Query-driven Node Selection in Distributed Learning Environments

Tahani Aladwani

Supervisor: Dr Christos Anagnostopoulos

University of Glasgow

School *of* Computing Science
Knowledge & Data Engineering Systems Group

Introduction

Distributed machine learning (DML)is a framework that aims to build a global model based on scattered data spread over a large number of nodes without gathering all the needed data in a central server. This kind of framework decreases the concerns about clients' data privacy, reduces the pressure on the central servers, and reduces communication costs. However, one of the key challenges in the DML is that we deal with heterogeneous nodes. Therefore, selecting unsuitable edge nodes can effect estimation. In addition, even in the most suitable nodes, the amounts of data could be significant, and not all of data samples could be relevant to improve the global models' performance.

Rationale

In this work, we investigate some values that can implicitly indicate the data distribution of the node. This could help predict superior nodes that accelerate and converge the global model accuracy. Naturally, some general information (such **as node ranking** (according to ***overlapping*** rate between a required analytic task and nodes that defines locally inside the node**, *the number of relevant data samples***)) is a good choice because sending this information to the server node cannot infer users' private information. Therefore, we conduct query-driven analysis to predict the impact of each available node on the global model. Then we choose the most appropriate set of nodes to participate in the training process [1].

Methods & Results

[Data Overlapping

Objective: Determine the **appropriateness** of a node for a query by exploring the node's available data overlapping between node's data and query.

Given a **query $\boldsymbol{q_k}$** and **node $\boldsymbol{n_k}$**, estimate the **percentage of data** (out of the whole **node's dataset $\boldsymbol{D_k}$**) required for executing that query.

Step 1: The node $\boldsymbol{n_k}$ quantizes its own data space $\boldsymbol{D_k}$, by e.g., adopting *k*-means into **K clusters**.

Step 2: Extract the **boundaries of each cluster** across all data dimensions, **the vector**., taking the minimum and maximum for each data dimension per cluster**, obtaining**

$$\boldsymbol{\mathcal{K} = [k_1^{min}, k_1^{max}, k_2^{min}, k_2^{max}, \dots, k_d^{min}, k_d^{max}]}$$

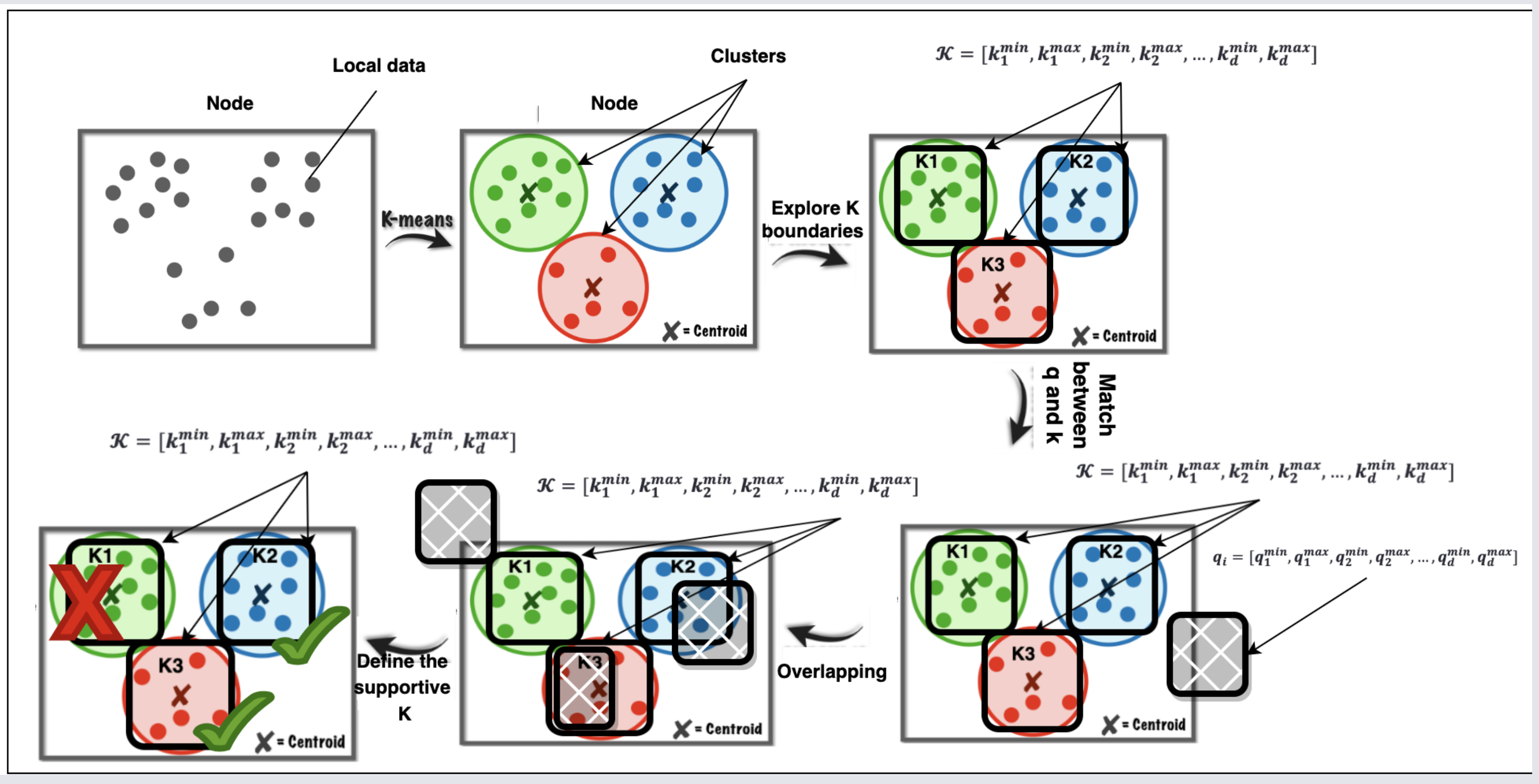
Step 3: Obtain the query $\boldsymbol{q_k}$ boundaries

