THE PROBABILISTIC METHOD

MATEO BRODY

Abstract. This expository paper discusses the application of the probabilistic method across various areas in combinatorics, elucidating the core principles of the method through a number of examples. Starting with introductory problems, we will build increasingly complex techniques and prove a collection of important theorems and lemmas. Ultimately, we will conclude with a brief discussion of how the probabilistic method facilitates structural understanding despite being inherently non-constructive.

1. Introduction

Questions in Combinatorics themselves with proving existence of objects satisfying a certain set of properties. Often, direct proof of the existence of such objects is complicated; explicit constructions may be difficult to generalize or be sufficiently complex that finding them is challenging. **The Probabilistic Method** gives an alternative means of proof: one considers a broader set of objects and, through argument, shows that a random object satisfies the desired properties with non-zero probability–thus implying the existence of the desired object. Notably, the probabilistic method is entirely non-constructive: nowhere in the process is the object explicitly constructed.

In effect, the probabilistic method is based on a relatively simple idea. First developed by Paul Erdos, the method stems from the following preposition: if a randomly chosen object satisfies a property with nonzero probability, then there must necessarily exist some object satisfying the property.

In many cases, demonstrating the above proves to be significantly easier than explicit construction. One such example is in Ramsey Numbers.

Definition 1.1. The Ramsey Number $R(s,t)$ is the minimum number of vertices n such that any two-coloring of the edges E on a complete graph of n vertices will contain a monochromatic clique of size s or of size t.

In fact, Erdos first conceived of the Probabilistic Method in order to bound certain families of Ramsey Numbers. His first use of the method in 1947 was to produce the following:

Theorem 1.2. *Erdos, 1947: For $n, r \in \mathbb{Z}^+$, if $\binom{n}{r}2^{1-\binom{r}{2}} < 1$, then $R(r,r) > n$*

Proof. Erdos' method involved imagining a graph where every edge was colored red or blue uniformly and independently at random. He then considered every clique K_r; using the union bound he showed that, if $\binom{n}{r}2^{1-\binom{r}{2}} < 1$, then with positive probability the coloring gave no monochromatic K_r. From this he concluded that there existed some coloring on n vertices such that no monochromatic K_r existed, and thus by definition, $R(r,r) > n$. We will rigorously prove this theorem in section 2. Right now, the important takeaway is that

Date: July 13, 2026.

through this probabilistic argument, we are able to prove the existence of a desired coloring – despite having no knowledge as to how one might construct it. ■

Since Erdos' creation of the method, the idea has been refined and adapted into numerous techniques and important theorems. Through these increasingly sophisticated tools, the probabilistic method has been gradually implemented to produce stronger bounds and tackle problems of greater complexity. Developed and explored by Alon and Spencer in their work *The Probabilistic Method,* [AS11] the technique is now an integral problem-solving tool across Combinatorics.

In this paper, we will discuss and gradually build a number of key techniques in the probabilistic method. Several interesting problems, applications, and examples will be provided to illustrate these methods. Although the probabilistic method is a rich and vast means of tackling questions in Combinatorics, this paper will focus namely on:

- The First Moment Method
- The Bollobas Two Families Theorem
- Alterations
- The Second Moment Method
- The Lovasz Local Lemma
- Martingale Inequalities
- Dependent Random Choice
- Containers

2. Preliminaries

We adopt the following basic conventions:

- $[n]$ is used to denote the set $\{1, 2, ..., n\}$
- Unless specified, the base of a particular logarithm is assumed to be e.

Moreover, a number of later examples will involve asymptotic analysis. For this, we introduce certain asymptotic notation. Each item below describes the behavior of functions f and g as the parameter X tends to infinity.

Definition 2.1. We say that a particular function $g(X)$is $o(f(X))$ if $\lim_{X\to\infty} \frac{g(X)}{f(X)} = 0$. We use this interchangeably with $f(X) \ll g(X)$, which has the same meaning.

Definition 2.2.
For functions $f(X)$ and $g(X)$, $f(X) \lesssim g(X)$ iff $f(X) \leq Cg(X)$ as X tends to infinity for some real constant $C > 0$.

Definition 2.3.
whp (with high probability) indicates probability $p = 1 - o(1)$

The Probabilistic Method borrows techniques from Probability Theory. Naturally, we reference a number of conventions and definitions relating to probability. While more advanced techniques will be introduced in subsequent sections, sufficient prerequisite knowledge on basic probability is required. In particular, we introduce the following definitions:

Definition 2.4. An Event A is a particular outcome or set of outcomes of some random experiment. We denote $\mathbb{P}(A)$ as the probability that event A occurs.

Definition 2.5. For two events A_1, A_2, define $\mathbb{P}(A_1 \wedge A_2)$ as the probability that both A_1 and A_2 occur. In general, let

$$\mathbb{P}(\bigwedge_{i=1}^{n} A_i)$$

denote the probability that events $A_1, A_2, ..., A_n$ occur. Moreover, let $\mathbb{P}(A_1 \cup A_2)$ be the probability that at least one of A_1 or A_2 occur.

Definition 2.6. For an event A, we let $\overline{A}$ denote the event that A does not occur. Since precisely one event out of $\{A, \overline{A}\}$ must occur, it follows that $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.

A basic understanding of indicator variables is also needed.

Definition 2.7. For an event A, we define the *indicator variable* X_A as a random variable that equals 1 when A occurs, and 0 otherwise. Observe that

$$\mathbb{E}[X_A] = \mathbb{P}(A)$$

since $\mathbb{E}[X_A]$ is just $\Pr(A \text{ occurs}) \times 1 + \Pr(A \text{ does not occur}) \times 0$

Later sections will also denote the notion of dependencies, for which we define the following:

Definition 2.8. For events A and B, let $\mathbb{P}(A|B)$ be the probability that A occurs, given that B occured. Explicitly, we define

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(B)}$$

The probabilistic method is frequently applied to the study of graphs, and hence we define some basic Graph Theory conventions.

Definition 2.9. Let v be a vertex in graph $G = (V, E)$. The neighbors of v, denoted $N(v)$, is the set of vertices which are connected to v by an edge.

For a particular $S \subset V(G)$, we let $N(S)$ be the set of vertices such that $a \in N(S)$ iff $a \in N(v)$ for all $v \in S$.

Definition 2.10. For some graph $G = (V, E)$, a subgraph is a graph $H = (V', E')$ satisfying $V' \subseteq V$ and $E' \subseteq E$, such that any edge in E' connects two vertices in V'.

Definition 2.11. A *clique* is a subgraph of some graph $G = (V, E)$ with n vertices and $\binom{n}{2}$ edges. That is, for any two vertices in the clique, there exists an edge connecting them.

3. Basic Applications and the First Moment

To reiterate, the probabilistic method proves the existence of a desirable object through non-constructive means. To do so, we formulate a random construction and show it satisfies the properties of the desired object with non-zero probability. We illustrate how this method may be applied with a basic example.

Definition 3.1. A graph is *bipartite* if the vertices can be partitioned into two disjoint sets such that every edge connects a vertex in one set to a vertex in the other.

Theorem 3.2. *Every graph with m edges has a bipartite subgraph with at least $\frac{m}{2}$ edges*

Proof. Define the graph $G = (V, E)$. Assign a random 2-coloring to the graph G by coloring every vertex $x \in V$ black or white, uniformly and independently at random. Denote the set $A \subseteq E$ as the set of edges with one black endpoint and one white endpoint. Then (V, A) is a bipartite subgraph of G, since the edges A exclusively connect from the set of black vertices to the set of white vertices.

Consider some edge k. Since every vertex is colored independently and uniformly at random, the probability that $k \in A$ is $1/2$. Let I_k be the indicator variable that is 1 if $k \in A$ and zero otherwise. Then

$$|A| = \sum_{k \in E} I_k$$

and by Linearity of Expectation,

$$\mathbb{E}[|A|] = \sum_{k \in E} \mathbb{E}[I_k] = \sum_{k \in E} \mathbb{P}(k \in A) = 1/2|E|$$

Observe that with positive probability, $X \geq \mathbb{E}[X]$ for a random variable X. Therefore there must exist some random coloring where the set A satisfies $A \geq \mathbb{E}[|A|] = 1/2|E|$. Then, (V, A) is the desired subgraph of G. ■

In essence, the argument in Proof 3.2 begins by selecting a random vertex-coloring of the graph and then analyzing the properties of the resulting coloring. In particular, we showed that the expected number of edges in A is $\frac{1}{2}|E|$. Using this expectation, we proved that the random coloring satisfies the desired inequality,

$$|A| \geq \frac{1}{2}|E|$$

with positive probability. Consequently, it followed that among all possible vertex-colorings, there must exist at least one coloring that satisfies the required condition.

Although applications of the probabilistic method are typically more involved, they generally retain this flavor–one constructs a random object, makes an observation about the object, and then uses this observation to show that it satisfies a certain set of desired conditions with positive probability.

In Theorem 3.2, our observations leveraged the fact that $X \geq \mathbb{E}[X]$ with positive probability. Likewise, $X \leq \mathbb{E}[X]$ with positive probability. These inequalities are sufficiently powerful that we formalize the into a proposition, to be referenced later.

Proposition 3.3. *For a random variable X, $X \geq \mathbb{E}[X]$ and $X \leq \mathbb{E}[X]$ with non-zero probability.*

if $X < \mathbb{E}[X]$ always, then $\mathbb{E}[X] < \mathbb{E}[X]$, a contradiction, from which the proposition follows. An analogous proof shows $X \leq \mathbb{E}[X]$ is always possible.

This proposition is one of many techniques that let us use expected value to make particular conclusions about probability. One of the most important of these techniques is known as the **First Moment Method**:

Theorem 3.4. ***The First Moment Method:*** *For a discrete non-negative random variable X, if $\mathbb{E}[X] < 1$, then $X = 0$ with positive probability.*

Proof. Suppose for the sake of contradiction that $X = 0$ occurs with probability 0.

Define the set $Q = \{p_0, p_1, p_2, ..., p_n, ...\}$ where p_i is the probability that X attains value i. Recall that by definition,

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} ip_i$$

Since $i \geq 1$, it follows that

$$\sum_{i=1}^{\infty} ip_i \geq \sum_{n=1}^{\infty} p_i$$

But note that the RHS is just the probability that $X > 0$, so

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} ip_i \geq \mathbb{P}(X > 0) = 1$$

Thus, $\mathbb{E}[X] \geq 1$, contradiction. It follows that if $\mathbb{E}[X] < 1$, then $X = 0$ with positive probability. ■

In practice, we apply the first moment method to perform an argument analogous to complementary counting: rather than considering "good" or desirable characteristics, we take a random construction and observe the number of "bad" properties that we wish to avoid. If we can show that the expected number of "bad" properties is less than one, the First Moment Method guarantees that some construction has no bad properties.

With this in mind, we revisit Theorem 1.1 from the Introduction.

Theorem 3.5. *Erdos, 1947: For $n, r \in \mathbb{Z}^+$, if $\binom{n}{r}2^{1-\binom{r}{2}} < 1$, then $R(r, r) > n$*

Remark 3.6. Recall that the Ramsey number $R(r, r)$ is the minimum number of vertices n such that any two-coloring of the edges E on a complete graph of n vertices will contain a monochromatic clique of size r.

Proof. Let $G = (V, E)$ be our complete graph with $|V| = n$. Assign a random 2-coloring to G by coloring each edge $e \in E$ Black or White, uniformly and independently at random.

Consider an arbitrary subset of r vertices. Denote X_K as the event that this particular subset is a monochromatic clique; that is, all $\binom{r}{2}$ edges connecting vertices in the subset are of the same color. These monochromatic cliques are the "bad" properties that we wish to avoid.

Out of $2^{\binom{r}{2}}$ possible colorings of these edges, only two give a monochromatic clique K. It follows that

$$\mathbb{P}(X_k) = \frac{2}{2^{\binom{r}{2}}} = 2^{1-\binom{r}{2}}$$

Let X be the total number of monochromatic cliques K. There are a total of $\binom{n}{r}$ subsets of the n vertices with size r. Hence by linearity of expectation,

$$\mathbb{E}[X] = \binom{n}{r}\mathbb{P}(X_K) = \binom{n}{r}2^{1-\binom{r}{2}}$$

If $\binom{n}{r}2^{1-\binom{r}{2}} = \mathbb{E}[X] < 1$, then the First Moment Method gives that there must exist some coloring with no monochromatic K. By definition, this forces $R(r, r) > n$.

■

Another subject of frequent study in Combinatorics is the enumeration of sets satisfying certain constraints. In particular, sum-free sets have been a popular area of inquiry.

Definition 3.7. A subset S of the real numbers $\mathbb{R}$ is sum-free if there are no elements $x, y, z \in S$ such that $x + y = z$.

