Asymptotic performance of the Grimmett–McDiarmid heuristic

Yuval Filmus

December 10, 2019

Abstract

Grimmett and McDiarmid suggested a simple heuristic for finding stable sets in random graphs. They showed that the heuristic finds a stable set of size $\sim \log_2 n$ (with high probability) on a $G(n, 1/2)$ random graph. We determine the asymptotic distribution of the size of the stable set found by the algorithm.

1 Introduction

Grimmett and McDiarmid [GM75] considered the problem of coloring $G(n, 1/2)$ random graphs. As part of their solution, they suggested the following simple heuristic for finding a large stable set: scan the vertices in random order, adding to the stable set any vertex which is not adjacent to the vertices added so far. They showed that this heuristic constructs a stable set of size asymptotically $\log_2 n$ (with high probability), in contrast to the maximum stable set, whose size is asymptotically $2 \log_2 n$ (with high probability).

Let us briefly indicate how to analyze the algorithm (for more details, consult any lecture notes on the subject). Denote by N_k the number of remaining vertices not adjacent to the first k vertices in the stable set constructed by the algorithm, or zero if the algorithm terminated before choosing k vertices. A simple induction shows that $\mathbb{E}[N_k] \leq n/2^k$, and so with high probability, the algorithm produces a stable set of size at most $\log_2 n + f(n)$, where $f(n)$ is *any* function satisfying $f(n) \to \infty$.

For the lower bound, let us imagine that there are infinitely many vertices (this idea already appears in [GM75]), let $i_0 = 0$, and let i_k be the index of the k'th chosen vertex in the random order of the vertices (starting with 1). Then $i_{k+1} - i_k \sim G(2^{-k})$ (geometric random variable with success probability 2^{-k}), and the size of the clique is the maximal k such that $i_k \leq n$. It is easy to calculate $\mathbb{E}[i_k] = 2^k - 1$, from which it easily follows that with high probability, the algorithm produces a stable set of size at least $\log_2 n - f(n)$, where $f(n)$ is *any* function satisfying $f(n) \to \infty$.

Let $\mathbf{k}$ be the size of the stable set produced by the algorithm. The foregoing suggests that $\mathbf{k} - \log_2 n$ approaches a limiting distribution, but there is a complication: $\mathbf{k}$ is always an integer, while the fractional part of $\log_2 n$ varies. We will show that if we fix the fractional part $\{\log_2 n\}$ then $\mathbf{k} - \log_2 n$ indeed approaches a limit; and furthermore, the various limits stem from the same continuous distribution.

Definition 1.1. The random variable $\mathbf{H}$ is given by the following sum of exponential distributions:

$$\mathbf{H} = \sum_{i=1}^{\infty} E(2^i).$$

(This defines a random variable due to Kolmogorov's two-series theorem.)

Theorem 1.2. *For a given n, define*

$$p_k = \Pr[\mathbf{k} = k], \quad q_k = \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right].$$

Then we have

$$\sum_{k=0}^{\infty} |p_k - q_k| = o(1).$$

Preliminaries The Wasserstein distance $W_1(X,Y)$ between two random variables is the minimum of $\mathbb{E}[|X-Y|]$ over all couplings of X,Y. This formula shows that $W_1(X_1+X_2,Y_1+Y_2) \le W_1(X_1,Y_1)+W_1(X_2,Y_2)$. The Wasserstein distance is also given by the explicit formula

$$W_1(X,Y) = \int_{-\infty}^{\infty} |\Pr[X<t]-\Pr[Y<t]|\,dt.$$

The Kolmogorov–Smirnov distance between X and Y is $\sup_t |\Pr[X<t]-|Pr[Y<t]|$. If Y is a continuous random variable with density bounded by C, then the Kolmogorov–Smirnov distance between X and Y is bounded by $2\sqrt{CW_1(X,Y)}$.

2 Proof

Recall that $\mathbf{k}$ is the size of the stable set produced by the Grimmett–McDiarmid algorithm. Grimmett and McDiarmid proved the following result, whose proof was outlined in the introduction.

Lemma 2.1.

$$\Pr[\mathbf{k}<k] = \Pr[G(1)+G(1/2)+\cdots+G(1/2^{k-1})>n] = \Pr[G(1/2)+\cdots+G(1/2^{k-1})\ge n].$$

Our main idea is to rewrite this formula as follows:

$$\Pr[\mathbf{k}<k] = \Pr\left[\frac{G(1/2^{k-1})}{n}+\frac{G(1/2^{k-2})}{n}+\cdots+\frac{G(1/2)}{n}\ge 1\right]. \tag{1}$$

It is known that the distribution $G(c/n)/n$ tends (in an appropriate sense) to an exponential random variable $E(c)$. We will show this quantitatively, in terms of the Wasserstein metric W_1.

