Midsem

Time : 180 minutes Full Marks :90

Submission Guidelines.

- The exam is open book and open notes but no electronic device can be used
- Keep your final answers brief and precise. Meaningless rambles fetch negative credits.

Problem 1. (25 points)

Let U be a universe with $|U| \geq 2$, and let $\mathcal{H}$ be a *universal family* of hash functions mapping U to $[m] = \{0, 1, \ldots, m-1\}$. That is, for all distinct $x, y \in U$,

$$\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m}.$$

Let $S \subseteq U$ be a set of m distinct elements. Choose $h \in \mathcal{H}$ uniformly at random and place each element of S into bin $h(x)$. Let X_b denote the number of elements hashed to bin b, and let $Y = \max_b X_b$ denote the maximum load.

a) **(10 points) Expected number of collisions.**
Define
$$T = \left|\{\{x, y\} \subseteq S : h(x) = h(y)\}\right|,$$
the total number of unordered colliding pairs. Show that
$$\mathbb{E}[T] \leq \frac{m-1}{2}.$$

Solution. There are $\binom{m}{2} = \frac{m(m-1)}{2}$ unordered pairs in S. By universality, for any fixed distinct x, y we have $\Pr[h(x) = h(y)] \leq 1/m$. Hence

$$\mathbb{E}[T] = \sum_{\{x,y\} \subset S} \mathbb{E}[I_{x,y}] = \sum_{\{x,y\} \subset S} \Pr[h(x) = h(y)] \leq \binom{m}{2} \cdot \frac{1}{m} = \frac{m-1}{2}.$$

b) **(5 points) Relating collisions to the maximum load.**
Explain why the number of collisions T is always at least the number of pairs in the fullest bin:
$$T \geq \binom{Y}{2}.$$

Hence, for any threshold $t > 0$,
$$\Pr[Y \geq t] \leq \Pr\left[T \geq \binom{t}{2}\right].$$

Solution. If some bin b contains X_b elements then the number of unordered colliding pairs inside that bin is $\binom{X_b}{2}$. Thus the fullest bin (one realizing Y) contributes $\binom{Y}{2}$ pairs, and therefore

$$T \;\geq\; \binom{Y}{2} \;=\; \frac{Y(Y-1)}{2}.$$

Consequently, for any threshold $t > 0$,

$$\Pr[Y \geq t] \;\leq\; \Pr\big[T \geq \binom{t}{2}\big],$$

because the event $\{Y \geq t\}$ implies $T \geq \binom{t}{2}$.

c) **(10 points) Bounding the load.**
Apply Markov's inequality to the random variable T and use your bound from part (a) to show that for $t = 1 + 2\sqrt{m}$,

$$\Pr[Y \geq 1 + 2\sqrt{m}] \leq \frac{1}{4}.$$

Conclude that, with probability at least 3/4, every bin contains at most $1 + 2\sqrt{m}$ elements.

Solution. Apply Markov's inequality to the nonnegative random variable T:

$$\Pr\big[T \geq \binom{t}{2}\big] \;\leq\; \frac{\mathbb{E}[T]}{\binom{t}{2}}.$$

Using the bound from part (a), $\mathbb{E}[T] \leq (m-1)/2$, we get

$$\Pr[Y \geq t] \;\leq\; \frac{(m-1)/2}{\binom{t}{2}}.$$

Now choose $t = 1 + 2\sqrt{m}$. Then $t - 1 = 2\sqrt{m}$ and

$$\binom{t}{2} \;=\; \frac{t(t-1)}{2} \;=\; \frac{(1+2\sqrt{m})(2\sqrt{m})}{2} \;=\; \sqrt{m} + 2m.$$

Therefore

$$\Pr\big[Y \geq 1 + 2\sqrt{m}\big] \;\leq\; \frac{(m-1)/2}{\sqrt{m} + 2m} \;\leq\; \frac{m/2}{2m} \;=\; \frac{1}{4},$$

where the penultimate inequality used $\sqrt{m} \geq 0$ to lower bound the denominator by $2m$. Hence with probability at least 3/4 we have

$$Y \;<\; 1 + 2\sqrt{m},$$

i.e. every bin contains at most $1 + 2\sqrt{m}$ elements.

This completes the proof.

Problem 2. **(5 x 2 = 10 points)** Let $G = (V, E)$ be a directed network with source s, sink t, and capacity function $c : E \to \mathbb{R}_+$. Let f be a feasible flow and G_f denote its residual graph.

(a) Prove or disprove: If every edge leaving s has zero residual capacity in G (not G_f), then the flow f is maximum. Justify your answer.

Solution. The statement is *false*. We exhibit a counterexample.

