# CSC473 Assignment 2

Jeff Blair: 1002177057
jeffrey.blair@mail.utoronto.ca

Jaryd Hunter: 1002725893
jaryd.hunter@mail.utoronto.ca

March 2020

## Question 1

1. We will modify the NEARNEIGHBOR algorithm by adding each y to a set of points instead of returning the value, only failing on timeout. Additionally, we will define a new value for L so that there are enough hash tables to make the probability of collision for all neighbors of x in some table to be at least 5/6.

$$L = n^{\rho} \log(6n)$$

We chose $L$, since the $P(\forall s \in D(r,x) : \exists l : s \text{ collides with } x \text{ in } T_l) \geq 1 - e^{\frac{-L}{n^{\rho}}}$. But, with the modification we need a lower bound on the probability that one of the elements of $D(x,r)$ is not in one of the hash tables. So, we can use the union bound for $P(\exists s \in D(x,r) : \forall l : s \text{ does not collide with } x \text{ in } T_l)$, and we chose $L$ so that we still get the required probability. See correctness section below for derivation of $L$.

2.
```
ReportNeighbors(D, x)
    num_checked = 0
    neighbors = {}
    for l=1 to L:
        i = h_l(g_Il(x))
        Set y to the head of T_l[i]
        while y != NIL
            if dist(x, y) <= Cr
                neighbors.add(y)
            else
                num_checked++ # only increment when we miss a neighbor
            if num_checked == 12L + 1:
                return neighbors
            else
                Set y to the next element in T_l[i]
    return neighbors
```

### Running Time

The only substantial modifications to the algorithm was to L - which grows at a rate proportional to $n^{\rho}\log(n)$, and how long we wait until timeout which is now $|D(r, Cx)| + 12L + 1$ in the worst case. Then the total number of iterations the algorithm takes before completion or timeout is at most $|D(x,Cr)|n^{\rho}log(6n) + 12n^{\rho}\log(6n)$. Therefore the total running time of the algorithm is $T(n) \in O((1 + |D(x,Cr)|)n^{\rho}\log(n))$.

## Correctness

Proof that choice of $L$ will give the necessary probability of success.

$$\begin{aligned} P(\exists s \in D(x,r) : \forall l : s \text{ does not collide with } x \text{ in } T_l) \leq & \sum_{s \in D(x,r)} e^{\frac{-L}{n^\rho}} && \text{by Union bound} \\ =& |D(x,r)| e^{\frac{-L}{n^\rho}} \\ =& |D(x,r)| e^{\frac{-n^\rho log(6n)}{n^\rho}} \\ =& \frac{|D(x,r)|}{e^{log(6n)}} \\ =& \frac{|D(x,r)|}{6n} \\ \leq & \frac{1}{6} \end{aligned}$$

Proof that there are at most $12L$ strings more than further than $Cr$ away from $x$ (from notes):
Let $F : \{y \in D \mid \text{dist}(y,x) > Cr$
$\forall y \in F$ :

$$P(g_{I_l}(x) = g_{I_L}(y)) \leq p_2^k \leq \frac{1}{n}$$

$$\forall u \neq v : P(h_l(u) = h_l(v)) \leq \frac{1}{m} \leq \frac{1}{n}$$

$$P(h_l(g_{I_l}(x)) = h_l(g_{I_l}(y))) \leq \frac{1}{n} + \frac{1}{n} \leq \frac{2}{n}$$

Let $X_{y,l}$ be the indicator random variable which equals 1 if y collides with x

$$E[X] = \sum_{y \in F} \sum_{l \in L} P(X_{y,l} = 1) \leq \frac{2|F|L}{n} \leq 2L$$

$$P(X > 12) < \frac{1}{6} \text{ by Markov's Inequality}$$

Then the total probability that the algorithm succeeds at producing every $y \in D(x,r)$ is at least $1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}$

# Question 2

a.

