We will explore applications of the variance of a random variable X today. While knowing the expected value of X is useful, this tells us nothing about the typical value of X. As a reminder,

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}X)^2\big] = \mathbb{E}\big[X^2\big] - \mathbb{E}\big[X\big]^2.$$

While the variance can be difficult to calculate, if $X = \sum_i X_i$, then we have a formula which can make life easier; namely,

$$\mathrm{Var}(X) = \sum_{i,j} \big(\mathbb{E}[X_i X_j] - \mathbb{E}X_i \mathbb{E}X_j\big).$$

Chebyshev's inequality. For any $\lambda > 0$,

$$\Pr\left[|X - \mathbb{E}X| \geq \lambda\sqrt{\mathrm{Var}(X)}\right] \leq \frac{1}{\lambda^2}.$$

The strength of Chebyshev's inequality is that it tells us that a random variable doesn't stray too far from its mean.

Let's start by proving something a bit silly. We know from Stirling's formula that

$$\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}},$$

but let's try to get a general (not asymptotic) lower bound by using probabilistic methods.

Claim 1.

$$\binom{2n}{n} \geq \frac{4^n}{4\sqrt{n} + 2}$$

Proof. Consider selecting a subset $S \subseteq [2n]$ uniformly at random; equivalently, independently for each $i \in [2n]$ include x in S with probability $1/2$. Now let X_i be the random variable which is 1 if $i \in S$ and 0 otherwise and let $X = |S|$. Clearly $X = \sum_{i=1}^{2n} X_i$, so $\mathbb{E}X = n$. On the other hand, as each element was added independently,

$$\mathbb{E}[X_i X_j] = \begin{cases} \frac{1}{2} & \text{if } i = j \\ \frac{1}{4} & \text{otherwise.} \end{cases}$$

Thereby, as $X = \sum_{i=1}^{2n} X_i$ and $\mathbb{E}X_i = 1/2$,

$$\mathrm{Var}(X) = \sum_{i,j \in [2n]} (\mathbb{E}[X_i X_j] - \mathbb{E}X_i \mathbb{E}X_j) = \sum_{i=1}^{2n} \left(\frac{1}{2} - \frac{1}{4}\right) = \frac{n}{2}.$$

By Chebyshev's inequality, we find that

$$\Pr\left[|X - n| \geq \lambda\sqrt{\frac{n}{2}}\right] \leq \frac{1}{\lambda^2}.$$

for all $\lambda > 0$. In other words, for $\lambda = \sqrt{2}$,

$$\Pr\left[|X - n| < \sqrt{n}\right] \geq \frac{1}{2}.$$

Now, as S was chosen uniformly at random,

$$\Pr[X = k] = \binom{2n}{k} 4^{-n}.$$

Hence,

$$\frac{1}{2} \leq \Pr\left[|X-n| < \sqrt{n}\right] = \sum_{|k|<\sqrt{n}} \Pr[X = n+k] = \sum_{|k|<\sqrt{n}} \binom{2n}{n+k} 4^{-n} \leq (2\sqrt{n}+1)\binom{2n}{n} 4^{-n},$$

from which the result follows. □

Claim 2. Let $G_1, \dots, G_k$ be graphs on the same vertex set each with m edges. There is a partition of the vertices (A, B) such that for each i, G_i has at least $\frac{m}{2} - c\sqrt{m}$ edges between A and B where c is a constant depending only on k.

Proof. Certainly we know that each of the G_i has a partition of the vertices for which there are at least $\frac{m}{2}$ edges crossing between the parts, but this partition need not be the same for each i. This is where we can use variance to show that there is a partition that works for *all* i that *almost* has half of the edges crossing.

To begin, we will consider only a single graph G with m edges. Independently for each vertex, flip a fair coin to decide whether the vertex is in A or B. Let X be the random variable which denotes the number of edges crossing between A and B and for each $e \in E(G)$, let X_e be 1 if e crosses between A and B and 0 otherwise. Of course, $X = \sum_{e \in E(G)} X_e$. In the homework, you verified that $\mathbb{E}X_e = \frac{1}{2}$, so $\mathbb{E}X = \frac{m}{2}$. Now let's calculate $\mathrm{Var}(X)$.

