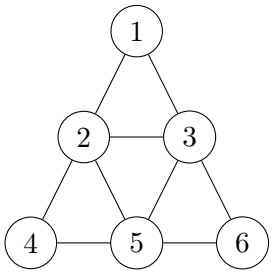
COMP 5711: Advanced Algorithms
Fall 2020 Final Exam (Part 2)

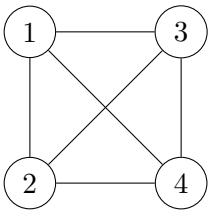
Note 1: $[n] = \{1, 2, \ldots, n\}$.

Note 2: For questions with multiple steps, if you don't know how to prove earlier steps, you can still try to prove later steps, assuming earlier steps have been proven.

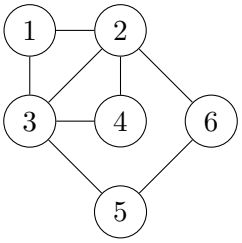
1. (9 pts) Draw a tree decomposition for each of the following graphs with minimal tree width.



(a)



(b)



(c)

2. (20 pts) In this problem, we will construct an LSH family for Euclidean distance, for the restricted domain of all points on the unit circle in the plane. Let $H = \{h_\alpha\}$ where $\alpha \in [0, \pi/2]$. (Yes, this hash family consists of infinitely many hash functions.) Note that each point on the unit circle corresponds to a unit vector. The hash function $h_\alpha(u)$ is defined as follows. For any point (unit vector) u, we rotate u counterclockwise by an angle of α. Let $f_\alpha(u)$ be the point after the rotation. Then we define $h_\alpha(u)$ as the quadrant that contains $f_\alpha(u)$, i.e., $h_\alpha(u)$ outputs a number in $\{1, 2, 3, 4\}$ indicating the quadrant. (What if $f_\alpha(u)$ falls on an axis? Well, this happens with probability 0.) This LSH family is (r, cr, p_1, p_2)-sensitive for $r = 1, c = \sqrt{2}$. Find the values of p_1 and p_2. You should make p_1 (resp. p_2) as large (resp. small) as possible. What if we change the range of α to $[0, 2\pi/3]$? Will the values of p_1 and p_2 be affected?

   The following facts may be useful:

   - Let θ be the angle between u and v. Then $d(u, v) = 2\sin(\theta/2)$. See Figure 2.
   - $\sin(\pi/2) = 1$, $\sin(\pi/3) = \sqrt{3}/2$, $\sin(\pi/4) = \sqrt{2}/2$, $\sin(\pi/6) = 1/2$, $\sin(0) = 0$.

3. (25 pts) The distinct count problem in the streaming model is defined as follows. Given a stream of elements, possibly with duplicates, approximately count the number of distinct elements using small memory. A simple algorithm is as follows. Take a truly random hash function h that maps each element uniformly and independently to the interval $(0, 1)$, and maintain the smallest and second smallest hash values ever seen from the stream. Note that if the same element appears multiple times, it will be mapped to the same hash value. Let X denote the second smallest hash value, and let n denote the number of distinct elements in the stream (assume $n \geq 2$). Note that n is unknown to the algorithm and is only used in the analysis.

   (a) Find $f_X(t)$, the PDF of X. Note that $f_X(t)$ depends on n. [Hint: First find the CDF $F_X(t) = \Pr(X \leq t)$ for $t \in (0, 1)$. Then the PDF is $f_X(t) = \frac{\mathrm{d}F_X(t)}{\mathrm{d}t}$.]

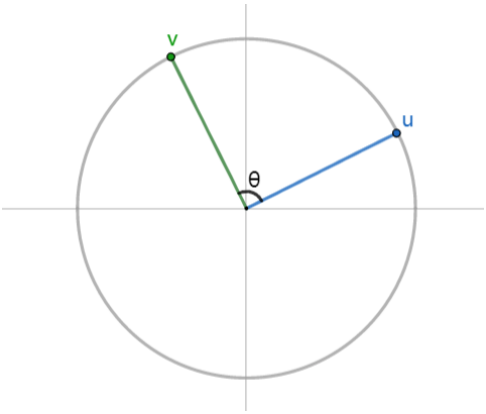


Figure 2

(b) We return $Y = 1/X$ as the estimator of n. Prove that Y is an unbiased estimator, i.e., $\mathbb{E}[Y] = n$. [Hint: For any function g, $\mathbb{E}[g(X)] = \int_0^1 g(t) f_X(t) \mathrm{d}t$.]

Forgot calculus? Here are the rules: $\frac{\mathrm{d}(y^n)}{\mathrm{d}t} = n y^{n-1} \frac{\mathrm{d}y}{\mathrm{d}t}$, $\int t^n \mathrm{d}t = \frac{1}{n+1} t^{n+1}$.

(c) However, it is not realistic to assume the existence of truly random hash functions. Show that even if h is just pairwise independent, Y is still a constant-factor approximation of n with constant probability, i.e., there exist constants $0 < c_1, c_2 < 1$ such that $\Pr[|Y - n| > c_1 n] < c_2$. [Hint: You can't apply Chebyshev inequality directly on X or Y as it's difficult to compute their variance. Try using indicator random variables.]