These are review questions for the actual Midterm 2.

---

1. Recall that a family $\mathcal{H}$ of hash functions is ***universal*** if $\Pr_{h\in\mathcal{H}}[h(x) = h(y)] \le 1/m$ for all distinct items $x \neq y$, where m is the size of the hash table. For any fixed hash function h, a ***collision*** is an unordered pair of distinct items $x \neq y$ such that $h(x) = h(y)$.

   Suppose we hash a set of n items into a table of size $m = 2n$, using a hash function h chosen uniformly at random from some universal family. Assume $\sqrt{n}$ is an integer.

   (a) ***Prove*** that the expected number of collisions is at most $n/4$.

   (b) ***Prove*** that the probability that there are at least $n/2$ collisions is at most $1/2$.

   (c) ***Prove*** that the probability that any subset of more than $\sqrt{n}$ items all hash to the same address is at most $1/2$. *[Hint: Use part (b).]*

   (d) Now suppose we choose h at random from a ***4-uniform*** family of hash functions, which means for all distinct items w, x, y, z and all addresses i, j, k, l, we have

   $$\Pr_{h\in\mathcal{H}}\big[h(w) = i \ \wedge\ h(x) = j \ \wedge\ h(y) = k \ \wedge\ h(z) = \ell\big] = \frac{1}{m^4}.$$

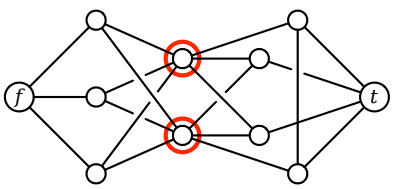
   Prove that the probability that any subset of more than $\sqrt{n}$ items all hash to the same address is at most $O(1/n)$.

   [Hint: All four statements have short elementary proofs via tail inequalities.]

2. The Island of Sodor is home to an extensive rail network. Recently, several cases of a deadly contagious disease have been reported in the village of Ffarquhar. The controller of the Sodor railway plans to close certain railway stations to prevent the disease from spreading to Tidmouth, his home town. No trains can pass through a closed station. To minimize expense (and public notice), he wants to close as few stations as possible. However, he doesn't want to close the Ffarquhar station, because that would expose him to the disease, and he *really* doesn't want to close the Tidmouth station, because then he couldn't visit his favorite pub in Tidmouth.

   The Sodor rail network is represented by an undirected graph, with a vertex for each station and an edge for each rail connection between two stations. Two special vertices f and t represent the stations in Ffarquhar and Tidmouth. Describe and analyze an algorithm to find the minimum number of stations *other than f and t* that must be closed to block all rail travel from Ffarquhar to Tidmouth.

   For example, given the following input graph, your algorithm should return the integer 2.



3. Let T be a treap with n vertices.

   (a) What is the *exact* expected number of leaves in T?

   (b) What is the *exact* expected number of nodes in T that have two children?

   (c) What is the *exact* expected number of nodes in T that have exactly one child?

   You do *not* need to prove that your answers are correct. *[Hint: What is the probability that the node with the kth smallest search key has no children, one child, or two children?]*

4. A *cyclic shift* of a string $A[1..n]$ is any string formed from A by moving a prefix of A to the end, or equivalently, moving a suffix of A to the beginning. For example, For example, the strings `RA!ABRACADAB` and `DABRA!ABRACA` and `ABRACADABRA!` are all cyclic shifts of the string `ABRACADABRA!`.

   (a) Describe and analyze an algorithm to determine, given two strings $A[1..n]$ and $B[1..n]$, whether A is a cyclic shift of B.

   (b) Describe a fast algorithm to determine, given two strings $A[1..m]$ and $B[1..n]$ with $m \leq n$, whether A is a substring of some cyclic shift of B.

Some Useful Inequalities

Suppose X is the sum of random indicator variables $X_1, X_2, \ldots, X_n$.
For each index i, let $p_i = \Pr[X_i = 1] = \mathrm{E}[X_i]$, and let $\mu = \sum_i p_i = \mathrm{E}[X]$.

- **Markov's Inequality:**

$$\Pr[X \geq x] \leq \frac{\mu}{x} \qquad \text{for all } x > 0, \text{ and therefore...}$$
$$\Pr[X \geq (1+\delta)\mu] \leq \frac{1}{1+\delta} \qquad \text{for all } \delta > 0$$

- **Chebyshev's Inequality:** If the variables X_i are pairwise independent, then...

$$\Pr[(X-\mu)^2 \geq z] < \frac{\mu}{z} \qquad \text{for all } z > 0, \text{ and therefore...}$$
$$\Pr[X \geq (1+\delta)\mu] < \frac{1}{\delta^2\mu} \qquad \text{for all } \delta > 0$$
$$\Pr[X \leq (1-\delta)\mu] < \frac{1}{\delta^2\mu} \qquad \text{for all } \delta > 0$$

- **Higher Moment Inequalities:** If the variables X_i are $2k$-wise independent, then...

$$\Pr[(X-\mu)^{2k} \geq z] = O\left(\frac{\mu^k}{z}\right) \qquad \text{for all } z > 0, \text{ and therefore...}$$
$$\Pr[X \geq (1+\delta)\mu] = O\left(\frac{1}{\delta^{2k}\mu^k}\right) \qquad \text{for all } \delta > 0$$
$$\Pr[X \leq (1-\delta)\mu] = O\left(\frac{1}{\delta^{2k}\mu^k}\right) \qquad \text{for all } \delta > 0$$

- **Chernoff's Inequality:** If the variables X_i are fully independent, then...

$$\Pr[X \geq x] \leq e^{x-\mu}\left(\frac{\mu}{x}\right)^x \qquad \text{for all } x \geq \mu, \text{ and therefore...}$$
$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\delta^2\mu/3} \qquad \text{for all } 0 < \delta < 1$$
$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\delta^2\mu/2} \qquad \text{for all } 0 < \delta < 1$$

- **The World's Most Useful Inequality:** $1 + x \leq e^x$ for all x
- **The World's Most Useful Limit:** $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e$

Hashing Properties

$\mathcal{H}$ is a set of functions from some universe $\mathcal{U}$ to $[m] = \{0, 1, 2, \ldots, m-1\}$.

- **Universal:** $\Pr_{h\in\mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m}$ for all distinct items $x \neq y$
- **Near-universal:** $\Pr_{h\in\mathcal{H}}[h(x) = h(y)] \leq O\left(\frac{1}{m}\right)$ for all distinct items $x \neq y$
- **Strongly universal:** $\Pr_{h\in\mathcal{H}}[h(x) = i \text{ and } h(y) = j] = \frac{1}{m^2}$ for all distinct $x \neq y$ and all i and j
- **2-uniform:** Same as strongly universal.
- **Ideal Random:** Fiction.