Questions on Clustering and Graph Mining

October 29, 2016

The following questions will be asked at the final exam. For each of them, the contribution to the grade of the final exam is indicated.

Question 1 (K-means++, 3/20 points) Consider the deterministic variant of the $k-$means++ algorithm where the set of initial centroids are selected in the following way. Let $\mathcal{X}$ be a set of points in $\mathbb{R}^d$ given in input. The first centroid is chosen arbitrarily in $\mathcal{X}$. Then at step t, given the set of centroids $C_t = c_1, \ldots, c_t$, $(1 \leq t \leq k-1)$ selected up to step t, we select the point at maximum distance from the centroids in C_t. Formally, we choose the point $p \in \mathcal{X}$ such that $\min_{c \in C_t} d(c, p)$ is maximum (if there are multiple choices we pick one of them arbitrarily). Give an example where such a variant of k-means++ does not perform well, i.e. it computes a set of centroids with SSE much larger than the optimum solution. Compare the performance of such a variant of k-means++ with the original k-means++ algorithm.

Question 2 (densest subgraphs, 3/20 points) Let $G = (V, E)$ be an undirected graph. Let $H_1 = (V_1, E_1)$ and $H_2 = (V_2, E_2)$ be two densest subgraphs in G, i.e., for any subgraph $H = (V_H, E_H)$ of G it holds that $\frac{|E(H)|}{|V(H)|} \leq \frac{|E_i|}{|V_i|}, i = 1, 2$. Let $\widehat{H} = (V_1 \cap V_2, E_1 \cap E_2)$ be the graph obtained by the intersection of H_1 and H_2. What can we say about the density of $\widehat{H}$?