Consider the directed graph with vertices s, a, b, t and unit capacities on the edges

$$s \to a, \qquad a \to b, \qquad b \to s, \qquad b \to t.$$

Define a feasible flow f by sending one unit of flow around the directed cycle $s \to a \to b \to s$:

$$f(s,a) = f(a,b) = f(b,s) = 1, \qquad \text{and } f(b,t) = 0.$$

The only original edge leaving s is $s \to a$, and it is saturated by f. Hence every edge leaving s in the original graph has zero forward residual capacity.

However, f is *not* a maximum $s-t$ flow. In the residual graph G_f, the reverse edge $s \to b$ exists with residual capacity 1 (coming from the flow on $b \to s$), and the forward edge $b \to t$ has residual capacity 1. Thus G_f contains an augmenting path $s \to b \to t$, so the flow value can be increased.

Therefore the condition that every *forward original* edge leaving s is saturated does *not* imply that the flow is maximum. The correct sufficient condition would be that *no* residual edge (forward or backward) leaves s in the residual graph G_f.

(b) Suppose two distinct maximum flows f_1 and f_2 exist in G. Is it true that for every s–t cut (S,T),

$$\sum_{u \in S,\, v \in T} f_1(u,v) \;=\; \sum_{u \in S,\, v \in T} f_2(u,v)?$$

Give a short justification or counterexample.

Solution. The statement is *false*. Use the same graph as in part (a): vertices s, a, b, t with unit capacities on $b \to s$ and $b \to t$ and capacity 2 on $s \to a, a \to b$.

Flow f_1. Send one unit along $s \to a \to b \to t$:

$$f_1(s,a) = 1, \quad f_1(a,b) = 1, \quad f_1(b,t) = 1,$$

This is feasible and $|f_1| = 1$.

Flow f_2. Add the unit circulation along the cycle $s \to a \to b \to s$ to f_1:

$$f_2(s,a) = 2, \quad f_2(a,b) = 2, \quad f_2(b,s) = 1, \quad f_2(b,t) = 1.$$

All edge flows respect the capacities above, and $|f_2| = 1$ (the circulation does not change the net flow to t).

Compare the forward flow across the cut $S = \{s\}$, $T = \{a, b, t\}$:

$$\sum_{u \in S, v \in T} f_1(u,v) = f_1(s,a) = 1, \qquad \sum_{u \in S, v \in T} f_2(u,v) = f_2(s,a) = 2.$$

The backward flow across the cut is 0 for f_1 and $f_2(b,s) = 1$ for f_2, so the *net* flow across the cut equals 1 for both flows. Thus the forward sums across the cut can differ between maximum flows.

Problem 3. (**25 points**) A graph is said to be *d-regular* if every vertex has exactly d incident edges.

(a) (**10 points**) Let $G = (A \cup B, E)$ be a d-regular bipartite graph, where A and B are the two vertex classes. Show, using **Hall's Theorem**, that G contains at least one perfect matching.

Solution. Let $G = (A \cup B, E)$ be a d-regular bipartite graph. Summing degrees over A gives $d|A|$ edges, and summing degrees over B gives $d|B|$ edges, so $d|A| = d|B|$ and hence $|A| = |B|$.

Take any subset $S \subseteq A$ and let $N(S)$ be its neighbourhood in B. The number of edges leaving S equals $d|S|$, since every vertex in S has degree d. All these edges end in $N(S)$, and each vertex in $N(S)$ has degree d, so the number of such edges is at most $d|N(S)|$. Thus

$$d|S| \le d|N(S)|, \qquad \text{so} \qquad |S| \le |N(S)|.$$

Therefore Hall's condition holds for all $S \subseteq A$, and G has a matching saturating A. Since $|A| = |B|$, this matching is perfect.

(b) (**10 points**) Prove using part (a) and induction on d that every d-regular bipartite graph can be written as the union of d edge-disjoint perfect matchings.

Solution. Proceed by induction on d. For $d = 1$ the graph is itself a perfect matching.

Let the claim hold for $d-1$. Let G be d-regular and bipartite. By part (a), G contains a perfect matching M. Removing M reduces every vertex degree by exactly 1, so $G-M$ is $(d-1)$-regular and bipartite. By the induction hypothesis, $G - M$ decomposes into $d - 1$ edge-disjoint perfect matchings. Adding M completes a decomposition of G into d perfect matchings.

(c) (**5 points**) Does the same statement hold for d-regular *non-bipartite* graphs? Either prove it or give a counterexample with justification.

Solution. The statement does *not* hold for non-bipartite graphs.

A simple counterexample is the triangle C_3, which is 2-regular on 3 vertices. Since it has an odd number of vertices, it has no perfect matching at all, and therefore cannot be decomposed into two perfect matchings. Thus the bipartiteness assumption is essential.