Lemma 2.2. *If $p\le 1/2$ then*

$$W_1(G(p)/n, E(pn)) = O\left(\frac{1}{n}\right).$$

Proof. Let $X=\lceil E(pn)n\rceil$. Then for integer t,

$$\Pr[X\ge t] = \Pr[E(pn)>(t-1)/n] = e^{-p(t-1)}.$$

In contrast,

$$\Pr[G(p)\ge t] = (1-p)^{t-1}.$$

By construction, $W_1(X/n,E(pn))\le 1/n$, and so

$$W_1(G(p)/n,E(pn)) \le \frac{1}{n}+W_1(G(p)/n,X/n) \le \frac{1}{n}+\int_0^{\infty}|\Pr[G(p)/n\ge s]-\Pr[X/n\ge s]|\,ds = \frac{1}{n}+\frac{1}{n}\sum_{r=0}^{\infty}|\Pr[G(p)\ge r]-\Pr[X\ge r]| = \frac{1}{n}+\frac{1}{n}\sum_{t=1}^{\infty}|(1-p)^t-e^{-pt}|.$$

Since $p\le 1/2$, we have $-p-O(p^2)\le \log(1-p)\le -p$, and so

$$e^{-pt-O(p^2t)} \le (1-p)^t \le e^{-pt}.$$

Therefore

$$|(1-p)^t-e^{-pt}| = e^{-pt}(1-e^{-O(p^2t)}) = O(p^2te^{-pt}).$$

We can thus bound

$$\sum_{t=1}^{\infty}|(1-p)^t-e^{-pt}| \le O(p^2)\sum_{t=1}^{\infty}\frac{t}{e^{pt}} = O\left(\frac{p^2e^p}{(e^p-1)^2}\right) = O(1). \qquad \square$$

Since W_1 is subadditive, we immediately conclude the following:

Lemma 2.3. *Let* $\mathbf{G}$ *be the random variable appearing in* (1). *Then*

$$W_1\left(\frac{n}{2^k}\mathbf{G}, \mathbf{H}\right) = O\left(\frac{k}{2^k}\right).$$

Proof. Lemma 2.2 shows that

$$W_1(\mathbf{G}, E(n/2^{k-1}) + \cdots + E(n/2)) = O\left(\frac{k}{n}\right),$$

which implies that

$$W_1\left(\frac{n}{2^k}\mathbf{G}, E(2) + \cdots + E(2^{k-1})\right) = O\left(\frac{k}{2^k}\right).$$

On the other hand,

$$W_1\left(\sum_{\ell=k}^{\infty} E(2^\ell), \mathbf{0}\right) = \mathbb{E}\left[\sum_{\ell=k}^{\infty} E(2^\ell)\right] = \frac{1}{2^{k-1}},$$

where $\mathbf{0}$ is the constant zero random variable. The lemma follows. $\square$

In order to convert this bound to a bound on the Kolmogorov–Smirnov distance, we need to know that $\mathbf{H}$ is continuous and has a bounded density function.

Lemma 2.4. *The random variable* $\mathbf{H}$ *is continuous, and has a bounded density function* f*:*

$$f(x) = 2C^{-1}\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2^i x}\prod_{r=1}^{i-1}\frac{2}{2^r-1}, \text{ where } C = \prod_{s=1}^{\infty}(1-2^{-s}) > 0.$$

(The constant C *is the limit of the probability that an* $n \times n$ *matrix over* $GF(2)$ *is regular.)*

Proof. Let $\mathbf{H}^{(\ell)} = \sum_{i=1}^{\ell} E(2^i)$. It is well-known that the density of $\mathbf{H}^{(\ell)}$ is

$$f_\ell(x) = \sum_{i=1}^{\ell} 2^i e^{-2^i x} K_{\ell,i}, \text{ where } K_{\ell,i} = \prod_{\substack{j=1 \\ j \neq i}}^{\ell} \frac{2^j}{2^j - 2^i}.$$

Note that

$$K_{\ell,i} = (-1)^{i-1}\prod_{j=1}^{i-1}\frac{1}{2^{i-j}-1} \times \prod_{j=i+1}^{\ell}\frac{1}{1-2^{i-j}} = (-1)^{i-1}\prod_{r=1}^{i-1}\frac{1}{2^r-1} \times \prod_{s=1}^{\ell-i}\frac{1}{1-2^{-s}}.$$