Here again, the Probabilistic Method can provide some interesting results.

Theorem 3.8. *Every set of k non-zero integers has a sum-free subset of size $\geq k/3$*

Proof. Denote our set of integers as A, with $|A| = k$. Let X be a variable chosen uniformly at random from $[0, 1]$. Construct a new subset A' such that

$$A' := \{a \in A : \{aX\} \in (1/3, 2/3)\}$$

That is, we include an element a of A in A' if the fractional part of aX is between $1/3$ and $2/3$.

We claim that A' is necessarily sum-free. For any three particular elements a_x, a_y, and a_z in A', consider the quantities a_xX, a_yX, and a_zX. By definition, all three of a_xX, a_yX, and a_zX must have fractional parts between $1/3$ and $2/3$. Observe that summing two fractional parts in the domain $(1/3, 2/3)$ never yields a fractional part in this domain. Therefore it is necessarily true that

$$a_xX + a_yX \neq a_zX$$

Dividing out the constant X, it follows that $a_x + a_y \neq a_z$. Performing this argument for all choices of a_x, a_y, and a_z in A' proves that A' is always sum-free.

Since X is chosen uniformly from $[0, 1]$, aX is also uniformly random between $[0, 1]$. Thus for any element $a \in A$, $P(a \in A') = 1/3$. By linearity of expectation,

$$\mathbb{E}[|A'|] = 1/3k$$

so by Proposition 3.3, there must exist some X such that $|A'| \geq 1/3k$. This set A' is our desired sum-free subset. ■

4. Set Families

In Section 3, a majority of the results hinged on Proposition 3.3 and Theorem 3.4 and bounded a particular quantity based on its expected value. However, leveraging expected value does not necessarily yield particularly strong bounds. For instance, suppose we wished to find the maximum of some variable X. By applying Proposition 3.3, we may only show $\max(X) \geq \mathbb{E}[X]$. But if X is on occasion significantly larger than $\mathbb{E}[X]$, this bound is weak. In effect, we require the maximal X to be near $\mathbb{E}[X]$ for this method to be reasonably applied. Combinatorially, this translates to desirable (in this example, maximal) objects being frequent and similar to the expected case. Hence, the first moment method is applicable only when objects are frequent but difficult to characterize through explicit construction (informally, these scenarios have often been characterized as "trying to find hay in a haystack"). This is the case in certain questions studying Set Families.

Definition 4.1. A Set family K is a collection of subsets of some particular "ground" set, typically $[n]$. Explicitly, every element of K is a subset of the ground set.

Probabilistic methods have often been used to bound and constrain set systems satisfying particular properties. One of the strongest results is the Bollobas two family theorem:

Theorem 4.2. ***Bollobas Two Family Theorem:*** *Let $A_1, A_2...A_m$ and $B_1, B_2....B_m$ be sets such that $A_i \cap B_i = \emptyset$ and $A_i \cap B_j \neq \emptyset$ for all $0 < i, j \leq m$, $i \neq j$. Then*

$$\sum_{i=1}^{m} \binom{|A_i| + |B_i|}{|A_i|}^{-1} \leq 1$$

Proof. Suppose that for particular sets $A_1, A_2...A_m$ and $B_1, B_2....B_m$, the hypothesis of the Bollobas Two Families theorem holds. That is, $A_i \cap B_i = \emptyset$ and $A_i \cap B_j \neq \emptyset$ for all $0 < i, j \leq m$, $i \neq j$.

Let $S = (\bigcup_{i=1}^{m} A_i) \cup (\bigcup_{i=1}^{m} B_i)$. In other words, S is the set of all elements contained in any of $A_1, A_2...A_m$ or $B_1, B_2..., B_m$. Consider a random permutation of the elements of S.

Denote P_i as the event that all the elements of A_i precede the elements of B_i in the permutation. By the hypothesis, the sets A_i and B_i are disjoint. In a random permutation, any order in which the elements in A_i and B_i appear in the permutation is equally likely. In total, there are

$$\binom{|A_i| + |B_i|}{|A_i|}$$

potential orderings. Among these orderings, there exists only one in which all elements in A_i precede all elements in B_i. Hence, it follows that

$$\mathbb{P}[P_i] = \binom{|A_i| + |B_i|}{|A_i|}^{-1}$$

.

We continue the proof with the following lemma:

Lemma 4.3. *if sets $A_1, A_2...A_m$ and $B_1, B_2....B_m$ satisfy the hypothesis of the Bollobas Family Theorem, then for all $i \neq j$, only one of P_i and P_j may occur in some permutation.*

Proof. Suppose for the sake of contradiction that P_i and P_j both occur in a particular random permutation. That is, A_i precedes B_i and A_j precedes B_j.

By definition, there must exist at least one element q such that $q \in A_i$ and $q \in B_j$. Similarly, there must exist at least one element r with $r \in A_j$ and $r \in B_i$.

Since A_i precedes B_i, $q \in A_i$ must precede $r \in B_i$. But since A_j precedes B_j, $r \in A_j$ must precede $q \in B_j$–hence, we arrive at a contradiction. It follows that only one such P_i may occur for a particular permutation. ■

Thus by Lemma 4.3, no P_i and P_j can occur simultaneously; i.e. the events are disjoint. If the sum across all i of $\mathbb{P}(P_i)$ was greater than one, then there must necessarily exist a permutation where P_i and P_j occur. But this is impossible by Lemma 4.3, so the inequality

$$\sum_{i=1}^{m} \mathbb{P}(P_i) = \sum_{i=1}^{m} \binom{|A_i| + |B_i|}{|A_i|}^{-1} \leq 1$$

follows as desired. ■

The Bollobas Two Family Theorem is an incredibly powerful problem-solving technique. In particular, it resolves a number of questions involving hypergraphs.

Definition 4.4. A Hypergraph is a graph $G = (V, E)$ where each edge $e \in E$ may connect between 1 and $|V|$ vertices. Note that E is a Set Family of subsets of vertices V.

Definition 4.5. An n-uniform hypergraph is a hypergraph $G = (V, E)$ where for each edge $e \in E$, $|e| = n$. For instance, a "standard" graph is effectively a 2-uniform hypergraph.

Definition 4.6. A hypergraph $G = (V, E)$ is *covered* by $M \subseteq V$ if for every edge $e \in E$, e overlaps with at least one vertex in M.

Theorem 4.7. *If every r-uniform hypergraph with at most $\binom{r+s}{r}$ edges can be covered by a set M of size s, then it is necessarily true that every r-uniform hypergraph, regardless of size, is covered by M.*

Proof. Let M be a set of vertices with $|M| = s$. Let M cover every r-uniform hypergraph with at most $\binom{r+s}{r}$ vertices.

Suppose for the sake of contradiction that there exists an r-uniform hypergraph G such that M does not cover G. Consider the procedure of iteratively removing edges from G until M suffices to cover the remaining graph.

We say that an r-uniform hypergraph graph is *critical* if the graph is not covered by M, but we may remove precisely one edge so that the remaining graph is covered by M.

At some point during the process of iterative removal, G must necessarily be critical. We denote Q as the number of edges in G at the moment when it is critical. Since we assumed that M covers every hypergraph with at most $\binom{r+s}{r}$ vertices, it follows that

$$Q > \binom{r+s}{r}$$

since by definition, M does not cover Q. Let the edges in G at the moment it is critical be E_1, E_2, E_3... E_q; in particular, let E_1 be the edge that may be removed so that the remaining graph is covered by M. Now, set $S_1 = M$, and $S_i = M/E_i$ for all $2 \leq i \leq Q$. Then the families $E_1, ...E_Q$ and $S_1...S_Q$ satisfy the hypothesis of the Bollobas Two Family Theorem. Explicitly,

$$E_i \cap S_i = \emptyset \text{ and } E_i \cap S_j \neq \emptyset \text{ for all } i \neq j$$

.

Applying the Two Family Theorem,

$$\sum_{i=1}^{Q} \binom{|E_i| + |S_i|}{|S_i|}^{-1} \leq 1$$

but $|E_i|$ is always r since the graph is r-uniform and we let $|S_i| = s$, so

$$\sum_{i=1}^{Q} \binom{|E_i| + |S_i|}{|S_i|}^{-1} = Q\binom{r+s}{s}^{-1} \leq 1$$

This implies $Q \leq \binom{r+s}{s}$, but we have already shown $Q > \binom{r+s}{s}$, contradiction.

It follows that there cannot exist any critical graph, and hence all graphs are covered by M. ■

For a more visual proof of Theorem 4.7, see [Pen17]

5. Alterations

Frequently, random probabilistic constructions come close to satisfying a particular property, but don't fully do so. Consequentially, very basic probabilistic method arguments are unable to produce strong results. Consider the following example:

Definition 5.1. In graph $G = (V, E)$, A subset of vertices $U \subseteq V$ is dominating if every vertex $v \in V$ with $v \notin U$ has a neighbor in U.

Question 5.2. *Given a graph $G = (V, E)$ with minimum degree θ, what is the smallest dominating subset U?*

Suppose we were to attack this question with an analogous approach to what has been done in previous sections. Following the line of reasoning of sections 2 and 3, we might construct a random vertex subset of size k and hope that, miraculously, we can show that the expected number of vertices with no neighbors in the subset is less than 1. Similar approaches might involve considering the probability that a particular vertex is excluded.

Unfortunately, these methods produce relatively weak bounds. Observe that showing that the expected number of vertices with no neighbors is less than 1 is analogous to showing that the average subset of size k is dominating. Yet, in this problem, one can intuitively envision that the best possible constructions are rare, and that much larger sizes k are needed until the average random subset catches up.

Hence, we desire a way to change our approach slightly. Rather than pursuing this strictly random construction, it is natural to apply the **Alteration Method**. Alteration (or Deletion) involves the construction of a random object which is then manually "fixed" to avoid any undesirable elements.

In practice, Alteration takes a very (small/large, depending on the goal) yet imperfect construction, and then weakens it by patching it up until it is the object that we want. Applying this method yields the following result:

Theorem 5.3. *For any graph $G = (V, E)$ on n vertices with minimum degree $\theta > 1$, there exists a dominating set with size*

$$\leq \left(\frac{\ln(\theta + 1) + 1}{\theta + 1}\right)n$$

Proof. We begin similarly to in previous questions; that is, we introduce a certain randomness to our construction.

Let $p \in [0, 1]$. We will decide what choice of p to make later. Denote $X \subseteq V$ as a random subset where each vertex is independently included with probability p.

Let Y be the set of vertices that are not in X and have no neighboring vertices in X. Consider an arbitrary vertex v; it is not in X with probability $1 - p$. Moreover, v has at least θ neighbors that are not in X with probability $1 - p$. Hence,

$$\mathbb{P}[v \in Y] \leq (1 - p)^{\theta+1}$$

Evidently, the set $X \cup Y$ is dominating. By definition, X and Y are disjoint sets, so $\mathbb{E}[|X \cup Y|] = \mathbb{E}[|X|] + \mathbb{E}[|Y|]$. By linearity of expectation,

$$\mathbb{E}[|X|] = np; \text{ and } \mathbb{E}[|Y|] \leq (1 - p)^{\theta+1}n$$

We continue to simplify with the following lemma:

Lemma 5.4. *for $p \in [0, 1]$ and $\theta > 2$, $(1 - p)^{\theta+1} \leq e^{-p(\theta+1)}$*

Proof. We implement the standard inequality

$$1 + x \leq e^x$$

for all real x. Setting $x = -p$ and raising both sides to $\theta + 1$, we obtain

$$(1-p)^{\theta+1} \leq e^{-p(\theta+1)}$$

as desired. ■

By Lemma 5.4,

$$\mathbb{E}[|X|] + \mathbb{E}[|Y|] \leq (p + e^{-p(1+\theta)})n$$

Setting $p = \frac{\ln\theta+1}{\theta+1}$ gives the desired inequality

$$\mathbb{E}[|X|] + \mathbb{E}[|Y|] \leq \big(\frac{\ln(\theta+1)+1}{\theta+1}\big)n$$

■

The "alteration" in this proof arises from the combination of the vertices in Y and the random set X to form a new, necessarily dominating set $X \cup Y$. In effect, alterations are the reintroduction of order and structure into a random construction by means of manually changing things around. This hybrid approach–which combines both randomness and order–is often stronger than purely random results. We present another result produced by alteration:

Definition 5.5. Let A be a matrix. Let S_x be a subset of column indices, and S_y be a subset of the row indices. The **Submatrix** B consists of all entries $a_{i,j}$, where $i \in S_y$, and $j \in S_x$

