TEL AVIV UNIVERSITY
Department of Computer Science
0368.3239 – Foundations of Data Mining
Fall Semester, 2017/2018

Homework 4, January 8, 2018

- **Due on January 22 23:59 IST. Each question has the same weight.**
- **Submission instructions: We are using `https://gradescope.com`. Gradescope entry code for the course is: MYVDZ5**
  Please prepare a PDF file with each problem starting on a new page. When uploading, you will need to indicate locations of each problem/section.
- **You may consult any sources or people but you must write and submit the solution yourself and state your collaborators.**

1. **Reverse-reachability with k-mins sketches.** For a node v in a graph $G = (V, E)$ denote the reverse reachability set of v by

$$\text{Reach}^{-1}(v) := \{u \in V \mid I_{u \rightsquigarrow v}\} \ .$$

a. (10 points) Describe an algorithm to compute, for every node $v \in V$ in the graph, a k-mins sketch of $\text{Reach}^{-1}(v)$.

b. (5 points) Explain how to use the sketches of two nodes u, v to obtain an estimate on $|\text{Reach}^{-1}(v) \cup \text{Reach}^{-1}(v)|$.

2. **Seed-set focused centrality.** The kernel density (centrality) value of a node v with respect to a subset of *seed* nodes $S \subset V$ and a decaying function α is defined as

$$C_{\alpha,S}(v) := \sum_{u \in S} \alpha(d_{vu}) \ .$$

a. (5 points) Describe an unbiased estimator for $C_{\alpha,S}(v)$ from $\text{ADS}(v)$ (assume that the ADS is with parameter k). (Here you get S only at query time. It is not available when you compute the sketches.) Hint: Use an unbiased estimator for the "presence", $I_{v \rightsquigarrow u}$ of each item u in $\text{ADS}(v)$.

b. (5 points) Assume that your estimators for $I_{v \rightsquigarrow u_1}$ and $I_{v \rightsquigarrow u_2}$ are independent for every $u_1 \neq u_2$ (They are not really independent but one can show that their covariances are non-positive).

I) What is the relation between the variance of your estimator for $C_{\alpha,S}(v)$ for some subset $S \not\subseteq V$ and your estimator for $C_{\alpha,V}(v)$. Explain.

II) In class we stated (without a proof) a bound of $1/\sqrt{2k-2}$ on the CV of the estimator for $C_{\alpha,V}(v)$, can such bound hold for $C_{\alpha,S}(v)$ for any $S \subseteq V$? Explain.

c. (5 points) Answer the following qualitatively. How does the relative error for your estimator for $C_{\alpha,S}(v)$ changes with I) the cardinality of the set S, and II) the distances of S from v relative to distances of other nodes from v? why?

d. (5 points) In the following, we assume that the set S is provided in advance, and can be used for computing the sketches.

Define ADS sketch structures of size $O(\epsilon^{-2})$ that allow us to estimate $C_{\alpha,S}(v)$ with CV ϵ. Outline how you compute these sketches efficiently and bound the computation of the algorithm by k, $|S|$, $|V| = n$ and m, the number of edges in the graph.

e. (5 points) Can you use these sketches to estimate with CV ϵ

$$\mathrm{INF}_{\alpha,S}(U) := \sum_{u \in S} \max_{v \in U} \alpha(d_{vu}) \ ?$$

Explain.

3. **Time-decay.** Consider a stream of events with timestamps $t_1 < t_2 < \cdots$. We are interested in maintaining a sketch of timestamps that would allow us to estimate the following two quantities at the current time t, with CV ϵ for any decay function α (as defined in class).

a. (10 points) Time-decaying sum with respect to start point:

$$T_\alpha = \sum_{i | t_i < t} \alpha(t_i)$$

b. (10 points) Time-decaying sum with respect to current time t:

$$T_\alpha = \sum_{i | t_i < t} \alpha(t - t_i)$$

For each of the two quantities describe the sketch, describe exactly how to update the sketch when a new time stamp arrives, and describe the estimator that uses the sketch.

4. **Submodularity.**

a. (10 points) Consider a weighted directed graph $G = (V, E, w)$ with a set of nodes V, a set of directed edges E, and positive weights $w : E \to R^+$. The directed cut function

$$f(S) := \sum_{(u,v) \in E | u \in S, v \in V \setminus S} w_{uv}$$

is defined for every $S \subseteq V$. Is f monotone? Is f submodular? prove.

b. (10 points) We have a weighted bipartite graph $G(V_1, V_2, E)$ where each edge in E is between a vertex $v_1 \in V_1$ and a vertex $v_2 \in V_2$. Each edge has a positive weight. We define a function f on subsets S of V_1 as follows. $f(S)$ is the sum over all vertices $u \in V_2$ of the weights of the largest two edges adjacent to u in the subgraph of G induced by S and V_2 (that is all edges between vertices of S and vertices of V_2.) Is f monotone? Is f submodular? prove.

5. **Greedy Bicriteria.** Consider a monotone submodular function and the greedy algorithm. Using the notation from the slides of lecture 11.

a. (10 points) What can you say about the quality of $f(G_{ck})$ for integer $c \geq 1$ compared with $f(\text{Opt}_k)$? Prove.

b. (10 points) Same question as in (a) but now we use the greedy algorithm with lazy updates and parameter ϵ to generate the sequence.