Walk through Combinatorics: Chromatic number of a random graph[*]

(Version 3b: revised 8 October 2021)

A *proper coloring* of a graph G is an assignment of colors to vertices of G such that no edge is monochromatic (its two vertices receive the same color). The *chromatic number* of G is the least number of colors in a proper coloring of G. An alternative way to think of chromatic number is that it is the fewest number of independent sets needed to cover all the vertices of G. The chromatic number of G is traditionally denoted by $\chi(G)$.

Let $G \sim G(n, 1/2)$, i.e G is a graph obtained by picking edges of G with probability $1/2$ independently of the other edges, or simpler put, G is a graph sampled uniformly among all graphs on n vertices. What is the chromatic number of G? It can be anything, of course, but most likely it is about $n/2\log_2 n$.

Theorem 1. *Let* $G \sim G(n, 1/2)$. *Then*

$$\Pr\Big[\chi(G) = \big(1 + o(1)\big)\frac{n}{2\log_2 n}\Big] \to 1 \qquad \textit{as } n \to \infty.$$

The theorem is a combination of a lower bound and an upper bound. We begin with the lower bound, as it is simpler of the two.

A proper coloring is a covering by independent sets. Hence, nonexistence of a proper coloring in at most n/k colors would follow from nonexistence of an independent set of size k. We will chose $k \sim 2\log_2 n$, and will show that G is unlikely to contain an independent set of size k. As independent sets become cliques upon taking the complement, it is equivalent to show that G is unlikely to contain a k-clique[1]

Let X be the number of k-cliques in G. We have

$$\mathbb{E}[X] = \sum_{S \in \binom{[n]}{k}} \Pr\big[S \text{ is a } k\text{-clique}\big] = \binom{n}{k} 2^{-\binom{k}{2}}.$$

[*]These notes are from `http://www.borisbukh.org/DiscreteMath12/notes_chromatic.pdf`

[1]The sole reason we prefer cliques to independent sets is that the former require 9 fewer keystrokes.

Write $f(k)$ for $\binom{n}{k}2^{-\binom{k}{2}}$. As $\Pr[X \geq 1] \leq \mathbb{E}[X]$, it follows that if $f(k) \to 0$, then $\Pr[X \geq 1] \to 0$. Let's find k for which this holds.

The equation $\binom{n}{k_0}2^{-\binom{k_0}{2}} = 1$ can be easily solved approximately: Taking the logarithms we obtain $k_0^2/2 = k_0 \log_2 n + O(k_0 \log k_0)$, which simplifies to $k_0 = 2\log_2 n + O(\log k_0)$. Self-substitution yields $k_0 = 2\log_2 n + O(\log\log n)$. Furthermore if $k = 2\log_2 n + O(\log\log n)$, then

$$f(k)/f(k+1) = \frac{\binom{n}{k}2^{-\binom{k}{2}}}{\binom{n}{k+1}2^{-\binom{k+1}{2}}} = \frac{k+1}{n-k} \cdot 2^k \geq \frac{1}{n} \cdot 2^{\bigl(2+o(1)\bigr)\log_2 n} = n^{1+o(1)}.$$

Hence, $\Pr[X \geq 1] \leq f(k) \to 0$ rapidly as k increases above k_0. This proves the lower bound.

The upper bound is more intricate. We need to show that $G(n,1/2)$ can be covered by $\bigl(1+o(1)\bigr)n/2\log_2 n$ cliques with high probability. There are two clever ideas to the proof. Let $m = \lfloor n/\log^2 n \rfloor$. The first idea is to reduce the problem to showing that the event

"all sets of size m contain a k-clique"

is very likely. Indeed, suppose the event holds. Then we can find the required covering in two stages. In the first stage, we select k-cliques, remove the vertices contained in them, and repeat until fewer than m vertices are left. In the second stage, we treat each vertex as a 1-clique. The total number of cliques would then be at most $n/k + m = n/2\log_2 n + O(n/\log^2 n)$, which is the desired upper bound.

A natural goal is to show that if S is an m-element set, then

$$p_{\text{small}} \stackrel{\text{def}}{=} \Pr\bigl[S \text{ contains no } k\text{-clique}\bigr]$$

is so tiny, as to make

$$\Pr[\exists S \text{ that contains no } k\text{-clique}] \leq \sum_{S \in \binom{[n]}{m}} \Pr\bigl[S \text{ contains no } k\text{-clique}\bigr] = p_{\text{small}}\binom{n}{m}$$

tend to zero. The problem is that $\binom{n}{m}$ is exponential, namely

$$\binom{n}{m} \leq n^m = e^{\Theta(n/\log n)}.$$

Our aim is thus to show that p_{small} is exponentially small, which we will do.

Proof that $\Pr[G(n,1/2)$ **contains no** k**-clique**$] = \exp\big(-n^{2-o(1)}\big)$**.** In this section we write n for what used to be m. Consider $G \sim G(n,1/2)$. Then $p_{\text{small}} = \Pr\big[G \text{ contains no } k\text{-clique}\big]$. The first impulse is to bound p_{small} via Chebyshev's inequality. Let X be the number of k-cliques in G. Then

$$\Pr[X=0] \le \Pr\big[|X-\mathbb{E}[X]| \ge E[X]\big] \le \frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2}.$$

Sadly, this inequality is just too weak because the dependence on $\mathbb{E}[X]$ is only polynomial. It is thus natural to use a large deviation inequality, such as Azuma's inequality.