Theorem 5.6. *For every positive integers $n \geq k \geq 2$, there exists an $n \times n$ matrix with $\{0,1\}$ entries, with at least $n^{\frac{2k}{k+1}}\big(1 - \frac{1}{k!^2}\big)$ 1's, such that there is no $k \times k$ submatrix consisting of all $1's$*

Proof. Again, let $p \in [0,1]$ to be decided later. Define Q as a random n by n matrix with $\{0,1\}$ entries formed by setting $a_{i,j} = 1$ with probability p independently for all $1 \leq i, j \leq n$. Denote X as the number of ones in Q. Then

$$\mathbb{E}[X] = n^2 p.$$

For a particular $k \times k$ submatrix Z, denote E_z as the event that Z contains all 1's. Since every one of the k^2 entries are chosen independently, it follows that $\mathbb{P}(E_z) = p^{k^2}$. Note that there are $\binom{n}{k}^2$ possible submatrices. Hence, letting Y be the set of submatrices with all 1's,

$$\mathbb{E}[|Y|] = \sum_{k\times k \text{ submatrix } Z} \mathbb{P}[E_z] = \binom{n}{k}^2 p^{k^2}$$

Consider the following procedure: Create a new matrix Q' by "deleting" (changing to 0) one entry in every $k \times k$ sub-matrix of Q that contains all ones. Then Q' has no $k \times k$ sub-matrices with all ones. Letting $\#(1(Q')$ denote the number of 1s in Q', since deleting a single entry may eliminate several $k \times k$ submatrices with all $1's$ simultaneously,

$$\#1(Q') \geq X - |Y|$$

Hence by taking the expectation of both sides,

$$\mathbb{E}[\#1(Q')] \geq \mathbb{E}[X - |Y|] = \mathbb{E}[X] - \mathbb{E}[|Y|] = n^2p - \binom{n}{k}^2 p^{k^2}$$

Setting $p = n^{-2/(k+1)}$ gives

$$\mathbb{E}[X - |Y|] = n^{2k/(k+1)} - \binom{n}{k}^2 n^{-2k^2/(k+1)}$$

Factoring $n^{2k/(k+1)}$ from both terms,

$$\mathbb{E}[X - |Y|] = n^{2k/k+1}\big(1 - \binom{n}{k}^2 n^{-2k}\big)$$

Note that $\binom{n}{k} = \frac{n\times(n-1)\times(n-2)....\times(n-k+1)}{k!} \leq \frac{n^k}{k!}$. Hence $\binom{n}{k}^2 n^{-2k} \leq \frac{1}{k!^2}$. Substituting this in, we arrive at

$$\mathbb{E}[\#1(Q')] \geq \mathbb{E}[X - |Y|] \geq n^{2k/k+1}\big(1 - \frac{1}{k!^2}\big)$$

By Proposition 3.3, there must exist some matrix Q' with at least $\mathbb{E}[X - |Y|]$ 1's and no $k \times k$ sub-matrices with all 1's; this is our desired matrix.

■

6. The Second Moment Method

Thus far, our results have generally leveraged some particular expected value expression to prove the existence of a certain object. However, expected value is often not sufficient to make strong conclusions about a bound. Specifically, considering expected value alone reveals nothing about the *structure* of the distribution for a variable X. A number of issues arise from this, especially when we desire to show that a variable X rarely attains a value k.

For instance, consider the scenario in which a non-negative discrete variable X_n has $\mathbb{E}[X_n] \to \infty$ for sufficiently large n. We might wish to show that X_n takes on a value greater than 0 *whp* (recall that *whp* is $1 - o(1)$). However, expectation alone does not guarantee this; $\mathbb{E}[X_n] \to \infty$ is not sufficient to conclude that X_n is rarely zero, since it is still possible that X_n is often zero but on occasion infinitely large.

To address these issues, it is natural to bring in techniques from probability theory that describe the concentration and density of a variable X around the mean–a high concentration would imply that outliers are unlikely. In particular, we use the **Variance** of a variable X.

Definition 6.1. The *variance* of random variable X is

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

The *covariance* of random variables X, Y is

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Remark 6.2. Since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for independent variables X, Y, the covariance is zero if two variables are independent.

Intuitively, covariance describes how variables X, Y tend to change together; that is, whether an increase in X tends to imply an increase/decrease in Y or vice versa. $\mathbb{E}[XY]$ tends to grow significantly larger than $\mathbb{E}[X]\mathbb{E}[Y]$ when Y increases with X, and grows significantly smaller when Y decreases with increases in X.

Moreover, it's important to note that by definition, Cov$[X, X]$=Var$[X]$.

Remark 6.3. We denote the mean as μ and the variance as σ^2. The Standard Deviation is σ.

The usage of covariance and variance to show a variable tends to cluster near its mean is called the **Second Moment Method**. Using the following results, we may apply the Second Moment Method to show that a variable is concentrated near its mean:

Lemma 6.4. ***Markov's Inequality****: If $X \geq 0$ is a nonnegative random variable, then for all a, $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$.*

Proof. Let $X_a = X$ if $X \geq a$ and 0 otherwise. Then

$$\mathbb{E}[X] \geq \mathbb{E}[X_a]$$

Let $Y = a$ if $X_a \geq a$ and 0 otherwise. Observe that $\mathbb{E}[Y] = a\mathbb{P}[X \geq a]$. It follows that

$$\mathbb{E}[X] \geq \mathbb{E}[X_a] \geq \mathbb{E}[Y] = a\mathbb{P}[X \geq a]$$

so $\mathbb{E}[X] \geq a\mathbb{P}[X \geq A]$ and $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$ as desired. ■

With this inequality in mind, we produce a result involving the second moment. The **Chebyshev Inequality** formalizes the idea that small variances imply that large deviations are improbable. In particular, it shows that the probability of attaining a value at least λ standard deviations away from the mean is inversely quadratic in λ.

Theorem 6.5. ***Chebyshev's Inequality:*** *Let X be a random variable with mean μ and variance σ^2. Then for any $\lambda \geq 0$,*

$$\mathbb{P}[|X - \mu| \geq \lambda\sigma] \leq \lambda^{-2}$$

Proof. Observe that we may square both sides of the inequality in $\mathbb{P}[|X - \mu| \geq \lambda\sigma]$ to show

$$\mathbb{P}[|X - \mu| \geq \lambda\sigma] = \mathbb{P}[|X - \mu|^2 \geq \lambda^2\sigma^2]$$

By Markov's Inequality,

$$\mathbb{P}[|X - \mu|^2 \geq \lambda^2\sigma^2] \leq \frac{\mathbb{E}[|X - \mu|^2]}{\lambda^2\sigma^2}$$

Since $\mu = \mathbb{E}[X]$, $\mathbb{E}[|X - \mu|^2]$ is just σ^2, by definition. Substituting this in, we have

$$\mathbb{P}[|X - \mu| \geq \lambda\sigma] = \mathbb{P}[|X - \mu|^2 \geq \lambda^2\sigma^2] \leq \frac{1}{\lambda^2}$$

as desired. ■

Corollary 6.6. *For any random variable X,*

$$\mathbb{P}(X = 0) \leq \frac{Var[X]}{\mathbb{E}[X]^2}$$

Proof. Writing $\mu = \mathbb{E}[X]$,

$$\mathbb{P}[X = 0] \leq \mathbb{P}(|X - \mathbb{E}[X]| \geq \mathbb{E}[X])$$

since the RHS is just the probability that $X \leq 0$ or $x \geq 2\mathbb{E}[X]$. But the RHS is also the LHS of the Chebyshev inequality for $\lambda = \mathbb{E}[X]/\sigma$, so

$$\mathbb{P}[X = 0] \leq \mathbb{P}(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$$

■

The following corollary is among the stronger parts of probabilistic machinery in the Second Moment Method. If we can deduce the variance of a sequence of variables, this corollary allows us to determine the asymptotic behavior of the sequence.

Corollary 6.7. *Let $X_1, X_2 ... X_n ...$ be a sequence of nonnegative random variables. if $\mathbb{E}[X_n] > 0$ and $Var(X) = o(\mathbb{E}[X_n^2])$ for sufficiently large n, then $X_n > 0$ whp and X_n grows asymptotically with $\mathbb{E}[X_n]$ whp*

Proof. The former statement ($X_n > 0$ *whp*) is guaranteed by Corollary 6.6, since

$$\mathbb{P}[X_n = 0] \leq \frac{\mathrm{Var}[X_n]}{\mathbb{E}[X_n]^2} = o(1)$$

Thus

$$\mathbb{P}[X > 0] = 1 - \mathbb{P}[X = 0] \geq 1 - o(1)$$

The second statement (X_n grows asymptotically to $\mathbb{E}[X_n]$) follows from the Chebyshev inequality. Setting $\lambda = \varepsilon\mathbb{E}[X_n]/\sigma$ and noting $\mu = \mathbb{E}[X_n]$, Theorem 6.5 gives

$$\mathbb{P}(|X - \mathbb{E}[X_n]| \geq \varepsilon\mathbb{E}[X_n]) \leq \frac{\sigma^2}{\varepsilon^2\mathbb{E}[X_n]^2}$$

We know that $\sigma^2 = \mathrm{Var}(X_n) = o(\mathbb{E}[X_n^2])$, so by definition,

$$\sigma^2/\mathbb{E}[X_n]^2 = o(1)$$

Substituting this back in, it follows that

$$\mathbb{P}(|X_n - \mathbb{E}[X_n]| \geq \varepsilon\mathbb{E}[X_n]) \leq \frac{1}{\varepsilon^2}o(1)$$

For all epsilon, the constant $1/\varepsilon^2$ preserves the asymptotic $o(1)$, thus

$$\mathbb{P}(|X_n - \mathbb{E}[X_n]| \geq \varepsilon\mathbb{E}[X_n]) \leq o(1)$$

Considering arbitrarily small ε, it follows that X deviates asymptotically from $\mathbb{E}[X_n]$ with probability $o(1)$, and thus X is asymptotically equivalent to $\mathbb{E}[X_n]$ *whp*. ■

Although these inequalities are powerful, we need a way to easily calculate the variance σ^2 of a particular variable. Frequently, we will deal with X being variance of the size of a random set. In such cases, it is natural to define $X = X_1 + X_2 + ...X_n$, where $X_i = 1$ if element i is included in our set and zero otherwise. Generally, it is much easier to work with the covariances of these indicator variables than to directly compute the variance of X. The following lemmas provide tools that we may apply to transform $\mathrm{Var}[X]$ into covariances of X_i and X_j.

Lemma 6.8. *For random variables $X_1, X_2, ..$ and $Y_1, Y_2, ..$*

$$Cov[\sum_i X_i, \sum_j Y_j] = \sum_{i,j} Cov[X_i, Y_j]$$

Proof. By definition,

$$\mathrm{Cov}[\sum_i X_i, \sum_j Y_j] = \mathbb{E}[\sum_i X_i \sum_j Y_j] - \mathbb{E}[\sum_i X_i]\mathbb{E}[\sum_j Y_j]$$

Expanding terms and using linearity of expectation,

$$\mathrm{Cov}\big[\sum_i X_i, \sum_j Y_j\big] = \mathbb{E}[\sum_{i,j} X_i Y_j] - (\sum_i \mathbb{E}[X_i])(\sum_j \mathbb{E}[Y_j]) = \mathbb{E}[\sum_{i,j} X_i Y_j] - \sum_{i,j} \mathbb{E}[X_i]\mathbb{E}[Y_j]$$

Taking out the summation from inside $\mathbb{E}[\sum_{i,j} X_i Y_j]$ using linearity of expectation,

$$\mathrm{Cov}\big[\sum_i X_i, \sum_j Y_j\big] = \sum_{i,j} \mathbb{E}[X_i Y_j] - \sum_{i,j} \mathbb{E}[X_i]\mathbb{E}[Y_j]$$

But the RHS is just

$$\mathrm{Cov}\big[\sum_i X_i, \sum_j Y_j\big] = \sum_{i,j} \mathrm{Cov}[X_i, Y_j]$$

so we are done. ■

Lemma 6.9. *For a random variable X,*

$$Cov[X, X] = Var[X]$$

Proof. by definition,

$$\mathrm{Cov}[X, X] = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

But this is just $\mathrm{Var}[X]$ by definition. ■

Together, these lemmas prove useful in playing with sums of co-variants. We will implement them in section 7. In particular, if we have some variable $X = X_1 + X_2 ... X_n$ where X_i is an indicator variable, then applying Lemma 6.9, $\mathrm{Var}[X] = \mathrm{Cov}[X, X]$. But we may write X as the sum of all X_i, so this becomes $\mathrm{Cov}[\sum_i X_i, \sum_j X_j]$, from where we may simplify using Lemma 6.8.

7. Applications of the Second Moment

The previous section built up numerous tools in the Second Moment Method that we can leverage. We dedicate this section to exploring a critical result obtained by applying the method.