$$\boldsymbol{q_k = [q_1^{min}, q_1^{max}, q_2^{min}, q_2^{max}, \dots, q_d^{min}, q_d^{max}]}$$

The **data overlapping $\boldsymbol{h_k}$ of cluster $\boldsymbol{k}$** is estimated through the overlapping of the intervals of each dimension in cluster *k* with these of query $\boldsymbol{q_k}$ at node $\boldsymbol{n_k}$.

Step 4: Define the **potential $\boldsymbol{p_k}$** for node $\boldsymbol{n_k}$ according to the number of **supporting clusters $\boldsymbol{K' \leq K}$** whose overlapping $h_k \geq \epsilon$, i.e.,

Step 5: The nodes **are sorted w.r.t. their rankings** $\{r_1, \dots, r_n\}$.

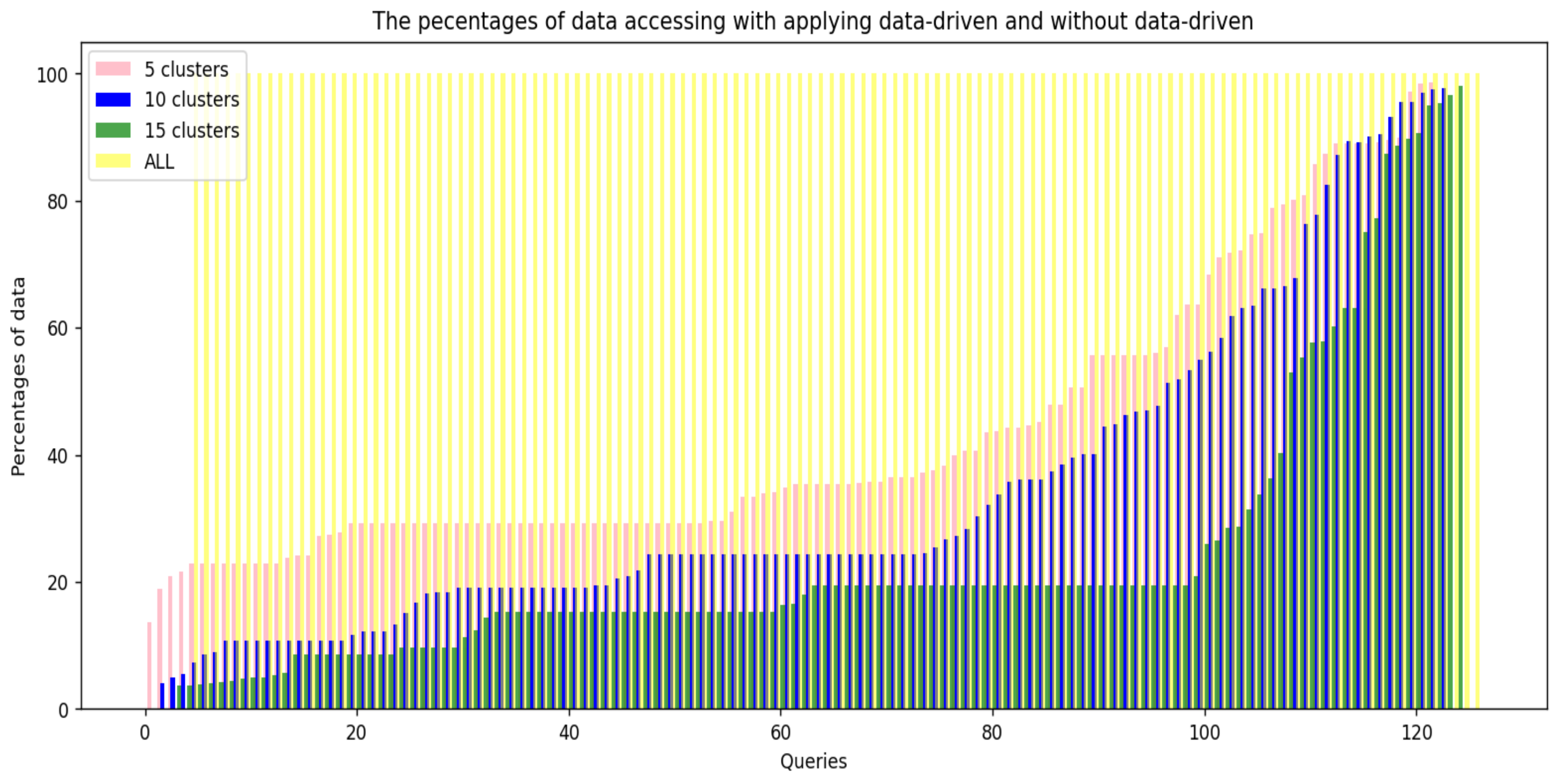

Local data
Node
Clusters
Node
K-means
Explore K boundaries
K1
K2
K3
= Centroid
$\mathcal{K} = [k_1^{min}, k_1^{max}, k_2^{min}, k_2^{max}, \dots, k_d^{min}, k_d^{max}]$
Match between q and k
$q_i = [q_1^{min}, q_1^{max}, q_2^{min}, q_2^{max}, \dots, q_d^{min}, q_d^{max}]$
Overlapping
Define the supportive K