Problem 4. **(30 points)**

Let $G = (V, E)$ be a directed flow network with source s, sink t, and **unit capacities** on all edges. Recall from Homework 2 that Dinic's algorithm proceeds in phases: in each phase it builds the level graph (by BFS from s) and augments a **blocking flow** in that level graph. In this problem, we are going to prove that Dinic's algorithm takes time $O(m^{3/2})$ on graphs with unit capacity.

(a) **(10 points)** Prove that if the distance between s and t in the residual graph is at least L, then the value of the maximum residual flow is at most $\dfrac{m}{L}$, where $m = |E|$.
Hint: Flow decomposition, use unit capacity crucially

Solution. Assume the distance from s to t in the residual graph is at least L. Let F be the value of the maximum residual $s-t$ flow. Because capacities are integral (unit) there exists an integral maximum flow, and any integral flow of value F can be decomposed into F simple $s-t$

paths (no edge is used by more than one such path, since every edge has capacity 1). Each of these paths has length at least L (number of edges $\geq L$). Since the paths are edge-disjoint, the total number of used edges is at least $F \cdot L$, but there are only m edges in the graph. Hence

$$F \cdot L \leq m,$$

so $F \leq \frac{m}{L}$, as required.

(b) **(10 points)** Using the above facts, show that after p blocking-flow phases the remaining residual flow is at most $m/(p+1)$. Conclude that the total number of blocking-flow phases is at most

$$p + \frac{m}{p+1}.$$

Hint: What happens to s-t distance after augmenting by one blocking flow ?

Solution. Let the algorithm perform p blocking-flow phases. Recall the property of Dinic's phases: after each blocking-flow augmentation the distance (length of shortest s–t path) in the residual graph strictly increases. More precisely, if the shortest-path distance before a blocking flow is d, then after augmenting a blocking flow the new shortest-path distance is at least $d+1$. (Reason: a blocking flow saturates every edge on all shortest $s-t$ paths in the current level graph, so no shortest path of the same length remains.)

Thus, after p blocking-flow phases the shortest s–t distance is at least $p+1$. Applying part (a) with $L = p+1$, the remaining residual flow value after p phases is at most

$$\frac{m}{p+1}.$$

Consequently, to eliminate all remaining flow we need at most an additional $m/(p+1)$ blocking-flow phases (each phase reduces the residual flow by at least 1). Therefore the total number of blocking-flow phases required is at most

$$p \; + \; \frac{m}{p+1}.$$

(c) **(5 points)** Choose a value of p to minimize this expression and deduce that Dinic's algorithm runs in $O(m^{3/2})$ time on unit-capacity graphs.

Solution. Choose p to (approximately) minimize the bound $p + \frac{m}{p+1}$. Treat p as a real variable and differentiate:

$$\frac{d}{dp}\left(p + \frac{m}{p+1}\right) = 1 - \frac{m}{(p+1)^2}.$$

Setting the derivative to zero gives $(p+1)^2 = m$, so $p \approx \sqrt{m} - 1$. Plugging $p = \lfloor \sqrt{m} \rfloor$ (or simply using the estimate) yields

$$p + \frac{m}{p+1} = O(\sqrt{m}).$$

Each blocking-flow phase in a unit-capacity graph can be implemented in $O(m)$ time (for example by the standard DFS-based blocking-flow routine on the level graph that always runs in linear time in the number of edges of the level graph; with unit capacities the total work per phase is $O(m)$). Multiplying phases by per-phase cost gives total running time

$$O(m) \cdot O(\sqrt{m}) = O(m^{3/2}).$$

Hence Dinic's algorithm runs in $O(m^{3/2})$ time on unit-capacity graphs.

(d) **(5 points)** Use some of the ideas above to design an algorithm for finding the maximum cardinality bipartite matching that runs in time $O(m\sqrt{n})$.

Solution. We are going to use the standard reduction of maximimum bipartite matching to the maximum-flow problem done in lectures. Now let us re-prove part (a) above with a tighter bound. Suppose the shortest path distance in the residual graph between s and t is L. We claim that then the maximum flow in the residual graph is n/L (in contrast to m/L proved above).

To see this, consider the residual graph and construct the BFS-layered graph done for blocking flows. Now notice that, at any iteration the incoming flow to any vertex on the left partite set is either 0 or 1. Consequently, if we take the union of all s-t paths in the residual graph, either the in-degree or the out degree of every vertex (other than s and t) is at most 1. Hence, for every layer in the BFS-layered graph, either the total incoming or outgoing flow is upper bounded by the number of vertices in that layer. Hence, the total flow that can be pushed from s to t is upper bounded by the minimum size of a layer - which is n/L by Pigeon Hole Principle. The rest of the analysis is the same as (a) and (b) with n replacing m everywhere except the time to compute a blocking flow. Hence, the runtime is $O(m\sqrt{n})$.