Fast distributed k-nn graph update

Thibault Debatty, Fabio Pulvirenti, Pietro Michiardi & Wim Mees
The context: Visual SPAM analysis

- Large, distributed k-NN graph
- E.g. Scalable k-NN based text clustering
  A Lulli, T Debatty M Dell'Amico, P Michiardi, L Ricci
- Subject similarity: Jaro–Winkler (not a metric)
The problem

How to efficiently add or remove nodes?

Naive algorithm:

- Add: $O(n)$ similarities
- Remove: $O(kn)$ similarities
Remove a node

- Use propagation to identify candidates $O[(k+1)^{\text{DEPTH}+1}]$
- Find new neighbor $O[k^{\text{DEPTH}+2}]$
Remove a node
Add a node

• Search neighbors of new node
  – Distributed graph based NN search
  – Graph partitioning:
    Distributed balanced k-medoids clustering

• Use propagation to update existing nodes
Sequential graph based NN search

- Hill climbing with restart
- Eager iteration
- Smart starting node selection
Distributed balanced k-medoids clustering

- Voronoï iteration
- Balanced: weight = 1 - size / capacity
- Distributed: randomized dataset
Add node

Correct edges and quality [%]

Synthetic - Correct edges
Synthetic - Quality
SPAM - Correct edges
SPAM - Quality

Number of nodes added to the graph [\cdot 1000]
Add node
Conclusions & future work...

• Fast add remove nodes

• Future:
  – Online algorithm and streaming framework
  – Simulated annealing based k-medoids clustering
Thank you!