Comparative Assessment

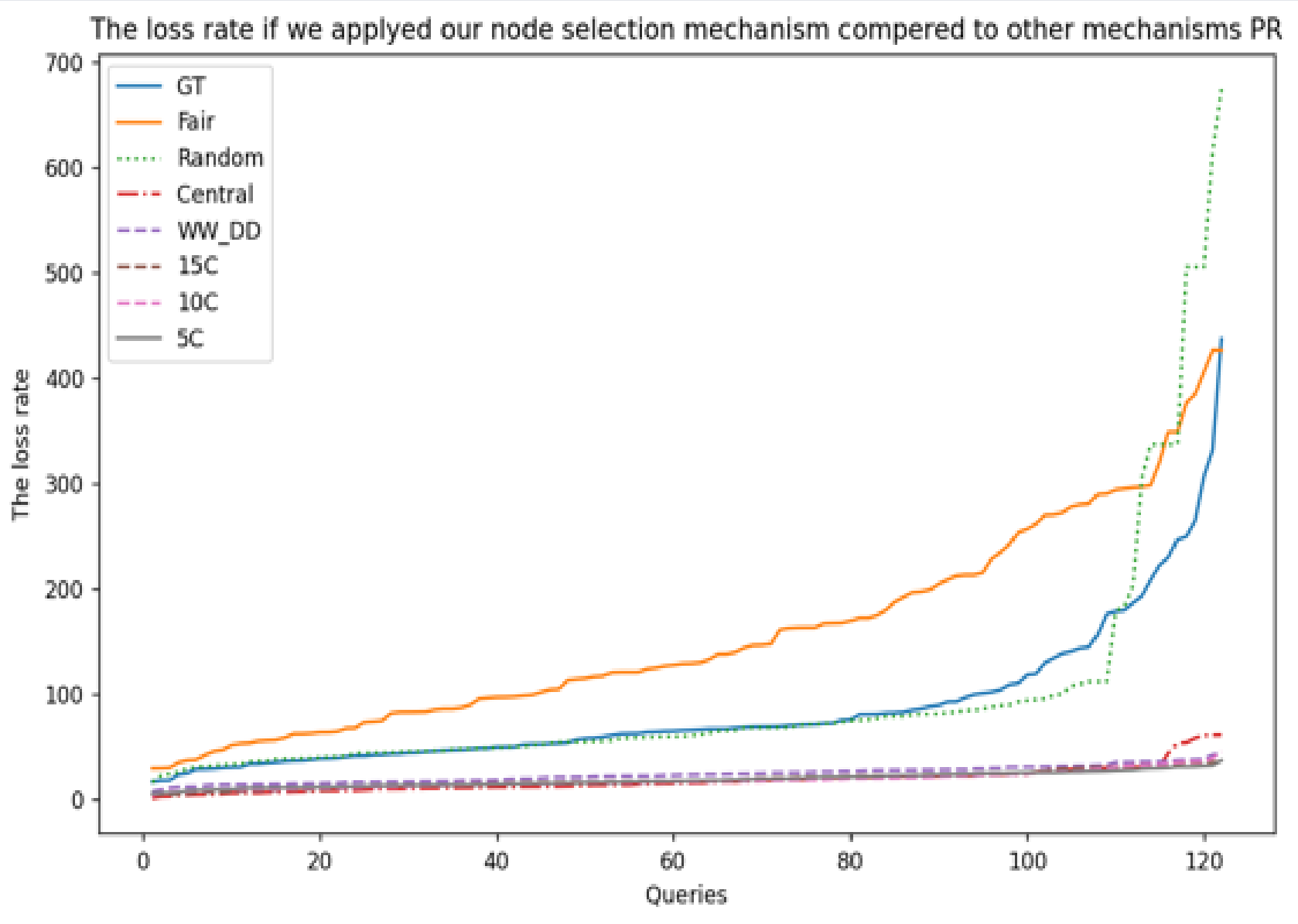
- **Random selection**: a node (or a subset of nodes) are randomly selected per query.
- **Game Theory (GT) selection mechanism**: nodes are selected based on their pre-trained models performances[2].
- **Fair selection:** First available set of nodes[2] .

