Problem 1

1. Assume our hash table has cm^2 slots for large constant c. Conditioned on the event that none of the first $i-1$ items collided when inserted in the table, the probability that the i^{th} item does not causes a collision is

$$1 - \frac{i-1}{cm^2}.$$

In particular, when the i^{th} item is inserted, the table has $i-1$ filled slots out of cm^2 total.

From the expression above, we can bound the probability there are no collisions by:

$$\left(1 - \frac{0}{cm^2}\right) \cdot \left(1 - \frac{1}{cm^2}\right) \cdot \left(1 - \frac{2}{cm^2}\right) \cdot \ldots \cdot \left(1 - \frac{m-1}{cm^2}\right) < \left(1 - \frac{1}{cm}\right)^m.$$

The inequality follow from the fact the $i/m < 1$ for all $i \leq m-1$.

As hinted, for $cm \geq 2$, $(1 - \frac{1}{cm})^{cm} \geq 1/2e$. So, $(1 - \frac{1}{cm})^m = \left((1 - \frac{1}{cm})^{cm}\right)^{1/c} \geq (1/2e)^{1/c}$. Choosing $c \geq 17$ yields $(1/2e)^{1/c} \geq 9/10$. So, as long as the table has $\geq 17m^2$ slots, there is no collision with probability $> 9/10$.

2. By linearity of expectation, we have the $\mathbb{E}[\|x\|_2^2] = \sum_{i=1}^d \mathbb{E}[x_i^2]$. So, we focus on bounding $\mathbb{E}[x_i^2]$ for a single i. We have that $x_i = \sum_{j=1}^n s_j$ where:

$$s_j = \begin{cases} 0 & \text{with probability } 1 - 1/d \\ -1 & \text{with probability } 1/2d \\ +1 & \text{with probability } 1/2d \end{cases}$$

So we calculate the expectation:

$$\begin{aligned} \mathbb{E}[x_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^n s_j\right)^2\right] &= \mathbb{E}\left[s_1^2 + \ldots + s_n^2 + s_1 s_2 + \ldots + s_j s_k + \ldots + s_{n-1} s_n\right] \\ &= \mathbb{E}[s_1^2] + \ldots + \mathbb{E}[s_n^2] + \mathbb{E}[s_1 s_2] + \ldots + \mathbb{E}[s_j s_k] + \ldots + \mathbb{E}[s_{n-1} s_n] \end{aligned}$$

We have that for each s_j, $\mathbb{E}[s_j] = 0$ and s_j and s_k are independent, so all of the terms of the form $\mathbb{E}[s_j s_k]$ equal 0. So we have that:

$$\mathbb{E}[x_i^2] = \mathbb{E}[s_1^2] + \ldots + \mathbb{E}[s_n^2].$$

It can be directly calculated that for each j, $\mathbb{E}[s_j^2] = 1 \cdot \frac{1}{2d} + 1 \cdot \frac{1}{2d} = \frac{1}{d}$. So $\mathbb{E}[x_i^2] = \frac{n}{d}$ and finally,

$$\mathbb{E}[\|x\|_2^2] = \sum_{i=1}^d \mathbb{E}[x_i^2] = d \cdot \frac{n}{d} = n.$$

Problem 2

There are many ways to solve this problem, including constructions that use multiple tables (like we did in class). I will provide a simple solution using just one table, A, with size cm, where c is a

large constant to be chosen later. Let $\tilde{f} = A(h(v))$ be the frequency estimate for item v obtained from just that table.

To get the improved bound, we split up $A(h(v))$ in a slightly refined way than what was done in class. The intuition is that, for large enough c, item v will not collide with *any* of the top m most frequent items with high probability. Call these items $v_1, \ldots, v_m$. Call the less frequent items $v_{m+1}, v_{m+2}, \ldots$. We have:

$$A(h(v)) = f(v) + \sum_{i=1,\ldots,m, v_i \neq v} \mathbb{1}[h(v) = h(v_i)] \cdot f(v_i) + \sum_{i>m, v_i \neq v} \mathbb{1}[h(v) = h(v_i)] f(y).$$