The difficulty with using Azuma's inequality is that X is far from being Lipschitz. Adding a single edge to G might increase the number of k-cliques by $\binom{n-2}{k-2}$. So, one replaces X by a random variable that behaves similarly to X, but that is Lipschitz. The second clever idea is the choice of this random variable.

Let Y be the maximal number of edge-disjoint k-cliques in G. It is clear that Y is Lipschitz — removing an edge destroys at most one k-clique. It remains to show that $\mathbb{E}[Y]$ is large. Let $\mathcal{F}$ be the family of all the k-cliques in G. Define an auxiliary graph H whose vertex set is $V(H)=\mathcal{F}$ and whose edges are given by the rule that $S_1, S_2 \in \mathcal{F}$ form an edge if k-cliques on S_1 and S_2 share an edge. In symbols, $S_1 \sim_H S_2 \iff |S_1 \cap S_2| \ge 2$. The random variable Y is precisely the independence number of H, and so if we show that H has only a few edges, we can invoke Turán's theorem to bound Y from below.

To use Turán theorem, we need to get an estimate on the number of vertices and on number of edge of H. We can use Chebyshev's inequality to show both $|V(H)|$ and $|E(H)|$ are concentrated near their expectations. Computing the variance of $|V(H)|$ and $|E(H)|$ is annoying, so it is simpler to use only the expectations and import the proof of existence of a large independence set in sparse graphs.

We have $\mathbb{E}\big[|V(H)|\big] = f(k)$, and

$$\begin{aligned}
\mathbb{E}\big[|E(H)|\big] &= \sum_{\substack{S_1,S_2 \in \binom{[n]}{k} \\ |S_1\cap S_2|\ge 2}} \Pr[S_1 \,\&\, S_2 \text{ are } k\text{-cliques}] \\
&= \sum_{S_1} \sum_{t\ge 2} \sum_{\substack{S_2 \\ |S_1\cap S_2|=t}} 2^{-2\binom{k}{2}+\binom{t}{2}} \\
&= \binom{n}{k} \sum_{t\ge 2} \binom{k}{t}\binom{n-k}{k-t} 2^{-2\binom{k}{2}+\binom{t}{2}} \\
&= f(k) 2^{-\binom{k}{2}} \sum_{t\ge 2} g(t),
\end{aligned}$$

where
$$g(t) = \binom{k}{t}\binom{n-k}{k-t}2^{\binom{t}{2}}.$$

The main contribution to the sum above is $g(2)$, for the other terms are much smaller. First of all, the term on the other extreme, $g(k)$ is smaller by direct calculation:
$$g(2)/g(k) = 2^{-\binom{k}{2}}\binom{k}{2}\binom{n-k}{k-2} = \binom{k}{2}f(k)\frac{\binom{n-k}{k-2}}{\binom{n}{k}} \approx f(k)n^{-2}k^4.$$
We shall choose k so that $f(k) = \Omega(n^3)$ to make this ratio $o(1)$.

It remains to show that $g(t)$ for $2 < t < k$ are small. For that we compute the ratio of consecutive terms to be
$$\frac{g(t)}{g(t+1)} = \frac{\binom{k}{t}\binom{n-k}{k-t}2^{\binom{t}{2}}}{\binom{k}{t+1}\binom{n-k}{k-t-1}2^{\binom{t+1}{2}}} = \frac{t+1}{k-t}\cdot\frac{n-2k+t+1}{k-t}2^{-t}.$$
This ratio is approximately $2^{-t}n/\log_2^2 n$ if t is small, and so the terms of the sum above decrease super-geometrically at first. At about $t = \log_2 n$ the ratio drops below 1, and terms of the sum start increasing. When t is close to $2\log_2 n$ the ratio is about n^{-1}. Hence, $g(2)$ is indeed the dominant term, and we conclude that
$$\mathbb{E}\bigl[|E(H)|\bigr] \approx 2f(k)2^{-\binom{k}{2}}\binom{k}{2}\binom{n-k}{k-2} \le f(k)^2k^4n^{-2}$$
provided that $f(k) = \Omega(n^3)$.

We are ready to select the independent set. Let $0 < p < 1$ be a parameter to be chosen later. Pick each vertex of H with probability p at random independently of the others. Let U be the resulting set. Let B be a set of vertices, one from each edge spanned by U. Then $U \setminus B$ is an independent set. We have
$$\mathbb{E}\bigl[|U \setminus B|\bigr] = pE\bigl[|V(G)|\bigr] - p^2E\bigl[|E(G)|\bigr] \ge pf(k) - p^2f(k)^2k^4n^{-2}.$$
We optimize this by choosing $p = \frac{1}{2}n^2k^{-4}f(k)^{-1}$, and obtain
$$\mathbb{E}[Y] \ge \mathbb{E}\bigl[|U \setminus B|\bigr] \ge \tfrac{1}{4}k^{-4}n^2.$$

Hence, by Azuma's inequality applied to the natural edge-exposure martingale we have
$$\Pr[Y=0] \le \Pr\bigl[|Y-\mathbb{E}[Y]| \ge \mathbb{E}[Y]\bigr] \le \exp\Bigl(-\mathbb{E}[Y]^2/2\binom{n}{2}\Bigr) \le \exp\bigl(-Cn^2/\log^8 n\bigr).$$

Conclusion of the proof It is simple:
$$p_{\text{small}}\binom{n}{m} \le \exp\bigl(-Cm^2/\log^8 m\bigr)2^n = \exp\bigl(-n^2/\log^{12} n\bigr)2^n < 1.$$