Definition 7.1. Let $G(n, p)$ be a random graph on n vertices where every edge is included with probability p independently. We say that a particular expression x is a threshold for some event k if:

(1) when $p \gg x$, $\mathbb{P}[k] = 1 - o(1)$
(2) when $p \ll x$, $\mathbb{P}[k] = o(1)$

That is, if p is asymptotically greater than x, then k occurs *whp*, and if p is asymptotically less than x, then k occurs with probability $o(1)$.

Importantly, note that the leading coefficient of x does not matter; any particular p satisfying $p \gg x$ also satisfies $p \gg Cx$ for constant $C \geq 0$ and the same follows for $p \ll x$.

Examining the Second Moment is particularly useful in identifying this threshold x. Notably, Corollary 6.7 is provides an easy way to identify the probability p such an event E occurs, since by setting $X_n = 1$ if E occurs and 0 otherwise, if we show $\mathbb{E}[X] > 0$ and $\mathrm{Var}[X_n] = o(\mathbb{E}[X_n]^2)$, then it follows that X_n is rarely zero and E occurs *whp*. In the following Lemmas, we build a general technique to identify the threshold of the event that a particular subgraph H appears in the random graph $G(n, p)$ with n vertices and each edge included with probability p.

The bulk of technique does not rely on the probabilistic method; instead, we use probability theory to find a simpler closed form and then conclude probabilistically.

Definition 7.2. Two events (A_i, A_j) are pairwise dependent if

$$\mathbb{P}(A_i \cap A_j) \neq \mathbb{P}(A_i) \times \mathbb{P}(A_j)$$

Conversely, they are independent if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \times \mathbb{P}(A_j)$$

Intuitively, two variables are pairwise dependent if a particular outcome in event A_i makes certain outcomes in A_j more probable and vice versa. We will label two events A_i, A_j as pairwise dependent by

$$A_i \not\perp\!\!\!\perp A_j$$

For the sake of this example, we will not consider $A_i \not\perp\!\!\!\perp A_i$; that is, if you see

$$a \not\perp\!\!\!\perp b$$

this necessarily implies $a \neq b$.

Covariance is not always trivial to evaluate. Thus, when evaluating complex sums, we wish to replace covariance with probability, which is frequently easier to determine. We may do so by applying the following result:

Lemma 7.3. *Let $A_1, A_2 ... A_n$ be events. Denote $X_1, X_2, ... X_n$ as the indicator variables such that $X_i = 1$ when A_i occurs and 0 otherwise. Then,*

$$Cov[X_i, X_j] \leq \mathbb{P}(A_i)\mathbb{P}(A_j|A_i)$$

Proof. By definition,

$$\text{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$$

Since indicator variables are nonnegative, $\mathbb{E}[X_i]\mathbb{E}[X_j] \geq 0$, thus

$$\text{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \leq \mathbb{E}[X_i X_j]$$

Observe that $X_i X_j = 1$ iff A_i and A_j occur, and zero otherwise. This implies $\mathbb{E}[X_i X_j] = \mathbb{P}[A_i A_j]$. Moreover we have that

$$\mathbb{P}[A_i]\mathbb{P}[A_j|A_i] = \mathbb{P}[A_i A_j]$$

since $\mathbb{P}[A_i|A_j] = \mathbb{P}[A_i A_j]/\mathbb{P}[A_i]$ by definition. Hence,

$$\text{Cov}[X_i, X_j] \leq \mathbb{E}[X_i X_j] = \mathbb{P}[A_i]\mathbb{P}[A_j|A_i]$$

as desired. ■

This inequality has the added benefit of letting us easily bound all covariances of X_i. In particular, we may apply the inequality on the sum across all X_j to determine

$$\sum_{j \neq i} \text{Cov}[X_i, X_j] \leq \mathbb{P}[A_i] \sum_{j \neq i} \mathbb{P}[A_j|A_i]$$

As we will see later, this inequality turns out to be useful in simplifying large groups of covariances. We continue with another Lemma which allows us to convert variances into probabilities:

Lemma 7.4. *Let X_i be an indicator variable that takes 1 if event A_i occurs and zero otherwise. Then, Var$[X_i] \leq \mathbb{P}[A_i]$*

Proof. By definition,

$$\text{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$$

Suppose that A_i happens with probability p. Then $\mathbb{E}[X_i] = p$ by the properties of indicator variables. Moreover, $\mathbb{E}[X_i^2] = \mathbb{E}[X_i]$, since $0^2 = 0$ and $1^2 = 1$ and so squaring X_i never changes its value. So

$$\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = p - p^2$$

But $p^2 > 0$, so $p - p^2 \leq p$ Hence,

$$\text{Var}[X] = p - p^2 \leq p = \mathbb{P}[A_i]$$

■

With these inequalities in mind, we begin the setup for our Theorem. It's important to note that the Second Moment Method weakens on the sum of indicator variables $X = X_1 + ... + X_n$ when indicator variables exhibit significant positive dependence. If the occurrence of a particular event A_i makes other events substantially more likely, then the variance of $X = \sum_i X_i$ may become large. Consequentially, we seek to define a quantity that indicates the amount of influence a particular variable may exert on the events dependent on it. For events A_i and A_j, the conditional probability $\mathbb{P}(A_j|A_i)$ quantifies this influence. In particular, summing across all A_j that are dependent on A_i produces a sort of "dependency mass" of A_i that reflects its impact on all other variables. For these reasons, we are motivated to define the quantity Δ as we do below:

Lemma 7.5. *Let $X = X_1 + X_2.... + X_n$ be the sum of the indicator variables for events $A_1, A_2, ...A_n$. Let Δ be defined as follows:*

$$\Delta = \max_i \sum_{j : j \not\perp i} \mathbb{P}(A_j|A_i)$$

That is, Δ is the maximum possible value across all events A_i of the sum of the expression $\mathbb{P}(A_j|A_i)$ for all events A_j dependent on A_i. Then

$$\textit{Var}[X] \leq \mathbb{E}[X](1 + \Delta)$$

Proof. By Lemma 6.9 and Lemma 6.8,

$$\text{Var}[X] = \text{Cov}[X, X] = \text{Cov}[\sum_i X_i, \sum_j X_j] = \sum_{i,j} \text{Cov}[X_i, X_j]$$

We simplify by splitting the summation across all i and j into two cases; one where $i = j$ and one where $i \neq j$. Doing this and applying Lemma 6.9 when $i = j$, we obtain

$$\text{Var}[X] = \sum_{i,j} \text{Cov}[X_i, X_j] = \sum_{i=1}^{n} \text{Var}[X_i] + \sum_{i \neq j} \text{Cov}[X_i, X_j]$$

Using Lemma 7.4 on the terms $\text{Var}[X_i]$, it follows that

$$\text{Var}[X] \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i \neq j} \text{Cov}[X_i, X_j]$$

Note that $\text{Cov}[X_i, X_j]$=0 for X_i, X_j independent, so we disregard such pairs X_i, X_j. Applying Lemma 7.3 on the remaining terms $\text{Cov}[X_i, X_j]$,

$$\text{Var}[X] \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i=1}^{n} \mathbb{P}[A_i] \sum_{j:j \not\perp i} \mathbb{P}[A_j|A_i]$$

Since we defined delta as the maximum possible value of $\sum_{j:j \not\perp i} \mathbb{P}(A_j|A_i)$, the expression $\sum_{j:j \not\perp i} \mathbb{P}[A_j|A_i]$ takes a maximum of Δ. Thus

$$\text{Var}[X] \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i=1}^{n} \mathbb{P}[A_i]\Delta = (1+\Delta) \sum_{i=1}^{n} \mathbb{P}[A_i]$$

But recall that for indicator variables, $\mathbb{E}[X_i] = \mathbb{P}[A_i]$, so we simplify

$$\text{Var}[X] \leq (1+\Delta) \sum_{i=1}^{n} \mathbb{E}[X_i] = (1+\Delta)\mathbb{E}[X]$$

by linearity of expectation as desired. ■

Theorem 7.6. *Using the above setup for Lemma 7.4, if* $\mathbb{E}[X] \to \infty$ *and* $\Delta = o(\mathbb{E}[X])$*, then* $X > 0$ *and* X *is asymptotically equivalent to* $\mathbb{E}[X]$ *whp.*

Proof. By Lemma 7.5, then

$$\text{Var}[X] \leq \mathbb{E}[X](1 + o(\mathbb{E}[X]) \leq o(\mathbb{E}[X]^2)$$

By Corollary 6.7, setting $X = X_n$, the above is sufficient to conclude that $X_n > 0$ and X_n grows asymptotically with $\mathbb{E}[X_n]$ whp. ■

Applying the above theorem "trivializes" finding the threshold for the existence a particular subgraph H. By letting i range across all subsets of vertices with size $|H|$ and setting A_i as the event that subset i induces the subgraph H, then by using the setup described in Theorem 7.6 and Lemma 7.5 it suffices to find the particular probability p of $G(n,p)$ such that $\Delta = o(\mathbb{E}[X])$.

Finding this probability is often not too challenging. Finally, here is where we employ probabilistic method arguments. Take the example where H is a complete graph with 4 vertices; i. e, the clique K_4.

Theorem 7.7. *For* $H = K_4$*, the threshold* x *of* $G(n,p)$ *is* $x = n^{-2/3}$

Proof. First, we show that if $p \ll n^{-2/3}$, then H occurs with probability $o(1)$.

Define X as the number subgraphs $H = K_4$ in $G(n,p)$. Let i range across all subsets of vertices in $G(n,p)$ of size 4, and set A_i to the event that subset i is connected; i.e, has edges equal to H. Letting the indicator variable $X_i = 1$ if A_i occurs and 0 otherwise, by linearity of expectation:

$$\mathbb{E}[X] = \sum_i X_i = \sum_i \mathbb{P}[A_i]$$

Note that every edge in a particular subgraph i is included within $G(n,p)$ with probability p. In total, there are $\binom{4}{2} = 6$ such edges for each subgraph; hence $\mathbb{P}[A_i] = p^6$. Moreover, there are a total of $\binom{n}{4}$ such i. Thus, summing across all i, we get

$$\mathbb{E}[X] = \binom{n}{4} p^6$$

$\binom{n}{4}$ is $O(n^4)$, so $p \ll n^{-2/3}$ gives $\mathbb{E}[X] = o(1)$. By Markov's inequality, any $\varepsilon > 0$ would have probability $\mathbb{P}(X > \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon} = o(1)$, so $X = 0$ whp.

It remains to show that if $p \gg n^{-2/3}$, then $X > 0$ whp. Note that from the above, $\mathbb{E}[X] \to \infty$ when $p \gg n^{-2/3}$. We now compute Δ.

Observe that $A_i \not\perp\!\!\!\perp A_j$ if an edge is shared between vertex sets i and j. This requires $|i \cap j| \geq 2$, since an edge is then shared between i and j. The number of j that share precisely two vertices with arbitrary i is $6\binom{n-4}{2} = O(n^2)$. There are 5 edges in j not shared with i, so $\mathbb{P}(A_j|A_i) = p^5$.

The number of j that share precisely 3 vertices with arbitrary i is $4\binom{n-4}{1} = O(n)$. In this case, there are 3 edges not shared with i in j, so $\mathbb{P}(A_j|A_i) = p^3$. Summing over all possible j,

$$\Delta = \sum_{j:|j\cap i|\in 2,3} \mathbb{P}(A_j|A_i) = O(n^2p^5 + np^3) \ll \mathbb{E}[X] = O(n^4p^6)$$

Since $\Delta \ll \mathbb{E}[X]$ and $\mathbb{E}[X] \to \infty$, by Theorem 7.6 we are done. ■

These results may seem relatively detached from the probabilistic method. In effect, the Second Moment Method leans more strongly into probability theory. Frequently, the probabilistic method borrows heavily from these topics in probability theory or, at certain points, analysis in order to produce stronger results. The probabilistic method is, after all, a method involving probability theory, and more complex results naturally hinge on more sophisticated probabilistic machinery. However, ultimately the conclusions are applied to a random graph $G(n, p)$, and the behavior of this graph is studied; the study of this graph involves the probabilistic method.

Theorem 7.6 is very broad and not unique to thresholds of random graphs. It is a sufficiently strong result to be widely applicable, and the setup is standard for any construction in which there exist few dependencies.

8. The Lovasz Local Lemma

Theorem 7.6 is a very strong result when we wish to prove that a particular event happens *whp* or with probability $o(1)$. However, it is not applicable when we desire to show that an event happens with non-zero probability (not necessarily *whp*) as we have in previous sections. Naturally, we seek an alternative to Theorem 7.6 which is applicable in these cases–hence, we arrive upon the **Lovasz Local Lemma**.