The second two terms are our error terms. To bound the first, note that the event $\mathbb{1}[h(v) = h(v_i)]$ only happens with probability $1/cm$ if our table has cm slots in it. Moreover, $\mathbb{1}[h(v) = h(v_i)]$ is independent from $\mathbb{1}[h(v) = h(v_j)]$ for $i \neq j$. So, none of the events $\mathbb{1}[h(v) = h(v_i)]$ happen with probability:

$$(1 - 1/cm)^m = ((1 - 1/cm)^{cm})^{1/c} \geq (1/2e)^{1/c}.$$

Choosing $c = 50$, we have that $(1/2e)^{1/c} \geq .95$, so

$$\Pr\left[\sum_{i=1,\ldots,m, v_i \neq v} \mathbb{1}[h(v) = h(v_i)] \cdot f(v_i) = 0\right] \geq 19/20.$$

I.e., with very high probability, the first error term is 0.

We now turn to the second error term. As in class, we have that:

$$\mathbb{E}\left[\sum_{i>m, v_i \neq v} \mathbb{1}[h(v) = h(v_i)] f(y)\right] = \frac{1}{cm} \sum_{i>m, v_i \neq v} f(y) \leq \frac{1}{cm} \cdot C = \frac{1}{50m} \cdot C.$$

Accordingly, by Markov's inequality,

$$\Pr\left[\sum_{i>m, v_i \neq v} \mathbb{1}[h(v) = h(v_i)] f(y) \geq \frac{1}{m} C\right] \leq 1/50.$$

We conclude by applying a union bound: the probability that the first error term is non-zero *or* the second error term is $\geq \frac{1}{m} C$ is upper bounded by $1/20 + 1/50 \leq 1/10$. Accordingly, we conclude

$$\Pr\left[\tilde{f}(v) - f(v) \geq \frac{1}{m} C\right] \leq 1/10.$$

Problem 3

1. Split the people being tested into C equal groups with n/C people each. At most k of the groups will test positive since, in the worst case, each infected person will be in a different group. For each positive group, we need to rerun n/C individual tests. So the number of tests run is at most $C + \frac{n}{C} \cdot k$. If we set $C = \sqrt{nk}$ then the total number of tests is:

$$C + \frac{n}{C} \cdot k = \sqrt{nk} + \frac{n}{\sqrt{n}\sqrt{k}} k = 2\sqrt{nk},$$

as desired.

2. Assume $q = c \log n$ for sufficiently large constant c and assume $C = 2k$. Consider any single individual who is negative. For this individual to be falsely reported positive, they would need to test positive in all q group tests they participates in.

What is the probability they test positive in any one group test? For this to happen, there would need to be a positive individual in that group. Let Z be the number of positive individuals in any given group. Each group has size $\frac{n}{2k}$, so by linearity of expectation, $\mathbb{E}[Z] = \frac{n}{2k} \cdot \frac{k}{n} = \frac{1}{2}$, since the probability that any group member is positive is k/n. Then by Markov's inequality, $\Pr[Z \geq 1] \leq \frac{1}{2}$. I.e., with probability $\geq \frac{1}{2}$ any given group contains *no infected individuals*.

So, the probability our negative individual has a positive person in all $q = c \log n$ of their groups is less than $\frac{1}{2}^q = \frac{1}{n^c}$, which is $\leq \frac{1}{10n}$ for sufficiently large c. Thus, the probability any given negative individual gets a false positive result is $\leq \frac{1}{10n}$. There are $n - k < n$ negative individuals and by a union bound, the probability *any* of them gets a false positive test is $\leq \frac{1}{10n} \cdot n \leq \frac{1}{10}$. In other words, we get no false positives with probability $\geq 9/10$.

3. **Informal**: The lower bound is via a counting argument. One way of framing our goal is that we need to return a bit string of length n which has a 1 in exactly k places and zeros everywhere else, with the 1's indicating the people we believe are infected, and the 0's indicating uninfected.

