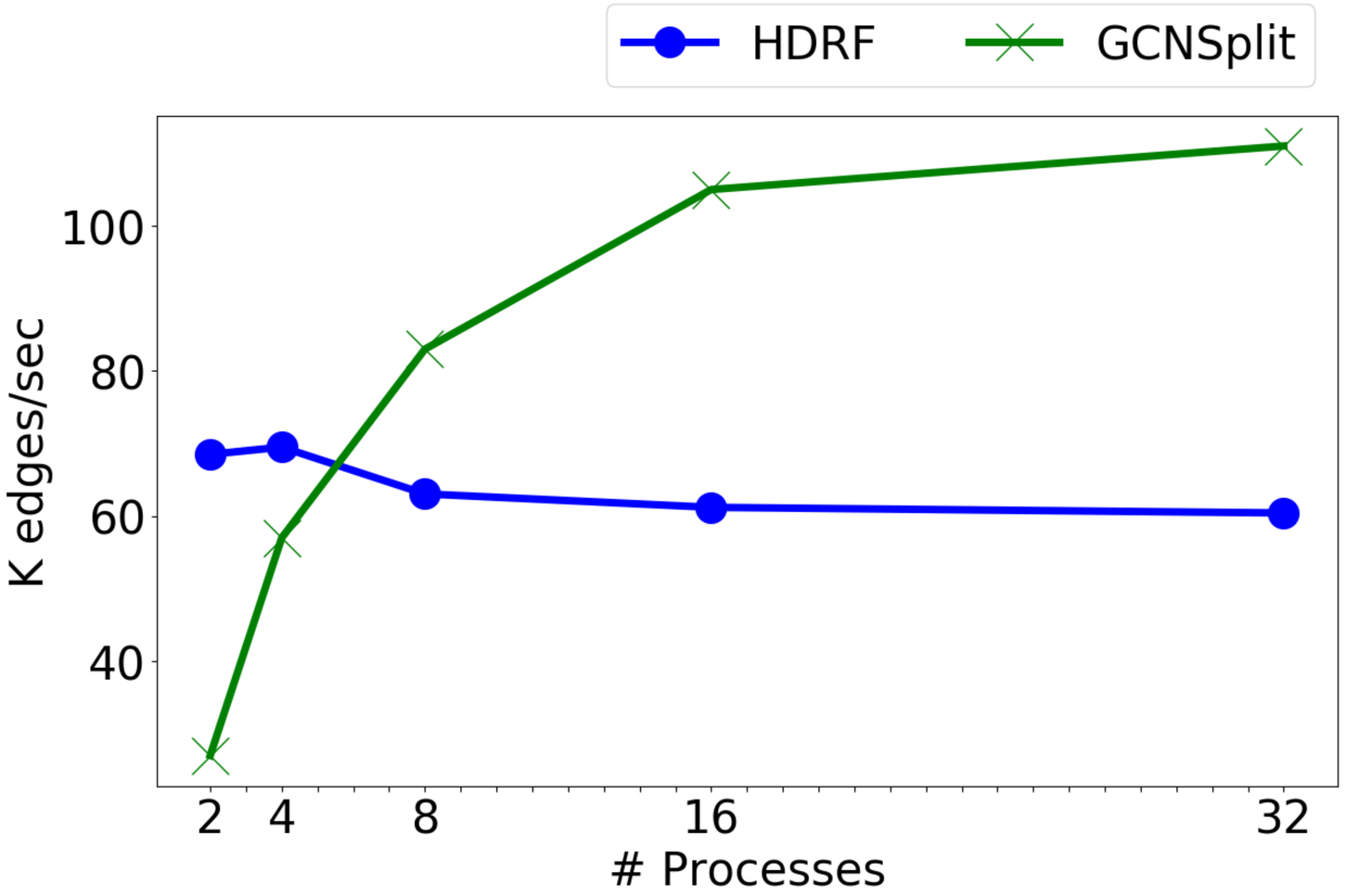




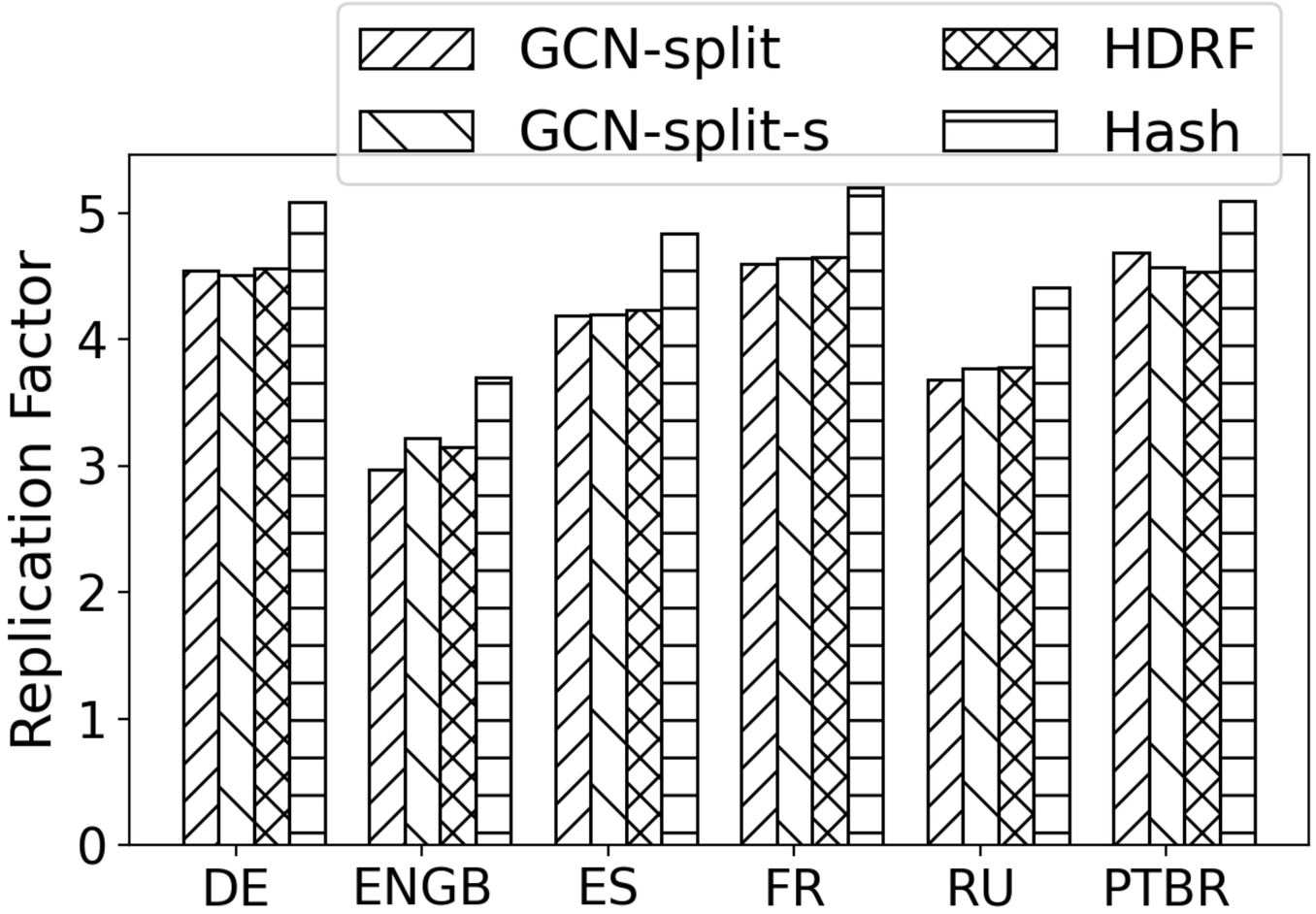
GCNSplit requires considerably smaller state

Dataset	Edges	GCNSplit state	HDRF state
Twitch	153K	1.6MB	4.1MB
Deezer	125K	126KB	5.4MB
Bitcoin	234K	166KB	19MB
Reddit	5.9M	385KB	47MB
Synthetic	**1.3B**	**115KB**	**>116GB**
Papers	**1.6B**	**147KB**	**>116GB**

GCNSplit can leverage parallelism to improve throughput



GCNSplit can generalize to unseen graph streams



- Twitch user-to-user networks speaking various languages
- Training on 10K edges sampled from the DE and RO networks

Limitations and future work

- Performance is highly dependent on the **quality of training data**
  - Rich feature sets lead to lower replication factor
  - High partitioning quality as long as the graph stream's characteristics do not change drastically
- In case of major concept drift GCNSplit behaves like **hash partitioning**
  - Constraints guarantee good load balance
  - Partitioning decisions equivalent to random assignment
- **Continual learning** methods can be used to update the model *incrementally*
  - Detect drift and use graph sampling to incorporate new knowledge while maintaining old one

Learning to partition unbounded graph streams

Vasiliki Kalavri, Boston University
vkalavri@bu.edu

Collaboration with Michal Zwolak, Zainab Abbas, Sonia Horchidan, Paris Carbone
(KTH Royal Institute of Technology)

LDBC TUC, Aug 16