Although Theorem 7.6 is a strong when considering scenarios with few dependencies, it weakens when variables are weakly dependent. To understand this, recall Lemma 7.3; in the process of bounding $\mathrm{Cov}[X_i, X_j]$ we crudely eliminate the term $\mathbb{E}[X_i]\mathbb{E}[X_j]$. When variables have dependencies that are not strong, this error term approaches $\mathbb{E}[X_iX_j] = \mathbb{P}[A_i]\mathbb{P}[A_j|A_i]$. As a consequence, the inequality we obtain in Lemma 7.3 becomes very loose, causing our eventual result in Theorem 7.6 to weaken. For similar reasons, the Lovasz Local Lemma is best when there are few, strong dependencies between variables.

To proceed with the Lemma, we must introduce a stronger notion of independency.

Definition 8.1. An event A_0 is mutually independent from a set of events $\{A_1, A_2, ...A_m\}$ if for all sets $\{B_1, B_2...B_m\}$ where $B_i = A_i$ or B_i is the event that A_i doesn't occur,

$$\mathbb{P}[A_0|B_1 \wedge B_2 \wedge .... \wedge B_m] = \mathbb{P}[A_0]$$

That is, the probability of A_0 is unchanged by any outcome of the events $A_1, ..., A_m$.

The following identity in probability theory is necessary in the Lemma.

Lemma 8.2. *For events X, Y, Z,*

$$\mathbb{P}(X|Y \wedge Z) = \frac{\mathbb{P}(X \wedge Y|Z)}{\mathbb{P}(Y \wedge Z)}$$

Proof. The definition of conditional probability gives

$$\mathbb{P}(X|Y \wedge Z) = \frac{\mathbb{P}(X \wedge Y \wedge Z)}{\mathbb{P}(Y \wedge Z)}$$

Simplifying the RHS by noting $P(X, Y|Z) = P(X \wedge Y \wedge Z)/P(Z)$ by the definition of probability and substituting in gives

$$P(X|Y, Z) = \frac{P(X, Y|Z)}{\frac{P(Y \wedge Z)}{Z}} = \frac{P(X, Y|Z)}{P(Y|Z)}$$

as desired. ■

We now establish the Lemma:

Lemma 8.3. ***Lovasz Local Lemma:*** *Let $A_1, A_2, ...A_n$ be events. For a particular event i, let $N(i)$ denote the set of events such that A_i is mutually independent from*

$$(A_j : j \notin \{i\} \cup N(i))$$

That is, $N(i)$ is the set of events that A_i is not mutually dependent from. If we can select a particular set $\{x_1, x_2, ...x_n\} \in [0, 1)$ such that the following is satisfied for all $i \in [n]$

$$\mathbb{P}[A_i] \leq x_i \prod_{j \in N(i)} (1 - x_j)$$

Then

$$\mathbb{P}[\textit{no event } A_i \textit{ occurs}] \geq \prod_{i \in [n]} (1 - x_i)$$

In particular, $\mathbb{P}[\textit{no event } A_i \textit{ occurs}] > 0$.

Proof. We begin by showing

$$\mathbb{P}(A_i| \bigwedge_{j \in S} \overline{A_j}) \leq x_i \text{ whenever } i \notin S \subseteq [n]$$

We do so by induction on $|S|$. Take base case $|S| = 0$. Here, we seek to show

$$\mathbb{P}[A_i] \leq x_i$$

But we know that $\mathbb{P}[A_i] \leq x_i \prod_{j \in N(i)}(1 - x_j)$; observe that $1 \geq \prod_{j \in N(i)}(1 - x_j)$, thus

$$\mathbb{P}[A_i] \leq x_i \prod_{j \in N(i)} (1 - x_j) \leq x_i$$

as desired. Hence, the hypothesis holds true for base case $|S| = 0$. We proceed with the inductive step. Consider some arbitrary $|S|$; suppose that the hypothesis holds true for all $|S'| < |S|$. We wish to show that the hypothesis holds for $|S|$.

Denote $S_1 = S \cap N(i)$ and $S_2 = S/N(i)$. Evidently, S_1 and S_2 are disjoint subsets with $S_1 \cup S_2 = S$.

Importantly, note that we may decompose

$$\mathbb{P}(A_i|\bigwedge_{j\in S}\overline{A_j}) = \frac{\mathbb{P}(A_i\bigwedge_{j\in S_1}\overline{A_j}|\bigwedge_{j\in S_2}\overline{A_j})}{\mathbb{P}(\bigwedge_{j\in S_1}\overline{A_j}|\bigwedge_{j\in S_2}\overline{A_j})}$$

since applying Lemma 8.2 with $A_i = X$, $\bigwedge_{j\in S_1}\overline{A_j} = Y$, and $\bigwedge_{j\in S_2}\overline{A_j} = Z$ gives the desired decomposition. Focusing on the numerator of our RHS, note that the probability that A_i but not any A_j with $j \in S_1$ occur is at best $\mathbb{P}[A_i]$. Hence, we bound

$$\text{numerator} = \mathbb{P}(A_i\bigwedge_{j\in S_1}\overline{A_j}|\bigwedge_{j\in S_2}\overline{A_j}) \leq \mathbb{P}(A_i|\bigwedge_{j\in S_2}\overline{A_j})$$

Recall that A_i is independent of the set $\{A_j : j \in S_2\}$, thus by definition $\mathbb{P}(A_i|\bigwedge_{j\in S_2}\overline{A_j}) = \mathbb{P}(A_i)$. Using $\mathbb{P}[A_i] \leq x_i\prod_{j\in N(i)}(1-x_j)$, it follows that

$$\text{numerator} \leq \mathbb{P}(A_i|\bigwedge_{j\in S_2}\overline{A_j}) = \mathbb{P}[A_i] \leq x_i\prod_{j\in N(i)}(1-x_j)$$

Plugging in our bound for the numerator,

$$\mathbb{P}(A_i|\bigwedge_{j\in S}\overline{A_j}) \leq \frac{x_i\prod_{j\in N(i)}(1-x_j)}{\mathbb{P}(\bigwedge_{j\in S_1}\overline{A_j}|\bigwedge_{j\in S_2}\overline{A_j})}$$

We now focus on the denominator. Denoting the elements of S_1 as $\{j_1, j_2, ...j_r\}$, since $\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$, we may iteratively use this property to construct the expansion

$$\text{denominator} = \mathbb{P}(\bigwedge_{j\in S_1}\overline{A_j}|\bigwedge_{j\in S_2}\overline{A_j}) = \mathbb{P}(\overline{j_1}|\bigwedge_{j\in S_2}\overline{A_j})\mathbb{P}(\overline{j_2}|\overline{j_1}\bigwedge_{j\in S_2}\overline{A_j}).....\mathbb{P}(\overline{j_r}|\overline{j_1}\wedge\overline{j_2}\wedge...\overline{j_{r-1}}\bigwedge_{j\in S_2}\overline{A_j})$$

But $|S_2| + |S_1| - 1 < |S|$, so for every term j_i, the inductive hypothesis guarantees

$$\mathbb{P}[j_i|\overline{j_1}\wedge\overline{j_2}..\wedge\overline{j_{i-1}}\bigwedge_{j\in S_2}\overline{A_j}] \leq x_{j_i}$$

since we supposed that it holds for every $|S'| < |S|$. Recall that for any event k, $\mathbb{P}(\overline{j_i}|k)$ is the same as $1 - \mathbb{P}(j_i|k)$, so we can bound the denominator with

$$\text{denominator} \geq \prod_{j_i\in[S_2]=N(i)}(1-x_{j_i})$$

Recombining this with our numerator, we achieve the bound

$$\mathbb{P}(A_i|\bigwedge_{j\in S}\overline{A_j}) \leq \frac{x_i\prod_{j\in N(i)}(1-x_j)}{\prod_{j_i\in N(i)}(1-x_{j_i})} = x_i$$

as desired. We have proved the base case and the inductive step, so by the principle of mathematical induction,

$$\mathbb{P}(A_i|\bigwedge_{j\in S}\overline{A_j}) \leq x_i \text{ whenever } i \notin S \subseteq [n]$$

With this in mind, again using $\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$ iteratively, we may write

$$\mathbb{P}(\bigwedge_{i=1}^{n} \overline{A_i}) = \mathbb{P}(\overline{A_1})\mathbb{P}(\overline{A_2}|\overline{A_1}).....\mathbb{P}(\overline{A_n}|\overline{A_1} \wedge \overline{A_2} \wedge \overline{A_3}... \wedge \overline{A_{n-1}})$$

For each factor with leading term A_i, note that the factor is congruent to $1 - \mathbb{P}(A_i|\overline{A_1} \wedge \overline{A_2} \wedge .... \wedge \overline{A_{i-1}})$. Thus, applying the inequality we just proved to each factor in this expression, we conclude that

$$\mathbb{P}(\bigwedge_{i=1}^{n} \overline{A_i}) \geq (1-x_1)(1-x_2)....(1-x_n)$$

as desired. ■

This is the general local Lemma. As a consequence of this result, we obtain the following, which is weaker but typically simpler to apply:

Corollary 8.4. *Let $A_1, ...A_n$ be events such that no event has probability larger than p of occuring. Suppose that A_i is mutually independent from a set of size at least $(n-d) < n$ events for all i. Then*

$$ep(d) \leq 1$$

implies that no events A_i occur with positive probability.

Proof. Set $x_i = 1/d$ for all i, and let $N(i)$ be defined as in Lemma 8.2. By definition, for all i, A_i has an independent set of size at least $n-d$. and $N(i)$ is effectively the set of $n-1$ A_j for $j \neq i$ minus the elements of an independent set, $|N(i)| \geq d-1$. Hence,

$$x_i \prod_{j \in N(i)} (1-x_j) = \frac{1}{d} \prod_{j \in N(i)} (1 - \frac{1}{d})$$

is bounded by

$$\frac{1}{d} \prod_{j \in N(i)} (1 - \frac{1}{d}) \geq \frac{1}{d} \prod_{j=1}^{d} (1 - \frac{1}{d}) = \frac{1}{d}(1 - \frac{1}{d})^d$$

Observe that $(1 - \frac{1}{d})^d$ is constantly decreasing as d tends towards infinity. Combining this with the known limit

$$\lim_{d \to \infty} (1 - \frac{1}{d})^d = \frac{1}{e}$$

gives that $(1 - \frac{1}{d})^d > \frac{1}{e}$ for all finite $d > 0$. Thus,

$$x_i \prod_{j \in N(i)} (1-x_j) \geq \frac{1}{d}(1 - \frac{1}{d})^d > \frac{1}{ed}$$

But we defined $epd \leq 1$, so $\frac{1}{ed} \geq p$. Hence

$$x_i \prod_{j \in N(i)} (1-x_j) > \frac{1}{ed} \geq p$$

so the hypothesis of the Lovasz Local Lemma is satisfied. By Lemma 8.2, no events A_i occur with positive probability, so we are done.

■

9. Applying the Lovasz Local Lemma

The Lovasz Local Lemma is among the more powerful techniques in the probabilistic toolkit. We present a few results which are produced by the LLL in order to illustrate how one might typically apply it. Several more interesting and more sophisticated examples are also presented in [Sas26].

Theorem 9.1. *Let $G = (V, E)$ be a graph with maximum degree Δ and let $V = V_1 \cup V_2 ... \cup V_r$ be a partition of the vertices, where each $|V_i| \geq 2e\Delta$. Then there exists an independent set in G containing one vertex in each V_i.*

Proof. We may assume that each V_i has magnitude $|V_i| = \lceil 2e\Delta \rceil$, or else we are able to remove vertices from V_i until this holds. This process of pruning the sets turns out to be necessary, since uneven sets make the Symmetric LLL (which we will apply) substantially weaker.

From each V_i, we select a vertex $v_i \in V_i$ uniformly at random. If we can show that $\{v_1, v_2 .... v_r\}$ is an independent set with positive probability, then we are done.