Before starting the testing scheme, we don't know what the correct bit string is, and there are $\binom{n}{k}$ different possibilities. When we run any testing scheme, we will receive a binary reponse to every test $t_1, \ldots, t_T$ and from that response will decide how to run the next test. In the end, we will decide on which bit string of the $\binom{n}{k}$ different possibilities is correct. The solution we return must be a function of $t_1, \ldots, t_T$: it could also depend on what subsets of people were tested, but at any time i, that must be a function of $t_1, \ldots, t_{i-1}$, via induction.

So, to obtain a correct answer after T tests, it must be that the number of possible test results $t_1, \ldots, t_T$ exceeds the number of possible solutions $\binom{n}{k}$. In other words that $2^t \geq \binom{n}{k}$. Taking logs on both sides, and noting that $\log \binom{n}{k} = O(k \log(n/k))$ gives the result.

Problem 4

1. Recall that $\tilde{n} = \frac{m(m-1)}{2D}$, where m is the number of samples collected, and D is the number of duplicates observed. To prove that $(1-\epsilon)n \leq \tilde{n} \leq (1+\epsilon)n$ we first claim that it suffices to show:

$$(1 - \epsilon/2)\,\mathbb{E}[D] \leq D \leq (1 + \epsilon/2)\,\mathbb{E}[D]. \tag{1}$$

In particular, if this is the case, then:

$$\frac{1}{1+\epsilon/2}\frac{1}{\mathbb{E}[D]} \leq \frac{1}{D} \leq \frac{1}{1-\epsilon/2}\frac{1}{\mathbb{E}[D]},$$

Using the inequalities from the start of Lecture 2, it follows that for $0 < \epsilon \leq 1$,

$$(1-\epsilon/2)\frac{1}{\mathbb{E}[D]} \leq \frac{1}{D} \leq (1+\epsilon)\frac{1}{\mathbb{E}[D]}.$$

Multiplying all sides of the inequalities by $m(m-1)/2$, we conclude that $(1-\epsilon)n \leq \tilde{n} \leq (1+\epsilon)n$

So, we focus on proving (1). To do so, we will bound the variance of D and use Chebyshev's inequality. Use $s_1, \ldots, s_m$ to denote our m samples. We have that:

$$D = \sum_{i<j} \mathbb{1}[s_i = s_j].$$

Note that the terms in the sum are pairwise independent. In particular, the event that $s_i = s_j$ does not effect the probability that $s_k = s_j$ for some other value of k. As discussed in class, the terms are not mutually independent, since the event that $s_i = s_j$ and $s_j = s_k$ implies that $s_i = s_k$. Regardless, pairwise independence is all we need to apply linearity of variance:

$$\text{Var}[D] = \sum_{i<j} \text{Var}[\mathbb{1}[s_i = s_j]].$$

$\mathbb{1}[s_i = s_j]$ is a binary random variable equal to 1 with probability $1/n$, so its variance is equal to $\frac{1}{n} - \frac{1}{n^2}$. We conclude that

$$\text{Var}[D] \leq \sum_{i<j} \frac{1}{n} = \binom{m}{2} \cdot \frac{1}{n} = \frac{m(m-1)}{2n}.$$

Note that this is exactly equal to $\mathbb{E}[D]$. So, we have via Chebyshev's inequality that:

$$\Pr\left[|D - \mathbb{E}[D] \leq \sqrt{10}\sqrt{\mathbb{E}[D]}\right] \leq \frac{1}{10}.$$

To prove (1), we need to choose m large enough so that $\sqrt{10}\sqrt{\mathbb{E}[D]} \leq \epsilon\,\mathbb{E}[D]$. I.e., we need

$$\sqrt{\frac{m(m-1)}{2n}} \geq \frac{\sqrt{10}}{\epsilon}$$
$$m(m-1) \geq \frac{20n}{\epsilon^2}$$
$$m \geq \sqrt{\frac{40n}{\epsilon^2}}.$$

This proves the claim.

3. For the mark-and-recapture estimator to be accurate as the number of samples goes to infinity, the key property we required was that $\mathbb{E}[D] = \frac{m(m-1)}{2n}$. Since we use the estimator $\tilde{n} = \frac{m(m-1)}{2D}$, we will obtain an *underestimate* in the limit if:

$$\mathbb{E}[D] > \frac{m(m-1)}{2n}.$$

We claim that this is the case for data items drawn from *any* non-uniform distribution – Wikipedia's particular distribution doesn't matter too much.