$$\forall n \in \{1, ..., k\} : P\left(\text{rank}(A[i_n]) \leq (\frac{1}{2} - \epsilon)n\right) = \frac{(\frac{1}{2} - \epsilon)n}{n} = (\frac{1}{2} - \epsilon)$$

$$\text{Then } X \sim \text{Binomial}(p = \tfrac{1}{2} - \epsilon, n = k)$$

$$\begin{aligned}
E[X] &= \sum_{x=1}^{k} xP(X = x) \\
&= \sum_{x=1}^{k} x\binom{k}{x} \cdot \left(\frac{1}{2} - \epsilon\right)^{x} \cdot \left(\frac{1}{2} + \epsilon\right)^{k-x} \\
&= \sum_{x=1}^{k} k\binom{k-1}{x-1} \cdot \left(\frac{1}{2} - \epsilon\right)^{x} \cdot \left(\frac{1}{2} + \epsilon\right)^{k-x} \\
&= k \cdot \left(\frac{1}{2} - \epsilon\right) \sum_{x=1}^{k} \binom{k-1}{x-1} \cdot \left(\frac{1}{2} - \epsilon\right)^{x-1} \cdot \left(\frac{1}{2} + \epsilon\right)^{(k-1)-(x-1)} \\
&= k \cdot \left(\frac{1}{2} - \epsilon\right) \cdot \left(\left(\frac{1}{2} - \epsilon\right) + \left(\frac{1}{2} + \epsilon\right)\right)^{k-1} \text{ since } (a+b)^k = \sum_{n=0}^{k} \binom{k}{n} a^k b^{n-k} \\
&= k \cdot \left(\frac{1}{2} - \epsilon\right) \cdot (1)^{k-1} \\
&= k \cdot \left(\frac{1}{2} - \epsilon\right)
\end{aligned}$$

$$
\begin{aligned}
E[X^2] &= \sum_{x=0}^{k} x^2 P(X=x) \\
&= \sum_{x=0}^{k} x^2 \binom{k}{x} \cdot \left(\frac{1}{2}-\epsilon\right)^{x} \cdot \left(\frac{1}{2}+\epsilon\right)^{k-x} \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \sum_{x=1}^{k} x \cdot \binom{k-1}{x-1} \cdot \left(\frac{1}{2}-\epsilon\right)^{x-1} \cdot \left(\frac{1}{2}+\epsilon\right)^{(k-1)-(x-1)} \\
&\quad \text{with } p = (\frac{1}{2}-\epsilon), a = x-1, b = k-1 \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \sum_{a=0}^{b} (a+1) \cdot \binom{b}{a} \cdot p^a \cdot (1-p)^{b-a} \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \sum_{a=0}^{b} a \cdot \binom{b}{a} \cdot p^a \cdot (1-p)^{b-a} + \binom{b}{a} \cdot p^a \cdot (1-p)^{b-a} \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( \sum_{a=0}^{b} b \cdot \binom{b-1}{a-1} \cdot p^a \cdot (1-p)^{b-a} + \sum_{a=0}^{b} \binom{b}{a} \cdot p^a \cdot (1-p)^{b-a} \right) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( bp \sum_{a=0}^{b} \cdot \binom{b-1}{a-1} \cdot p^{a-1} \cdot (1-p)^{b-a} + \sum_{a=0}^{b} \binom{b}{a} \cdot p^a \cdot (1-p)^{b-a} \right) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( bp(p+(1-p))^{b-1} + (p+(1-p))^b \right) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) (bp+1) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( (k-1)\left(\frac{1}{2}-\epsilon\right) + 1 \right) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( (k-1)\left(\frac{1}{2}-\epsilon\right) + 1 \right) \\
&= k \cdot \left(\frac{1}{2}-\epsilon\right) \left( k\left(\frac{1}{2}-\epsilon\right) - \left(\frac{1}{2}-\epsilon\right) + 1 \right) \\
&= k^2 \cdot \left(\frac{1}{2}-\epsilon\right)^2 - k\left(\frac{1}{2}-\epsilon\right)^2 + k\left(\frac{1}{2}-\epsilon\right) \\
&= k^2 \cdot \left(\frac{1}{2}-\epsilon\right)^2 + k\left(\frac{1}{2}-\epsilon\right)\left(1-\left(\frac{1}{2}-\epsilon\right)\right)
\end{aligned}
$$