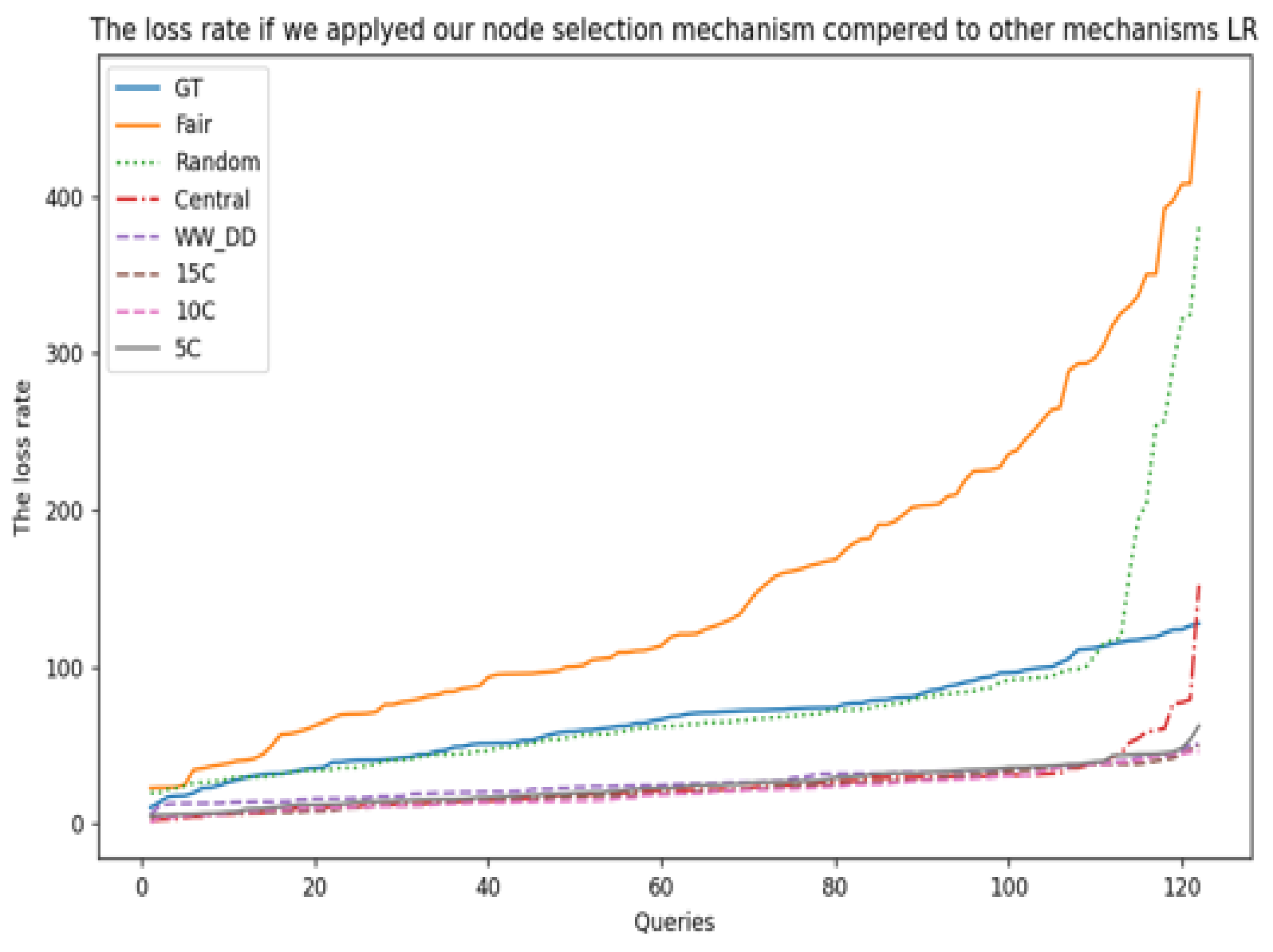
Indicating Results

[1]- Percentage/Amount of data accessed per query



[2]-ML Model Loss (over all queries)





Conclusions

We propose a **query-driven node selection mechanism** in distributed learning environments. We contributed with a mechanism **to determine the most appropriate subset of nodes** to be engaged in a model building **per analytics query**. Our mechanism **minimizes the data needed by each participant** to train the model **only** over the query-driven supporting clusters' data. Our experimental results and comparative assessment showcase that our selection mechanism is deemed appropriate in distributed edge learning environments.

References

1. Deng, Y., Lyu, F., Ren, J., Wu, H., Zhou, Y., Zhang, Y., Shen, X., 2021.Auction: Automated and quality-aware client selection framework for efficient federated learning. IEEE Transactions on Parallel and Distributed Systems 33, 1996–2009..
2. Hammoud, A., Mourad, A., Otrok, H., Dziong, Z., 2022. Data-driven federated autonomous driving, in: International Conference on Mobile Web and Intelligent Information Systems, Springer. pp. 79–90.
3. Savva, F., Anagnostopoulos, C., Triantafillou, P., Kolomvatsos, K., 2020. Large-scale data exploration using explanatory regression functions. ACM Transactions on Knowledge Discovery from Data (TKDD) 14,1–33