Problem 1

a) We can show that $E[C^2] = F_2$ as follows:

$$C = \sum_{i=1}^{n} a_i H(x_i)$$
$$E[C^2] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j H(x_i)H(x_j)\right]$$
$$E[C^2] = \sum_{i=1}^{n}\sum_{j=1}^{n} E\left[a_i a_j H(x_i)H(x_j)\right]$$

We first note that $E[a_i H(x_i)]$ is 0 since there is equal probability that $H(x_i)$ is -1 and 1. We will use this fact throughout this problem. We can see that if $i \neq j$, then the terms are independent (since $\mathscr{H}$ is 2-independent and a of each element is independent of each other)

$$\begin{aligned} E\left[a_i a_j H(x_i)H(x_j)\right] &= E[a_i H(x_i)]E[a_j H(x_j)] \text{ if } i \neq j \\ &= 0 \cdot 0 = 0 \end{aligned}$$

Thus, the expression evaluates to

$$\begin{aligned} E[C^2] &= \sum_{i=1}^{n}\sum_{j=1}^{n} E\left[a_i a_j H(x_i)H(x_j)\right] \\ &= \sum_{i=1}^{n} E\left[a_i a_i H(x_i)H(x_i)\right] \\ &= \sum_{i=1}^{n} E\left[a_i^2\right] \\ &= \sum_{i=1}^{n} a_i^2 = F_2 \end{aligned}$$

b) We can bound the variance as follows:

$$\begin{aligned} \mathrm{Var}(C^2) &= E\left[C^4\right] - E\left[C^2\right]^2 \\ &= E\left[C^4\right] - F_2^2 \end{aligned}$$

Looking at the first term individually, we have

$$\begin{aligned} E\left[C^4\right] &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} a_i a_j a_k a_l H(x_i)H(x_j)H(x_k)H(x_l)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left[a_i a_j a_k a_l H(x_i)H(x_j)H(x_k)H(x_l)\right] \end{aligned}$$

We can see that for any $i \neq j$, $i \neq k$, and $i \neq l$, we can treat the terms involving the index i as independent since $\mathscr{H}$ is 4-independent and a of every element is independent of each other:

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left[a_i a_j a_k a_l H(x_i)H(x_j)H(x_k)H(x_l)\right] &= \sum_{i=1}^{n} E\left[a_i H(x_i)\right]\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left[a_j a_k a_l H(x_j)H(x_k)H(x_l)\right] \\ &= 0 \cdot \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left[a_j a_k a_l H(x_j)H(x_k)H(x_l)\right] \\ &= 0 \end{aligned}$$

Thus, we can eliminate (as in lecture) any terms that have one different element or three of the same elemement (since that implies one different element). This leaves us with the terms that contain either all different elements or two pairs of elements.

$$\sum_{i=1}^{n} E\left[(a_i H(x_i))^4\right] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right]$$

where the first term represents all the terms in the original summation with all distinct i, j, k, l and the second represents the terms with two pairs of elements (sum of all pairs times 4 choose 2 permutations of them). We can then simplify as shown in lecture (converting second inner summation from [i+1, n] to [1,n] with $i \neq j$ and doubling):

$$\begin{aligned}
\sum_{i=1}^{n} E\left[(a_i H(x_i))^4\right] + 6\sum_{i=1}^{n} \sum_{j=i+1}^{n} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] &= \sum_{i=1}^{n} E\left[(a_i H(x_i))^4\right] + 3\sum_{i \neq j} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] \\
&= \sum_{i=j} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] + 3\sum_{i \neq j} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] \\
&\leq 3(\sum_{i=1}^{n} E\left[(a_i H(x_i))^2\right])^2
\end{aligned}$$

where the last step follows since $\sum_{i=j} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] + \sum_{i \neq j} E\left[(a_i H(x_i))^2 (a_j H(x_j))^2\right] = (\sum_{i=1}^{n} E\left[(a_i H(x_i))^2\right])^2$ due to 4-independence of $\mathscr{H}$ and a. This bound is similar to the one in lecture 12. We can then continue to simplify:

$$\begin{aligned}
3(\sum_{i=1}^{n} E\left[(a_i H(x_i))^2\right])^2 &= 3(\sum_{i=1}^{n} E\left[a_i^2\right])^2 \\
&= 3(\sum_{i=1}^{n} (a_i^2))^2 \\
&= 3F_2^2
\end{aligned}$$