Then,

$$
\begin{aligned}
Var(X) =& E[X^2] - E[X]^2 \\
=& k^2 \cdot \left(\frac{1}{2}-\epsilon\right)^2 + k\left(\frac{1}{2}-\epsilon\right)\left(1-\left(\frac{1}{2}-\epsilon\right)\right) - \left(k \cdot \left(\frac{1}{2}-\epsilon\right)\right)^2 \\
=& k\left(\frac{1}{2}-\epsilon\right)\left(1-\left(\frac{1}{2}-\epsilon\right)\right) \\
=& k\left(\frac{1}{2}-\epsilon\right)\left(\frac{1}{2}+\epsilon\right)
\end{aligned}
$$

b. We need to calculate $P\left[(\frac{1}{2}-\epsilon)n \leq \text{rank}(Z) \leq (\frac{1}{2}m+\epsilon)n\right]$, but we can calculate this probability by writing it in terms of distributions similar to $X$ which we solved in part a. Where we know if there are $\frac{k}{2}$ elements with rank less than the lower bound, or greater than the larger bound then we know rank$(Z)$ will be outside that range, we will designate two random variables that will represent these two events separately.

Let $L$ be the random variable of how many elements in the sample have rank less than or equal to $(\frac{1}{2}-\epsilon)n$, and let $H$ be the random variable of how many elements in the sample have rank greater than or equal to $(\frac{1}{2}+\epsilon)n+1$. $L$ Follows the same distribution as $X$ from part a, so $E[L] = k(\frac{1}{2}-\epsilon)$, and $V[L] = k(\frac{1}{2}-\epsilon)(\frac{1}{2}+\epsilon)$. $H$ also follows a binomial distribution with

$$p = \frac{n-(\frac{1}{2}+\epsilon)n}{n} = 1 - \frac{1}{2} - \epsilon = \frac{1}{2} - \epsilon$$

Then, $E[H] = k(\frac{1}{2}-\epsilon)$ and $V[H] = k(\frac{1}{2}-\epsilon)(\frac{1}{2}+\epsilon)$.

Note,

$$\begin{aligned}
P[L > \frac{k}{2}] =& P[L - E[L] > \frac{k}{2} - E[L]] \\
<& \frac{k(\frac{1}{2}-\epsilon)(\frac{1}{2}+\epsilon)}{(\frac{k}{2} - k(\frac{1}{2}-\epsilon))^2} && \text{Using Chebyshev inequality} \\
=& \frac{k(\frac{1}{2}-\epsilon)(\frac{1}{2}+\epsilon)}{k^2\epsilon^2} \\
=& \frac{\frac{1}{4}-\epsilon^2}{k\epsilon^2} \\
\leq& \frac{\frac{1}{4}-\epsilon^2}{2} && \text{Let, } k \geq \frac{2}{\epsilon^2} \\
=& \frac{1}{8} - \frac{\epsilon^2}{2} \\
\leq& \frac{1}{8}
\end{aligned}$$

Since, $E[H] = E[L]$ and $V[H] = V[L]$ the same result applies to $P[H > \frac{k}{2}]$.

$$\begin{aligned}
P\left[(\frac{1}{2}-\epsilon)n + 1 \leq \text{rank}(Z) \leq (\frac{1}{2}m+\epsilon)n\right] =& P\left[\text{rank}(Z) \geq ((\frac{1}{2}-\epsilon)n)+1\right] - P\left[\text{rank}(Z) > (\frac{1}{2}m+\epsilon)n\right] \\
=& 1 - P[L > \frac{k}{2}] - P[H > \frac{k}{2}] \\
\geq& 1 - \frac{1}{8} - \frac{1}{8} \\
\geq& \frac{3}{4} \geq \frac{1}{2}
\end{aligned}$$