For an edge $e \in E$, let A_e denote the probability that both endpoints in e are in $\{v_1, v_2 .... v_r\}$. There are a total of r choices v_i from a set of $r \times 2e\Delta$ vertices. Hence,

$$\mathbb{P}(A_e) \leq (\frac{r}{r \times 2e\Delta})^2 = \frac{1}{(2e\Delta)^2}$$

Each A_e is dependent on some other A_f if there exists a V_i such that e and f are incident to vertices in V_i. For a particular e, there are at most $2(2e\Delta)$ ways to choose the vertex of f that is incident to V_i. There are at most Δ ways to choose the second vertex of f, since the maximum degree of a vertex is Δ. Thus A_e is dependent on at most $2(2e\Delta)(\Delta)$ other events. Applying the symmetric LLL on the events A_e,

$$epd \leq e\frac{1}{(2e\Delta)^2} 2(2e\Delta)(\Delta) = 1$$

So by the Symmetric LLL, with positive probability there does not exist an edge e such that both endpoints of e are selected as a v_i. Thus $\{v_1, v_2 ... v_r\}$ is independent with positive probability, and so there must exist a particular choice $\{v_1, v_2 ... v_r\}$ that is independent. ■

10. Martingale Inequalities and Bounded Differences

In Section 5 we introduced the Chebyshev Inequality, which provided a bound for deviations of size λ from the expectation that was inversely quadratic in λ. Although this result is useful, it is often not enough to show that a variable strays from its mean with inversely polynomial probability. In particular, we frequently require for the probability of a deviation of size λ to be inversely exponential. For instance, when we are bounding several independent random variables individually, if the number of variables becomes sufficiently large, then Chebyshev gives an exponentially small probability of every variable being concentrated. If we could instead prove a sufficiently powerful inversely exponential tail bound, we would be able to guarantee little deviation with high probability.

This is substantial improvement over the Chebyshev Inequality, but consequentially, results proving inversely exponential tail bounds tend to be less general. One of the strongest exponential tail bounds is known as the **Bounded Differences Inequality**, which leverages the fact that if perturbations in a single choice minimally affects the outcome, then the entire variable must be concentrated. We roughly follow the argument of [Car19]

To prove the Bounded Differences Inequality, we use **Martingales**.

Definition 10.1. A sequence of random variables $Z_0, Z_1 ... Z_n$ is a *Martingale* if for all i,

$$\mathbb{E}[Z_i | Z_0, Z_1 ... Z_{i-1}] = Z_{i-1}$$

More formally, a sequence $Z_0, Z_1 ... Z_n$ of random variables on the probability space $(\Omega, \mathcal{F}, P)$, along with a filtration $\mathcal{F}_i$ that represents the history of the process before Z_i, which satisfies

- for all i, $\mathbb{E}[Z_i] < \infty$
- For all i, the conditional value expectation of the next value given the history up to time i, is equal to the current value. That is, $\mathbb{E}[Z_{i+1} | \mathcal{F}_i] = Z_i$.

We begin with a Lemma, which we will later apply to the study of martingales in order to produce exponential bounds on martingale processes, from which we will deduce the bounded differences inequality.

Lemma 10.2. ***Hoeffding's Lemma:*** *Let X be a random variable which takes on maximum b and minimum a such that $b - a = \ell$. Then supposing $\mathbb{E}[X] = 0$, it must be true that*

$$\mathbb{E}[e^X] \leq e^{\ell^2/8}$$

Proof. Since e^x is convex, on the interval $[a, b]$ it is always smaller than the line passing through the points (a, e^a) and (b, e^b). Then by writing the equation of this line,

$$e^x \leq e^a + \frac{e^b - e^a}{b - a}(x - a) = \frac{b - x}{\ell} e^a + \frac{x - a}{\ell} e^b, \text{ for all } x \in [a, b]$$

Taking the expectation of both sides and recalling $\mathbb{E}X = 0$,

$$\mathbb{E}[e^x] \leq \frac{b}{\ell} e^a + \frac{-a}{\ell} e^b$$

Let $p = -a/\ell$, so that $a = -p\ell$ and $b = (1 - p)\ell$. Then substituting in p and taking the logarithm of both sides, we obtain

$$\log \mathbb{E}[e^x] \leq \log\Big((1 - p)e^{-p\ell} + pe^{(1-p)\ell}\Big) = -p\ell + \log\Big(1 - p + pe^{\ell}\Big)$$

Define the function

$$f(\ell) = -p\ell + \log\left(1 - p + pe^{\ell}\right)$$

and fix $p \in [0, 1]$. Then it remains to show that $f(\ell) \leq \ell^2/8$ for all $\ell \geq 0$, since this would imply $\log \mathbb{E}[e^x] \leq \ell^2/8$, and hence $\mathbb{E}[e^X] \leq e^{\ell^2/8}$. Observe $f(0) = 0$, and computation gives

$$f'(\ell) = -p + \frac{pe^{\ell}}{1 - p + pe^{\ell}}$$

and hence $f'(0) = 0$. Taking the second derivative, we obtain

$$f''(\ell) = \frac{pe^{\ell}(1 - p)}{(1 - p + pe^{\ell})^2} = \Big(\frac{pe^{\ell}}{1 - p + pe^{\ell}}\Big)\Big(1 - \frac{pe^{\ell}}{1 - p + pe^{\ell}}\Big)$$

Since $t(1 - t) \leq 1/4$ for $t \in [0, 1]$, this means $f''(\ell) \leq 1/4$. Since $f(0)$ and $f'(0)$ are the same as the corresponding values of $\ell^2/8$, but $f''(\ell)$ is always smaller than the second derivative of $\ell^2/8$, it follows that $f(\ell) \leq \ell^2/8$ for all $\ell \geq 0$. This completes the proof. ■

We now proceed with an inequality that bounds Martingale differences.

Theorem 10.3. ***Azuma's Inequality:*** *Let $Z_0, Z_1, ...Z_n$ be a martingale satisfying*

$$|Z_i - Z_{i-1}| \leq c_i \text{ for each } i \in [n]$$

Then for any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + c_2^2... + c_n^2)}\right).$$

Proof. Conditioned on the values of $Z_0, ...Z_{i-1}$, Z_i must lie in the interval $\{Z_{i-1} - c, Z_{i-1} + c\}$ by the hypothesis. Then since $\mathbb{E}[Z_i | Z_{i-1}...Z_0]] = Z_{i-1}$, conditioning on $Z_0, ...Z_{i-1}$, we have $\mathbb{E}[Z_i - Z_{i-1}] = 0$. Let t be decided later. We may apply Hoeffding's Lemma on $t(Z_i - Z_{i-1})$ to obtain

$$\mathbb{E}[e^{t(Z_i - Z_{i-1})} | Z_0, ...Z_{i-1}] \leq e^{t^2 c_i^2/2}$$

With this in mind, we consider the expression $\mathbb{E}[e^{t(Z_i - Z_0}]$. By separating the exponent, we find

$$\mathbb{E}[e^{t(Z_i - Z_0}] = \mathbb{E}[e^{t(Z_i - Z_{i-1})} e^{t(Z_{i-1} - Z_0)}].$$

By the tower property of conditional expectation, and noting that $e^{t(Z_{i-1} - Z_0)}$ is constant given the outcomes of Z_{i-1} and Z_0,

$$\mathbb{E}[e^{t(Z_i - Z_{i-1})} e^{t(Z_{i-1} - Z_0}] = \mathbb{E}[\mathbb{E}[e^{t(Z_i - Z_{i-1})} | Z_0, ...Z_{i-1}] e^{t(Z_{i-1} - Z_0}] \leq e^{t^2 c_i^2/2} \mathbb{E}[e^{t(Z_{i-1} - Z_0)}]$$

Using the bound $\mathbb{E}[e^{t(Z_i - Z_0}] \leq e^{t^2 c_i^2/2} \mathbb{E}[e^{t(Z_{i-1} - Z_0)}]$, we may descend down from $i = n$ to $i = 0$ to obtain

$$\mathbb{E}[e^{t(Z_n - Z_0)}] \leq \prod_{i=1}^{n} e^{t^2 c_i^2/2} = e^{t^2(c_1^2 + c_2^2... + c_n^2)/2}$$

Since $\mathbb{P}(Z_n - Z_0 \geq \lambda) = \mathbb{P}(e^{t(Z_n - Z_0)} \geq e^{t\lambda})$, Markov's Inequality implies

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) = \mathbb{P}(e^{t(Z_n - Z_0)} \geq e^{t\lambda}) \leq e^{-t\lambda} \mathbb{E}[e^{t(Z_i - Z_0}] \leq e^{-t\lambda + \frac{t^2}{2}(c_1^2 + c_2^2... c_n^2)}$$

Substituting $t = \lambda / c_1^2 + ... + c_n^2$ gives the desired inequality. ■

With this in mind, we can finally proceed with the Bounded Differences Inequality. Intuitively, the inequality follows from the fact we can construct a martingale using the conditions of the hypothesis, from which Azuma yields the inequality.

Theorem 10.4. ***Bounded Differences Inequality:*** *Let $X_1 \in \Omega_1 .... X_n \in \Omega_n$ be Independent random variables. Suppose the function $f : \Omega_1 \times .... \times \Omega_n \to \mathbb{R}$ satisfies*

$$|f(x_1 .....x_n) - f(x_1' ....x_n')| \leq c_i$$

Whenever $\{x_1 ...x_n\}$ and $\{x_1' .....x_n'\}$ differ only on the ith index for constants c_i. Then the random variable $Z = f(X_1 ...X_n)$ satisfies, for all $\lambda \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + c_2^2... + c_n^2)}\right)$$

and likewise

$$\mathbb{P}(Z - \mathbb{E}[Z] \leq -\lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + c_2^2... + c_n^2)}\right)$$

Proof. Consider the sequence of random variables $Z_0, ...Z_n$ such that $Z_i = \mathbb{E}[f(X_1....X_n)|X_1...X_i]$. Then $Z_0 = \mathbb{E}[f(X_1....X_n)]$ and $Z_n = f(X_1....X_n)$. We claim that this sequence is a martingale. Indeed, $\mathbb{E}[Z_i|Z_,.....Z_{i-1}] = Z_{i-1}$, since the expected value of f after fixing the value X_i, taken across all outcomes X_i, is just the same as Z_{i-1}.

Now, we show that $|Z_i - Z_{i-1}| \leq c_i$. Fix the values of the random variables $X_0 = x_0$, $....X_i = x_i$. Then $Z_i = \mathbb{E}[f(x_0..x_i, X_{i+1}.., X_n)]$. Introduce the function

$$g_i(a) = \mathbb{E}[f(x_1, ..., x_{i-1}, a, X_{i+1}, ..., X_n]$$

Since changing only one input X_i changes f by at most c_i, it follows that for all constants a, b,

$$|g_i(a) - g_i(b)| \leq c_i$$

Indeed, this follows from

$$\begin{aligned} |g_i(a) - g_i(b)| &= |\mathbb{E}[f(....a....) - f(.....b....)]| \\ &= \mathbb{E}[|f(....a....) - f(.....b....)|] \\ &\leq \mathbb{E}[c_i] = c_i \end{aligned}$$

This implies that all possible values of $g_i(x)$ lie inside an interval of length at most c_i. Then its expectation must lie in this interval as well. Note that $Z_i = g_i(x_i)$ and by the properties of martingales, $Z_{i-1} = \mathbb{E}[Z_i] = \mathbb{E}[g_i(x_i)]$. Hence, it follows that

$$|g_i(x_i) - \mathbb{E}[g_i(x_i)]| \leq c_i$$

and thus

$$|Z_i - Z_{i-1}| \leq c_i$$

So the sequence $Z_0....Z_n$ satisfies the hypothesis of Azuma's Inequality, and so

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + c_2^2... + c_n^2)}\right)$$

We obtain the Bounded Differences inequality by recalling $Z_n = Z$ and $Z_0 = \mathbb{E}[Z]$. ■

11. Applying Bounded Differences

We present a brief example of how one may leverage Bounded Differences to produce concentration results, and consequentially bound a particular quantity. Recall that $G(n, p)$ is a graph where each edge on n vertices is included with probability p, and that $\chi(G)$ is the chromatic number of G; that is, the minimum number of colors needed such that there exists a valid coloring of G that avoids monochromatic edges.

Theorem 11.1. *Let $\alpha > 5/6$ be fixed. Then for $p < n^{-a}$, $\chi(G(n,p))$ is concentrated on 4 values with probability $1 - o(1)$. That is, there exists a $u = u(n,p)$ such that as $n \to \infty$*

$$\mathbb{P}(u \leq \chi(G(n,p) \leq u + 3) = 1 - o(1).$$

Proof. It suffices to show that for all $\varepsilon > 0$, we may find a $u = u(n, p, \varepsilon)$ such that with $p < n^{-a}$ and n sufficiently large,

$$\mathbb{P}(u \leq \chi(G(n,p) \leq u + 3) = 1 - 3\varepsilon.$$

Set u as the smallest integer satisfying $\mathbb{P}(\chi(G(n,p) \geq u) \geq \varepsilon$. Let G be a graph produced by the random procedure $G(n,p)$, and let $Y = Y(G)$ denote the minimum possible size of a subset $S \subseteq V(G)$, such removing S from G yields a u-colorable graph.