Taking it back to the original expression, we have

$$\begin{aligned}
\text{Var}(C^2) &= E\left[C^4\right] - F_2^2 \\
&\leq 3F_2^2 - F_2^2 \\
&= 2F_2^2
\end{aligned}$$

c) Chebyshev's inequality is as follows:

$$Pr\left(|x - \mu| \geq c\sigma\right) \leq \frac{1}{c^2}$$

If we let x be C^2, then μ is F_2 $\sigma^2 \leq 2F_2^2$, and $\sigma \leq \sqrt{2}F_2$ This gives us the following

$$Pr\left(\left|C^2 - F_2\right| \geq c\sqrt{2}F_2\right) \leq \frac{1}{c^2}$$

If we let $c = \frac{\epsilon}{\sqrt{2}}$ we have

$$Pr\left(\left|C^2 - F_2\right| \geq \epsilon F_2\right) \leq \frac{2}{\epsilon^2}$$

This isn't a particularly useful bound because in order for the bounding probability to be less than 1, ϵ has to be at least $\sqrt{2}$. We want to be able to bound the probability to a certain value given any ϵ.

d) We can show expectance as follows:

$$\begin{aligned}
E[D] &= E\left[\frac{1}{w}\sum_{i=1}^{w} C_i^2\right] \\
&= \frac{1}{w}E\left[\sum_{i=1}^{w} C_i^2\right] \\
&= \frac{1}{w}\sum_{i=1}^{w} E\left[C_i^2\right] \\
&= \frac{1}{w}w \cdot F_2 \\
&= F_2
\end{aligned}$$

We can show variance as follows:

$$
\begin{align}
\text{Var}(D) &= \text{Var}\left(\frac{1}{w}\sum_{i=1}^{w} C_i^2\right) \tag{1}\\
&= \frac{1}{w^2}\text{Var}\left(\sum_{i=1}^{w} C_i^2\right) \tag{2}\\
&= \frac{1}{w^2}\left(\sum_{i=1}^{w} \text{Var}(C_i^2)\right) \tag{3}\\
&\leq \frac{1}{w^2}\left(\sum_{i=1}^{w} 2F_2^2\right) \tag{4}\\
&= \frac{1}{w^2}\left(w \cdot 2F_2^2\right) \tag{5}\\
&= \frac{2F_2^2}{w} \tag{6}
\end{align}
$$

where (3) follows because each run is independent (i.e. all C_i^2 are independent).

e) We can again use Chebyshev's inequality to bound the probability.

$$Pr\left(|x-\mu| \geq c\sigma\right) \leq \frac{1}{c^2}$$

If we let x be D, then μ is F_2 and $\sigma^2 \leq \frac{2F_2^2}{w}$, which means $\sigma \leq F_2\sqrt{\frac{2}{w}}$. We can then show that

$$
\begin{aligned}
Pr\left(|x-\mu| \geq c\sigma\right) &\leq \frac{1}{c^2}\\
Pr\left(|D-F_2| \geq cF_2\sqrt{\frac{2}{w}}\right) &\leq \frac{1}{c^2}
\end{aligned}
$$

If we let $c = \epsilon\sqrt{\frac{w}{2}}$, we have

$$Pr\left(|D-F_2| \geq \epsilon F_2\right) \leq \frac{2}{w\epsilon^2}$$

We then bound the inverse of this probability to be at least p and solve for our desired w.

$$
\begin{aligned}
1 - \frac{2}{w\epsilon^2} &\geq p\\
1 - p &\geq \frac{2}{w\epsilon^2}\\
w &\geq \frac{2}{(1-p)\epsilon^2}\\
&= O(\epsilon^{-2})
\end{aligned}
$$

f) We can choose the median value of all the results again as in a count sketch and use that as the estimate. We let $p = 1 - \frac{1}{e}$ Then we can use the Chernoff Bound as in lecture 11. The bound states:

$$Pr\left[X > n/2\right] < e^{\frac{-n(1/2-q)^2}{2q}}$$

if X $\sim$ Binom(n,q) and $q < 1/2$. We need at least half of the estimates to be out of the bound to return the wrong median and we can model X as a binomial distribution since each of the copies of the data structure is independent. Since p is the probability of getting a value that is within the error bound, we let $q = 1 - p = \frac{1}{e}$ in the Chernoff Equation. We then get

$$
\begin{aligned}
Pr\left[X > d/2\right] &< e^{\frac{-n(1/2-1/e)^2}{2(1/e)}}\\
&= e^{(-O(1)\cdot d)}
\end{aligned}
$$

If we choose $d = O(\log \delta^{-1})$, then $Pr[X > d/2] \leq \delta$ as desired and the success probability is at least $1 - \delta$.