THEORY QUESTIONS

1. [**2 points**] Indicate two features of the MapReduce framework, which make it suitable for developing big data applications.

2. [**3 points**] Briefly describe the agglomerative hierarchical clustering with single linkage, pointing out one advantage over center-based clusterings.

3. [**4 points**] Let T be a dataset of transactions over a set of items I and let $X \subseteq I$. Define the itemset Closure(X) and show that Supp(Closure(X)) = Supp(X).

4. [**4 points**] Let $G = (V, E)$ be an undirected graph with n nodes and $m = n^{1+c}$ edges, for some constant $c \in (0, 1]$. The MapReduce algorithm presented in class for computing a Minimum Spanning Forest (MSF) for G partitions E into ℓ subsets $E_1, \ldots, E_\ell$ of m/ℓ edges each, computes a MSF F_i independently in each E_i, and then computes the final MSF on the union of the F_i's. Indicate a suitable choice for ℓ determining the amount of local space required by the algorithm with the chosen value for ℓ. Justify your answer.

EXERCISES

1. [**7 points**] Let P be a set of N points in a metric space (M, d), and let $\mathcal{C} = (C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$ be a k-clustering of P. Design and analyze an efficient MapReduce algorithm that for each cluster center c_i determines the most distant point among those belonging to the cluster C_i. (Assume that all distances between pairs of points are distinct.) Initially, each point $q \in P$ is represented by a pair $(\mathrm{ID}(q), (q, c(q)))$, where $\mathrm{ID}(q)$ is a distinct key in $[0, N-1]$ and $c(q) \in \{c_1, \ldots, c_k\}$ is the center of the cluster of q. Specify map and reduce phases, and intermediate and output pairs of each round. To get full score, the algorithm must use $o(N)$ local space and linear aggregate space.

2. [**6 points**] Let T be a dataset of N transactions over a set of items I, and let $\epsilon > 0$ be a parameter. For each itemset X, let s_X be an approximation of its true support $\mathrm{Supp}_T(X)$ such that

$$\mathrm{Supp}_T(X) - \epsilon \le s_X \le \mathrm{Supp}_T(X) + \epsilon$$

Consider an ordering of all itemsets $X_1, X_2, X_3, \ldots$ such that $s_{X_1} \ge s_{X_2} \ge s_{X_3} \ldots$, and let $K < K'$ be two positive indices for which $s_{X_K} > s_{X_{K'}} + 2\epsilon$.

   (a) Show that for each pair of indices $i, j \ge 1$, with $i \le K < K' \le j$, we have $\mathrm{Supp}_T(X_i) > \mathrm{Supp}_T(X_j)$

   (b) Based on the previous points, show that the set $\{X_1, X_2, \ldots, X_{K'}\}$ contains the Top-K frequent itemsets with respect to the true support.

Total time: 2 hours