Exercises on Graph Analytics

Exercise. Let $G = (V, E)$ be an undirected graph with n nodes and $m = n^{1+c}$ edges, for some constant $c > 0$. Fix $M = n^{1+\epsilon}$, for some $\epsilon \in [0, c/2]$. The following MapReduce algorithm computes a Minimum Spanning Forest for G using $O(m)$ aggregate space and $O(M)$ local space. Let $m_0 = m$.

- **Round 1:** Partition E into m_0/M subsets of size M each, and compute a MSF separately for each subset. Let m_1 be the number of residual edges (i.e., edges that belong to the computed forests).
- **Round 2:** Partition the residual set of m_1 edges into m_1/M subsets of size M each, and compute a MSF separately for each subset. Let m_2 be the number of residual edges.
- ... Continue until the first **Round** i where m_{i-1} residual edges are partitioned into m_{i-1}/M subsets of size M each, a MSF is computed separately for each subset, and $m_i \leq M$ residual edges survive. Then, in the next round (**Round** $i+1$) compute the final MSF on these m_i residual edges.

Determine the number of rounds required by the algorithm as a function of ϵ and c.

Solution. Since a spanning forest of any subgraph of G has at most $n-1$ edges we have that

$$\begin{array}{lllllll} m_1 & \leq & \frac{m_0}{M}(n-1) & < & \frac{m_0}{M}n & = & n^{1+c-\epsilon} \\ m_2 & \leq & \frac{m_1}{M}(n-1) & < & \frac{m_1}{M}n & = & n^{1+c-2\epsilon} \\ \cdots & & & & & & \\ m_i & \leq & \frac{m_{i-1}}{M}(n-1) & < & \frac{m_{i-1}}{M}n & = & n^{1+c-i\epsilon}. \end{array}$$

Then, in order to have $m_i \leq M_i$ it is sufficient that $n^{1+c-i\epsilon} \leq M = n^{1+\epsilon}$, that is $i \geq c/\epsilon - 1$. Since i is an integer, we conclude that at the end of Round $\lceil c/\epsilon \rceil - 1$, there are at most M residual edges, hence the number of rounds is at most $\lceil c/\epsilon \rceil$. Note that if $\epsilon = c/2$, the result is consistent with the analysis of the 2-round algorithm done in class. □

Exercise. Consider a graph $G = (V, E)$ with n nodes and $m = n^{1+c}$ edges, for some constant $c > 0$. Suppose that the edge set E is randomly partitioned into $\ell = n^{c/2}$ subsets $E_1, E_2, \ldots, E_\ell$, where an edge e is assigned to a subset E_i with probability $1/\ell$ independently of the other edges. Show that with probability that tends to 1 as n goes to ∞, every subset E_i has size $O(m/\ell)$.

Hint: Use the Chernoff Bound that states that for a Binomial r.v. X with $E[X] = \mu$, $\Pr(X \geq 6\mu) \leq 2^{-6\mu}$.

Solution.
Consider an arbitrary subset E_i and let $X = |E_i|$. Since each edge is assigned to E_i with probability $1/\ell$ independently of the other edges, X can be regarded as the sum of m i.i.d. Bernoulli variables I_e, one for each edge e, where $I_e = 1$ if $e \in E_i$ and 0 otherwise, and $\Pr(I_e = 1) = 1/\ell$. Hence, X is a Binomial r.v. with expectation

$$\mu = \frac{m}{\ell} = n^{1+c/2}.$$

By the Chernoff Bound we have that

$$\Pr(X \geq 6n^{1+c/2}) \leq 2^{-6n^{1+c/2}}.$$

Since there are ℓ subsets, by the union bound, the probability that there exists a subset E_i with $\geq 6m/\ell = 6n^{1+c/2}$ edges is at most

$$\ell 2^{-6n^{1+c/2}} = n^{c/2} 2^{-6n^{1+c/2}},$$

which tends to 0 as n tends to ∞. Therefore, all subsets have size $< 6m/\ell$ with probability that tends to 1 as n goes to ∞.

N.B. *A similar argument was used in the analysis of Word Count 2. Check Slide 25 in the slides on MapReduce.* □

Exercise. Let $G = (V, E)$ be a connected, undirected graph with n nodes. For an arbitrary node $v \in V$, let $\Delta = \max_{u \in V} \text{dist}(v, u)$. Show that $\text{Diameter}(G) \in [\Delta, 2\Delta]$.

Solution. By definition,

$$\text{Diameter}(G) = \max_{x,y \in V} \text{dist}(x, y).$$

Since by the definition of Δ we know that there exists a pair of nodes at distance Δ, we have that $\text{Diameter}(G) \geq \Delta$. Let x and y be two nodes such that $\text{dist}(x, y) = \text{Diameter}(G)$. One can go from x to y passing through v, hence we have that

$$\text{Diameter}(G) = \text{dist}(x, y) \leq \text{dist}(x, v) + \text{dist}(v, y) \leq 2\Delta.$$

Thus,

$$\Delta \leq \text{Diameter}(G) \leq 2\Delta.$$

□

Exercise. Let $G = (V, E)$ be a connected, undirected graph with n nodes. Suppose that a BFS is executed from each of $\ell > 1$ distinct pivotal nodes $v_1, v_2, \ldots, v_\ell \in V$ and that the following two values are computed:

$$\begin{aligned} R &= \max_{u \in V} \min_{1 \leq i \leq \ell} \text{dist}(u, v_i) \\ \Delta &= \max_{1 \leq i,j \leq \ell} \text{dist}(v_i, v_j). \end{aligned}$$

Show that $\text{Diameter}(G) \in [\Delta, \Delta + 2R]$.

Solution. By definition,

$$\text{Diameter}(G) = \max_{x,y \in V} \text{dist}(x, y).$$

Since by the definition of Δ we know that there exists a pair of nodes at distance Δ, we have that $\text{Diameter}(G) \geq \Delta$. Let x and y be two nodes such that $\text{dist}(x, y) = \text{Diameter}(G)$, and let v_i be the pivotal node closest to x, and v_j the pivotal node closest to y. One can go from x to y passing first through v_i and then through v_j, hence we have that

$$\text{Diameter}(G) = \text{dist}(x, y) \leq \text{dist}(x, v_i) + \text{dist}(v_i, v_j) + \text{dist}(v_j, y) \leq 2R + \Delta.$$

Thus,

$$\Delta \leq \text{Diameter}(G) \leq 2R + \Delta.$$

$\square$