Note that Y changes the by at most 1 if we change the edges connecting to one vertex, since we will either need to add or remove the vertex from S. Fix a particular ordering of the vertices, and let x_i denote the set of edges connected to the ith vertex in the ordering, that are not contained in any $x_j, j < i$. Note that all x_i are independent, in the sense that no two x_i, x_j contain the same edge. Then define

$$f(x_1....x_n) = Y$$

and since changing one parameter in f changes Y by at most 1, bounded differences gives

$$\mathbb{P}(Y - \mathbb{E}[Y] \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2} \text{ and } \mathbb{P}(Y - \mathbb{E}[Y] \leq -\lambda\sqrt{n}) \leq e^{-\lambda^2/2}$$

for all $\lambda > 0$. Choose $\lambda = \lambda(\varepsilon)$ such that $e^{-\lambda^2/2} = \varepsilon$. We first show that $\mathbb{E}[Y]$ must be small. We have

$$e^{-\lambda^2/2} = \varepsilon < \mathbb{P}(\chi(G) \leq u)$$

But note that $\chi(G) \leq u$ implies $Y = 0$, hence

$$e^{-\lambda^2/2} \leq \mathbb{P}(Y = 0) = \mathbb{P}(Y - \mathbb{E}[Y] \leq \mathbb{E}[Y]) = \exp(\frac{-\mathbb{E}[Y]^2}{2n})$$

And thus $-\lambda^2/2 \leq -\mathbb{E}[Y]^2/2n$, which gives $\mathbb{E}[Y] \leq \lambda\sqrt{n}$. Now, we apply the bound to show that Y is rarely large. Using the obtained bound for $\mathbb{E}[Y]$,

$$\mathbb{P}(Y \geq 2\lambda\sqrt{n}) \leq \mathbb{P}(Y - \mathbb{E}[Y] \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2} = \varepsilon$$

It follows that with probability $1-\varepsilon$, there exists a $Y \subseteq V(G)$ with $|S| \leq 2\lambda\sqrt{n}$ such that G is u-colorable when S is removed. Moreover, recall that with probability $1-\varepsilon$, $\chi(G) \geq u$ by our choice of u in the beginning. If we can 3-color S, then it follows that $\chi(G) \leq u+3$ by u-coloring $G-S$ and then using 3 new colors for S. Hence, if we prove that with probability at least $1-\varepsilon$, S is 3 colorable, it follows that with probability at most $1-3\varepsilon$, $u \leq \chi(G) \leq u+3$. We do so in the following Lemma:

Lemma 11.2. *Fix $a > 5/6$, constant C and $p = n^{-a}$. Then with probability $1 - o(1)$, every subset of at most $C\sqrt{n}$ vertices can be properly 3-colored.*

Proof. Let G be the graph produced by the random process, and assume that G is not 3-colorable, or else the lemma follows trivially. Choose the smallest possible $T \subseteq V(G)$ such that $G[T]$ is not 3-colorable. We claim that $G[T]$ has minimum degree ≥ 3. Indeed, if for some vertex x, $\deg_{G[T]}(x) < 3$, then we may remove x from T and produce a graph that is not 3-colorable, violating minimality. This follows from the fact that, if $T - x$ were 3-colorable, we could extend the coloring to x since it is bounded by ≤ 2 other vertices.

Therefore $G[T]$ must have at least $3|T|/2$ edges in order to force a minimum degree greater than 2. We now compute the probability that G has an induced subgraph on $t \leq C\sqrt{n}$ vertices and $\geq 3t/2$ edges; let the event that this occurs be A. The probability that some subgraph exists is at most the sum of the probabilities that a particular subgraph H satisfies the conditions, and thus

$$\mathbb{P}(A) \leq \sum_{t=4}^{\lfloor C\sqrt{n} \rfloor} \binom{n}{t} \binom{\binom{t}{2}}{3t/2} p^{3t/2}$$

We obtain this summation by noting that $\binom{n}{t}\binom{\binom{t}{2}}{3t/2}$ generates all possible induced subgraphs on t vertices with precisely $3t/2$ edges. Taking precisely $3t/2$ edges rather than $\geq 3t/2$ edges suffices, since for sufficiently large n and $a > 5/6$, ignoring graphs with more than $3t/2$ edges

changes the expression by a constant factor $o(1)$. Simplifying with the well-known inequality $\binom{n}{k} \leq (\frac{ne}{k})^k$ and substituting $p = n^{-\alpha}$

$$\mathbb{P}(A) \leq \sum_{t=4}^{\lfloor C\sqrt{n} \rfloor} (\frac{ne}{t})^t (\frac{te}{3})^{3t/2} n^{-\alpha 3t/2} = \sum_{t=4}^{\lfloor C\sqrt{n} \rfloor} n^{t-a3t/2} e^{5t/2} 3^{-3t/2} t^{1/2t}$$

Taking asymptotics, we see that

$$\mathbb{P}(A) \leq \sum_{t=4}^{\lfloor C\sqrt{n} \rfloor} O\big(n^{1-3\alpha/2} t^{1/2}\big)^t$$

With $t \leq C\sqrt{n}$, it follows that

$$\mathbb{P}(A) \leq \sum_{t=4}^{\lfloor C\sqrt{n} \rfloor} O\big(n^{(1-3\alpha/2+1/4)}\big)^t$$

Provided $\alpha > 5/6$, the exponent of n inside the asymptotic is ≤ 0, and with $t > 1$, this implies the summation is $o(1)$. ■

By Lemma 11.2, S is of size $2\lambda\sqrt{n} = C\sqrt{n}$ with probability $1-\varepsilon$, and hence is 3-colorable with high probability. Therefore, we are done. ■

Remark 11.3. Note that this result is significantly stronger than what we can produce with the setup of Theorem 7.6, since 7.6 only guarantees asymptotic equivalence, and doesn't bound the variable in a particular interval. In general, Martingale Inequalities and Bounded Differences produce substantially more powerful concentration results.

12. Dependent Random Choice

In Graph Theory, one of the most natural questions to ask for a particular dense graph G is whether a certain graph H exists as a subgraph of G. For instance, in Section 7, we found a threshold for the existence of the subgraph K_4 in the random graph $G(n,p)$. Questions of this nature are connected to numerous other characteristics of graphs, and hence it is highly desirable to develop methods to attack them. For instance, determining if the subgraph K_n exists in $G = (V, E)$ or its complement H is analogous to bounding the Ramsey number $R(k,k)$. Proving the existence of certain subgraphs also facilitates the decomposition of dense graphs into smaller, sparser subgraphs. When we describe the search for a subgraph K inside G, we frequently discuss it in terms of an "embedding".

Definition 12.1. An *embedding* of graph K onto G is just an injective mapping of K onto G; that is, a graph K is *embedded* in G if it is a subgraph of G

Simple probabilistic arguments tend to be very weak at proving that there exists an embedding. For particular graphs G and H, the number of possible areas where H may be embedded grows like $v(G)^{v(H)}$ for sufficiently small $v(H)$. The number of possible cases makes a naive first moment or alteration approach intractable, and for reasonably sized $v(H)$, the number of dependencies grows large enough for the Second Moment and LLL to weaken substantially. As a consequence, we seek a stronger probabilistic technique; we find this in Dependent Random Choice (DRC) which offers a powerful tool for finding Graph Embeddings.

One way to help find embeddings is to find and characterize regions which are very well-connected and have many common neighbors. If we are able to find such an area in a graph G, then we may greedily search for an embedding of H by using the common neighbors.

Therefore it is particularly useful to isolate local regions of a particular dense graph with sufficiently large amounts of common neighbors. To force the existence of these local areas, we may apply **Dependent Random Choice**. In a certain dense graph, DRC involves taking a random subset of the vertices T and then evaluating their common neighbors, $N(T)$. This process makes it disproportionately likely that this set $N(T)$ has a large number of common neighbors; this is because the probability of a certain subset showing up in $N(T)$ corresponds to the number of elements of $V(G)$ that all elements of the subset's neighbors. By pruning the resulting set $N(T)$ through alteration, we may entirely avoid the infrequent undesirable subsets with few shared neighboring vertices, and hence produce the desired vertex subset with many common neighbors. For a more rigorous treatment of Dependent Random Choice and the results it produces, see [FS10]

The following theorem uses the DRC to give a condition for finding sets with many common neighbors based on a general density statement.

Theorem 12.2. *Let a, d, m, n, r be positive integers, and let $G = (V, E)$ be a graph on n vertices with average degree d. If there exists a positive integer t such that*

$$\frac{d^t}{n^{t-1}} - \binom{n}{r}\left(\frac{m}{n}\right)^t \geq a$$

then G must contain a subset U of at least a vertices such that every set of r vertices in U has at least m common neighbors.

Proof. Choose a set T of t vertices uniformly and independently at random, with repetition. Let X denote the size of the set $N(T)$. For each $v \in V(G)$, the probability that v is contained within $N(T)$ is equivalent to the probability that every vertex of T lies in $N(v)$. Since we select every element of T independently, it follows that $\mathbb{P}(v \in N(T)) = \left(\frac{N(v)}{n}\right)^n$ Hence by linearity of expectation,

$$\mathbb{E}[X] = \sum_{v \in V(G)} \mathbb{P}(v \in N(T)) = \sum_{v \in V(G)} \left(\frac{|N(v)|}{n}\right)^t = n^{-t} \sum_{v \in V(G)} |N(v)|^t$$

Jensen's Inequality gives that $f\left(\frac{x_1+x_2..+x_k}{k}\right) \leq \frac{f(x_1)+...+f(x_k)}{k}$ for convex functions $f(x)$. Applying Jensen's inequality on the convex function $f(|N(v)|) = |N(v)|^t$ with $k = n$, it follows that

$$n^{-t} \sum_{v \in V(G)} |N(v)|^t \geq n^{1-t}\left(\frac{\sum_{v \in V(G)} |N(v)|}{n}\right)^t$$

Note that $|N(v)|$ is the degree of vertex v. Then summing $|N(v)|$ across all vertices and dividing by n gives the average degree, and so $\frac{\sum_{v \in V(G)} |N(v)|}{n} = d$, from which it follows that

$$\mathbb{E}[X] \geq n^{1-t}\left(\frac{\sum_{v \in V(G)} |N(v)|}{n}\right)^t = n^{1-t}d^t = \frac{d^t}{n^{t-1}}$$

We perform alteration method on the random set $N(T)$ to avoid subsets with few common neighbors. Let Y be the number of subsets $S \subset N(T)$ of size r with $N(S) < m$. For a particular random subset S_{random} of size r in $V(G)$ and a certain $v \in T$, the probability that

S_{random} lies strictly within $N(v)$ is $(|N(S_{random})|/n)$. For $S_{random} \in N(T)$, this condition must hold for all $v \in T$, hence

$$\mathbb{P}(S_{random} \in N(T)) = \Big(\frac{N(S_{random})}{n}\Big)^t \text{ and thus, } \mathbb{E}[Y] = \sum_{S_{random} \subset V(G)} \Big(\frac{N(S_{random})}{n}\Big)^t$$

There are at most $\binom{n}{r}$ subsets S_{random} of size r for which $m > N(S_{random})$. Thus

$$\mathbb{E}[Y] \leq \binom{n}{r}(\frac{m}{n})^t$$

For each subset $S \subset N(T)$ of size r and fewer than m common neighbors, we delete an element in S from $N(T)$. When this process terminates, we are left with

$$\mathbb{E}[X - Y] \geq \frac{d^t}{n^{t-1}} - \binom{n}{r}(\frac{m}{n})^t \geq a$$

vertices in $N(T)$. Moreover, there must no longer exist any subsets of size r and fewer than m common neighbors in $N(T)$. Then by Proposition 3.3, there must exist a random T such that the corresponding set $U = N(T)$ has size $\geq \mathbb{E}[X - Y] \geq a$, and has no subsets of size r with fewer than m neighbors after alteration. ■

Satisfying the density condition of Theorem 12.1 tends to be uncomplicated. Consequentially, the theorem is broadly applicable when graphs are relatively dense. A number of results follow trivially; in particular, consider the following proof of a conjecture of Erdos.

Definition 12.3. A 1-subdivision of a graph H is the graph formed by placing a vertex in the middle of every edge in H; that is, we transform the edge (u, v) to $(u, x), (x, v)$ for some vertex x.