In particular, suppose the sample procedure selects item z with probability p_z. If we collect samples $s_1, \ldots, s_m$, the expected number of duplicates is:

$$\mathbb{E}[D] = \mathbb{E}\left[\sum_{i<j} \mathbb{1}[s_i = s_j]\right] = \frac{m(m-1)}{2} \cdot \mathbb{E}\left[\mathbb{1}[s_1 = s_2]\right].$$

Here I am using that $\mathbb{1}[s_i = s_j]$ is the idenitically distributed for all i, j, so I can just fixed i and j to be $1, 2$ arbitrarily.

What is $\mathbb{E}[\mathbb{1}[s_1 = s_2]]$? We can write this as follows:

$$\mathbb{E}[\mathbb{1}[s_1 = s_2]] = \Pr[s_1 = s_2] = \sum_{z=1}^{n} \Pr[s_1 = z \text{ and } s_2 = z] = \sum_{z=1}^{n} p_z^2.$$

So, we conclude that:

$$\mathbb{E}[D] = \frac{m(m-1)}{2} \cdot \sum_{z=1}^{n} p_z^2.$$

We will obtain an underestimate if $\sum_{z=1}^{n} p_z^2 > n$. Note that it is exactly equal to n in the uniform case when $p_1, \ldots, p_n = \frac{1}{n}$. There are a number of ways to see that it is larger for any other choice of n.

One way is via an "exchange argument". Suppose our probabilities are not uniform, so there are at least two items a and b such that $p_a \neq p_b$. Now, consider modifying the distribution so that p_a and p_b are both replaced with their average $p'_a = p'_b = \frac{p_a + p_b}{2}$. Then, after this change, the summation $\sum_{z=1}^{n} p_z^2$ changes by:

$$p_a'^2 + p_b'^2 - p_a^2 - p_b^2 = 2 \cdot \frac{p_a^2 + 2p_a p_b + p_b^2}{4} - p_a^2 - p_b^2 = p_a p_b - p_a^2/2 - p_b^2/2 = -(p_a/\sqrt{2} - p_b/\sqrt{2})^2.$$

This change is always *negative*, meaning that the sum decreases. So, take any non-uniform distribution and make it more uniform by averaging two of the non-equal probabilities and you will *strictly decrease* $\sum_{z=1}^{n} p_z^2$. We conclude that the sum is always strictly larger than the value of n obtained by the uniform distribution.

4. In Wikipedia's case, the probabilities $p_1, \ldots, p_n$ are themselves random. So, to understand, $\sum_{z=1}^{n} p_z^2$, it makes sense to consider:

$$\mathbb{E}\left[\sum_{z=1}^{n} p_z^2\right] = n \cdot \mathbb{E}[p_z^2].$$

For simplicity, imagine that we are assigning Wikipedia pages random numbers between $[0, 1]$ We can compute the expecation of p_z using a similar approach as in Lecture 1 by first trying to understand $\Pr[p_z \geq t]$ for a fixed threshold t. In particular, for p_z to be $\geq t$, it must be that no other page lands within distance t of r_z. Imageine r_z is chosen first. Then the chance this happens is $(1-t)^{n-1}$. So have have:

$$\Pr[p_z \geq t] = (1-t)^{n-1}$$

or equivalently

$$\Pr[p_z^2 \geq t] = (1-\sqrt{t})^{n-1}$$

We thus have:

$$\mathbb{E}[p_z^2] = \int_0^1 \Pr[p_z^2 \geq t] dt = \int_0^1 (1-\sqrt{t})^{n-1} dt = \frac{2}{n^2+n}.$$

It follows that:

$$\mathbb{E}\left[\sum_{z=1}^{n} p_z^2\right] = n \cdot \mathbb{E}[p_z^2] = \frac{2}{n+1}.$$

For large n, this is very close to $2/n$. So, $\mathbb{E}[D]$ will be almost exactly twice as large as when the probabilities are exactly uniform (in which case $\sum_{z=1}^{n} p_z^2 = 1/n$).