We begin by noting that

$$\begin{aligned}
\mathbb{E}[X_e X_s] &= \Pr[e \text{ crosses and } s \text{ crosses}] \\
&= \Pr[e \text{ crosses} | s \text{ crosses}] \Pr[s \text{ crosses}] \\
&= \begin{cases} \frac{1}{2} & \text{if } e = s \\ \frac{1}{4} & \text{otherwise.} \end{cases}
\end{aligned}$$

As such,

$$\mathrm{Var}(X) = \sum_{e,s \in E(G)} \left(\mathbb{E}[X_e X_s] - \mathbb{E}X_e \mathbb{E}X_s\right) = \sum_{e \in E(G)} \left(\frac{1}{2} - \frac{1}{4}\right) + \sum_{e \neq s} \left(\frac{1}{4} - \frac{1}{4}\right) = \frac{m}{4}.$$

By Chebyshev's inequality, for any $\lambda > 0$,

$$\Pr\left[\left|X - \frac{m}{2}\right| \geq \lambda\sqrt{\frac{m}{4}}\right] \leq \frac{1}{\lambda^2},$$

so by taking only one side of the absolute value,

$$\Pr\left[X \leq \frac{m}{2} - \lambda\sqrt{\frac{m}{4}}\right] \leq \frac{1}{\lambda^2}.$$

Now that we have done this calculation, we return to the case of multiple graphs. Again partition the common vertex set of $G_1, \dots, G_k$ as before and let $X^{(i)}$ be the random variable which denotes the number of edges crossing between A and B in G_i. As each $X^{(i)}$ is distributed according to the X from earlier, we can apply the union bound to find,

$$\begin{aligned}
\Pr\left[\bigvee_{i=1}^{k} \left(X^{(i)} \leq \frac{m}{2} - \lambda\sqrt{\frac{m}{4}}\right)\right] &\leq \sum_{i=1}^{k} \Pr\left[X^{(i)} \leq \frac{m}{2} - \lambda\sqrt{\frac{m}{4}}\right] \\
&\leq \sum_{i=1}^{k} \frac{1}{\lambda^2} = \frac{k}{\lambda^2}
\end{aligned}$$

Choosing $\lambda > \sqrt{k}$, this probability is strictly less than 1. Hence, for any $c > \frac{\sqrt{k}}{2}$, there is a positive probability that every G_i has at least $\frac{m}{2} - c\sqrt{m}$ edges crossing the partition. □

Let's end our discussion of discrete probability with a fun little problem. A birthday cake starts with n lit candles. Uniformly at random select a number k between 1 and n and blow out any k of the candles. Now there are $n-k$ candles lit, so uniformly at random select a number between 1 and $n-k$ and blow out that many candles. Repeat this process until all candles have been blown out. Let X_n be the random variable denoting how many turns it takes to blow out all n candles. What is $\mathbb{E}X_n$? Naïvely, we would expect $\mathbb{E}X_n$ to be logarithmic in n as at each stage, we expect to blow out around half the remaining candles. This intuition is correct as we will see. Firstly, $X_0 = 0$ always, so $\mathbb{E}X_0 = 0$. $X_1 = 1$ as we will always blow out the only candle there, so $\mathbb{E}X_1 = 1$. On the other hand, by the law of total probability, we can condition the expected value on the outcome of the first selected number (let Y be the random variable denoting the value of this number). We find that

$$\begin{aligned}\mathbb{E}X_n &= 1 + \sum_{i=1}^{n} \mathbb{E}[X_n | Y = i] \Pr[Y = i] \\ &= 1 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_{n-i} = 1 + \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}X_i.\end{aligned}$$

Let's make the guess that in general $\mathbb{E}X_n = H_n$, which is reasonable as we expect $\mathbb{E}X_n$ to be logarithmic and we have the above recurrence. Certainly $\mathbb{E}X_0 = H_0$ and $\mathbb{E}X_1 = H_1$, so suppose that $\mathbb{E}X_i = H_i$ for all $i < n$. Then

$$\begin{aligned}\mathbb{E}X_n &= 1 + \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}X_i = 1 + \frac{1}{n}\sum_{i=1}^{n-1} H_i \\ &= 1 + \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{i} \frac{1}{j} = 1 + \frac{1}{n}\sum_{j=1}^{n-1} \frac{n-j}{j} \\ &= 1 + \sum_{j=1}^{n-1} \frac{1}{j} - \frac{n-1}{n} = H_{n-1} + \frac{1}{n} = H_n.\end{aligned}$$

As such, it is the case that $\mathbb{E}X_n = H_n \sim \log n$ as we predicted.