Theorem 12.4. ***Conjectured by Erdos, Resolved by Alon et al. :*** *Let $G = (V, E)$ be a graph with n vertices and εn^2 edges. Then it must be true that G contains the $1-$subdivision of a complete graph with $a = \varepsilon^{3/2} n^{1/2}$ edges.*

Proof. The average degree d of G is just $2\varepsilon n$. Set $r = 2$, $t = \frac{\log n}{2 \log 1/\varepsilon}$, and let m be the number of vertices of the 1-subdivision of the complete graph on a vertices. Observe $m = \binom{a}{2} + a \leq a^2$, and with $\varepsilon \leq 1/2$, computation gives

$$\frac{d^t}{n^{t-1}} - \binom{n}{r}(\frac{m}{n})^t \geq (2\varepsilon)^t n - \frac{n^2}{2}\varepsilon^{3t} = 2^t n^{1/2} - \frac{n^{1/2}}{2} \geq n^{1/2} \geq a$$

So the hypothesis of theorem 12.2 holds. Then by Theorem 12.2, there exists a vertex subset U of G with $|U| = a$ such that every pair of vertices have at least m common neighbors. Now that we have shown there must exist such a set U, we proceed by applying the following lemma:

Lemma 12.5. *Let $H = (A \cup B, F)$ be a bipartite graph in which all edges are of the form $(x \in A, y \in B)$. Suppose $|A| = a$ and $|B| = b$, and let the vertices in B have maximum degree r. If a graph G contains a vertex subset U with $|U| = a$ such that all subsets of U of size r have at least $a + b$ common neighbors, then H is a subgraph of G.*

Proof. We find an embedding in G of H. We start by arbitrarily embedding the vertices in A by randomly mapping them onto vertices in U; that is, we construct an arbitrary injective function $f : A \to U$. Label the vertices of B as $v_1, v_2, ...v_b$. We greedily embed the vertices

of B in this order, one at a time. Suppose that the current vertex to embed is v_i, and let $N(i) \subset A$ be the vertices in H that neighbor i, so that $|N(i)| < r$. Since $f(N_i)$ is a subset of U with size at most r, it has at least $a + b$ common neighbors. Since the current number of embedded vertices is always less than $a + b$, there must exist a vertex $w \in V(G)$ which neighbors all $f(N_i)$ and is not yet used in the embedding. Therefore, we may embed v_i to w. We may repeat this process for all vertices in B, and thus produce a valid embedding. ■

Note that the 1-subdivision of G is just a bipartite graph $(A \cup B, F)$ where A are the original vertices, and B are the vertices generated by the 1-subdivision. Then by Lemma 12.5, we may embed G into U and its neighbors, so we are done. ■

13. Containers

Th study of random combinatorial objects becomes increasingly difficult as the set of potential objects grows. In particular, exponential quantities become increasingly difficult, since they make outliers more likely and are harder to bound with measure concentration inequalities. However, recent work has found a way to circumvent the challenges of exponential numbers of cases in the study of independent sets and hypergraphs through **Containers**.

Much of the recent progress from the probabilistic method is the result of a new technique, known as **The Container Method**. Using Containers–which bound independent sets in (hyper)graphs–a number of fruitful results have been produced.

Remark 13.1. Recall that a hypergraph is a graph $G = (V, E)$ where a particular edge may connect any number of vertices, rather than strictly 2–as is the case in a conventional graph.

Numerous open problems in combinatorics admit a natural rephrasing into finding independent sets in graphs and hypergraphs. In particular, consider the following example:

Definition 13.2. The *Chromatic Number* $\chi(G)$ of a graph is the minimum number of colors needed to proper color its vertices. A coloring is considered "proper" if no edge connects vertices that are all the same color.

Consider a particular coloring of G; if it is proper, then it is analogous to a partition of G into independent sets, from which we color each independent set the same color.

Given this, it is desirable to have simple ways to bound the size and number of independent sets in a particular graph. However, the number of possible independent sets is exponential, since there are 2^n vertex subsets one could examine. But this exponential growth is slightly deceptive, since majority of independent sets tend to cluster together in groups; to understand this, consider a maximal independent set of size n–by removing any number of vertices, we can generate $2^n - 1$ distinct independent sets.

This structural clustering provides the intuition for *Containers*. Intuitively, containers generate a small number of vertex subsets that must necessarily contain these clusters. Through iterative applications of containers, we can gradually constrict these contains until they near-perfectly mimic the independent sets of the graph. Moreover, the individual containers provide a way to isolate and focus on certain cluster. The number of containers turns out to be sub-exponential, and hence it is dramatically easier to study containers as compared to independent sets. In this paper, we present only the Container Lemma for regular graphs; to learn its extension to hypergraphs, see [Mor16].

Definition 13.3. We let $I(H)$ denote the set of all independent sets in a graph H.

Lemma 13.4. ***Graph Container Lemma:*** *for every $c > 0$, there exists a $\delta > 0$ such that the following holds. Let G be a graph with average degree d and maximum degree $\Delta(G) \leq c \cdot g$. Set $r = 2\delta/d$. Then there exists a collection K of subsets of $V(G)$, with*

$$|K| \leq \binom{v(G)}{\lceil v(G)r \rceil}$$

such that

(1) *For every independent set $I \in I(G)$, there exists a $C \in K$ such that $I \subset C$.*
(2) *$|C| \leq (1-\delta)|V(H)|$ for every $C \in K$.*

We postpone a proof of this Lemma until later.

We may construct each container $C \in K$ through the Graph Container Algorithm, which we apply to each independent set $I \in I(G)$.

13.1. **The Graph Container Algorithm.** Input: the graph G, an independent set I.

Output: A partition $V(G) = A(I) \cup S(I) \cup X(I)$.

Let G be a graph, and let $I \in I(G)$ be an independent set. We define 3 sets that partition $V(H)$: the "fingerprint" S, the "available" vertices A, and a set X of "forbidden" vertices, satisfying $A \cup X \cup S = V(G)$. We begin with $A = V(G)$ and S, X empty. While $|X| \leq \delta v(G)$ repeat the following process:

(1) Assign an ordering to the remaining vertices in A based on degree, such that the leftmost vertices in $V(A)$ have the largest degree.
(2) Let u be the leftmost vertex of I in the ordering of A. We move u into S; that is, set $S := S \cup \{u\}$ and $V(A) := V(A) \backslash u$.
(3) Move the neighbors of u into X; that is, set $X := X \cup N(u)$ and $A := A \backslash N(u)$
(4) Remove from A every vertex which precedes u in the ordering of A, and add it to X.

When this process has terminated, define $A(I) := A$, $S(I) := S$, and $G(I) := G$. Output these three sets.

Proof. of the Graph Container Lemma: We claim that running the Graph Container Algorithm for all $I \in I(H)$ and setting $C(S) = S \cup A(I)$ yields a collection

$$K : \{C(S) : S = S(I) \text{ for some } I \in I(G)\}$$

with the desired properties.

First, we show that if I and I' in $I(G)$ have the same fingerprint $S(I') = S(I)$, then $A(I) = A(I')$. Indeed, if we ever query a vertex and find that it is in only one of I and I', then we place it in only one of $S(I)$ and $S(I')$, which implies the sets are different. Hence, since we query the vertices of G in the same order until we find such a vertex, and the algorithm is entirely determined by these queries, it follows that process is identical for the two independent sets, and thus the outcomes $A(I)$ and $A(I')$ are identical. This implies that the fingerprint of the container $C(S)$ uniquely determines the container.

Moreover, since vertices are moved into $X(I)$ only if they are neighbors of some $v \in I$ or if they are queried and found not to be in I, it must be true that $I \cap X(I) = \emptyset$. Hence $I \subset C(S(I)) = A(I) \cup (S(I))$, and so each I is a contained within some set $C(S(I)) \in K$. Since the algorithm terminates with $|X(I)| \geq \delta v(G)$, it follows that $|C| \leq (1-\delta)v(G)$ for all $C \in K$.

It remains to show that $|S(I)| \leq \lceil v(G)r \rceil$, as this implies $|K| \leq \binom{v(H)}{\lceil v(H)r \rceil}$ by the fact that $S(I)$ uniquely determines the container C. Explicitly, if $|S(I)| \leq \lceil v(G)r \rceil$, then there exist at most $\binom{v(H)}{\lceil v(H)r \rceil}$ possible $S(I)$ and hence at most that many possible C.

To prove this, we will demonstrate that when a vertex u is added to $S(I)$, at least $d/2$ vertices are added to $X(I)$. This is sufficient to prove the bound on $|S(I)|$, since after $r \cdot v(G)$ vertices have been added to $S(I)$, we will have $|X(I)| \geq \delta v(G)$, at which point the process terminates.

Suppose that during a certain point in the algorithm, $|X| \leq \delta v(G)$, and the leftmost $u \in I, A$ is less than $d/2$ spots from the left. Let W denote the set of vertices in A to the left of u in the ordering. Note that we may take $|S| \leq rv(G)$, since if the bound holds for $|S|$ of this size, it follows that there cannot exist any $|S|$ greater than this size. Let $e(G[X])$ be the number of edges in G contained within the induced subgraph X. Then since at most $(|S| + |X| + |W|) \cdot \Delta(G)$ edges have an endpoint in S, X, or W, this implies

$$e(G[A \backslash W]) \geq e(G) - (|S| + |X| + |W|) \cdot \Delta(G) \geq e(G) - \big((r + \delta)v(G) + d/2)\big) \cdot \Delta(G)$$

Substituting in $\Delta(G) \leq c \cdot d$ and assuming that $d \leq 2\delta v(G)$, since otherwise $|S(I)| = 1$,

$$e(G[A \backslash W]) \geq e(G) - \big((r + \delta)v(G) + \delta v(G)\big) \cdot c \cdot d$$

Recalling $r = 2\delta/d$ and hence $r \leq 2\delta c$, we simplify

$$e(G[A \backslash W]) \geq e(G) - \big((2c\delta + \delta)v(G) + \delta v(G)\big) \cdot c \cdot d = e(G) - \big(2(c + 1)\big)\delta v(G) \cdot c \cdot d$$

But $v(G) \cdot d = e(G)$, and taking $\delta < 1/(4(c)(c + 1))$,

$$e(G[A \backslash W]) \geq e(G)\big(1 - 2c(c + 1)\delta\big) \geq e(G)/2$$

By definition, u has the highest degree in $A \backslash W$. Then by the pigeonhole principle,

$$\deg(u) \geq e(G[A \backslash W])/|A \backslash W| \geq e(G)/2v(G) = d/2$$

This implies that u has at least $d/2$ neighbors in A, all of which are moved to X when u is queried by the algorithm. Thus, we have proved the claim that when a vertex u is added to $S(I)$, at least $d/2$ vertices are added to $X(I)$. It therefore follows that the Container Algorithm creates a collection K satisfying the conditions of the Container Lemma, and so we are done.

■

14. Closing Remarks

The probabilistic method is based on the fundamentally simple idea that introducing randomness can facilitate non-constructive proofs of existence. Despite this simplicity, the probabilistic method has evolved into a powerful tool across Combinatorics, particularly in Graph Theory. The method has gradually been built upon over time, with new techniques introduced to deal with the shortcomings of older techniques. In particular, the Second Moment Method was introduced as a consequence of the First Moment not revealing the structure of a distribution; Alteration was introduced because purely random constructions were weak and could be further developed upon by manual changes; Containers were introduced to handle the intractable number of cases in the study of independent sets.

Paul Erdős is said to have remarked that probabilistic method would long outlive him, and it is not difficult to see why he felt this way. A simple yet remarkably powerful problem-solving technique, the probabilistic method continues to produce strong results throughout Combinatorics.

An important thing to note is that in mathematics, the goal is frequently not just to solve a problem, but to uncover some profound structural understanding about a particular object. As the probabilistic method is inherently nonconstructive, one might assume it fails in this–if we never describe a particular object, how could we glean some new understanding about its properties?

However, despite being nonconstructive, the probabilistic method *does* give structural understanding, particularly through the problem-solving process. More specifically, the particular probabilistic tool one attacks a question with is largely motivated by structural understanding. For instance, one may only apply the Lovasz Local Lemma when one knows that the object possesses few dependencies–a structural conclusion. Moreover, if a particular probabilistic technique fails, it can offer structural insight. If the Second Moment setup of 7.6 fails, then one knows that the object likely possesses a high number of dependencies and consequentially a relatively high variance.

Herein lies the beauty of the Probabilistic Method: it doesn't substitute in for structural understanding, but instead complements it.

References

[AS11] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley, 2011.

[Car19] Carnegie Mellon University Statistics Ph.D. Program. Lecture notes on the bounded differences inequality. Lecture notes for 36-709: Advanced Statistical Theory I, February 2019. Course instructor: Alessandro Rinaldo.

[FS10] Jacob Fox and Benny Sudakov. Dependent random choice, 2010.

[Mor16] Robert Morris. The method of hypergraph containers, 2016. Lecture notes, São Paulo School of Advanced Science on Algorithms, Combinatorics and Optimization, IME–USP.

[Pen17] Paolo Penna. Lecture 7: Vectors in generic position and a proof of bollobás' theorem. Lecture notes for "Linear Algebra Methods in Combinatorics", ETH Zürich, 2017. Accessed: 2026-07-12.

[Sas26] Igal Sason. The lovász local lemma: Foundations and applications, 2026.