We can therefore write

$$f_\ell(x) = \sum_{i=1}^{\ell} 2e^{-2^i x} \times (-1)^{i-1}\prod_{r=1}^{i-1}\frac{2}{2^r-1} \times \prod_{s=1}^{\ell-i}\frac{1}{1-2^{-s}}.$$

This allows us to bound

$$|f_\ell(x)| \le 2C^{-1}e^{-2x}\sum_{i=1}^{\ell}\prod_{r=1}^{i-1}\frac{2}{2^r-1},$$

where C is the constant in the statement of the lemma. Bounding the sum by a geometric series, we conclude that $|f_\ell(x)| = O(e^{-2x})$, where the bound is independent of ℓ. Applying dominated convergence, we obtain the formula in the statement of the lemma. $\square$

Armed with this information, we can finally estimate $\Pr[\mathbf{k} < k]$.

Lemma 2.5.

$$\Pr[\mathbf{k} < k] = \Pr\left[\mathbf{H} \ge \frac{n}{2^k}\right] \pm O\left(\sqrt{\frac{k}{2^k}}\right).$$

Proof. Since $\mathbf{H}$ has bounded density by Lemma 2.4, we can bound the Kolmogorov–Smirnov distance between $\frac{n}{2^k}\mathbf{G}$ and $\mathbf{H}$ by $O(\sqrt{W_1(\frac{n}{2^k}\mathbf{G}, \mathbf{H})}) = O(\sqrt{k/2^k})$, using Lemma 2.3. It follows that

$$\Pr[\mathbf{k} < k] = \Pr\left[\frac{n}{2^k}\mathbf{G} \geq \frac{n}{2^k}\right] = \Pr\left[\mathbf{H} \geq \frac{n}{2^k}\right] \pm O\left(\sqrt{\frac{k}{2^k}}\right). \qquad \square$$

Theorem 1.2 now easily follows:

Proof of Theorem 1.2. Lemma 2.5 shows that for each k,

$$\Pr[\mathbf{k} = k] = \Pr[\mathbf{k} < k+1] - \Pr[\mathbf{k} < k] = \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right] \pm O\left(\sqrt{\frac{k}{2^k}}\right).$$

This implies that

$$\sum_{k=\ell}^{\infty} \left|\Pr[\mathbf{k} = k] - \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right]\right| = O\left(\sqrt{\frac{\ell}{2^\ell}}\right).$$

Lemma 2.1 shows that

$$\Pr[\mathbf{k} < \ell] = \Pr[G(1/2) + \cdots + G(1/2^{\ell-1}) \geq n] \leq \frac{\mathbb{E}[G(1/2) + \cdots + G(1/2^{\ell-1})]}{n} < \frac{2^\ell}{n},$$

and so choosing $\ell := \frac{2}{3}\log_2 n$, we have

$$\Pr[\mathbf{k} < \ell] \leq \frac{1}{n^{1/3}}.$$

Lemma 2.5 shows that

$$\Pr\left[\mathbf{H} \geq \frac{n}{2^\ell}\right] = O\left(\frac{\sqrt{\log n}}{n^{1/3}}\right),$$

and so

$$\sum_{k=0}^{\ell-1} \left|\Pr[\mathbf{k} = k] - \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right]\right| \leq \sum_{k=0}^{\ell-1} \left(\Pr[\mathbf{k} = k] + \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right]\right) = O\left(\frac{\sqrt{\log n}}{n^{1/3}}\right).$$

In total, we conclude that

$$\sum_{k=0}^{\infty} \left|\Pr[\mathbf{k} = k] - \Pr\left[\frac{n}{2^{k+1}} \leq \mathbf{H} < \frac{n}{2^k}\right]\right| = O\left(\frac{\sqrt{\log n}}{n^{1/3}}\right). \qquad \square$$

We can also express Theorem 1.2 in terms of the variation distance between $\mathbf{k}$ and an appropriate random variable.

Let $\theta = \{\log_2 n\} = \log_2 n - \lfloor \log_2 n \rfloor$, and let $k = \lfloor \log_2 n \rfloor + c$. Then $n/2^k = 2^{\theta - c}$, and so the quantity q_k in Theorem 1.2 is

$$\Pr[2^{-(c+1)} \leq 2^{-\theta}\mathbf{H} < 2^{-c}] = \Pr[2^{-(c+1)} < 2^{-\theta}\mathbf{H} \leq 2^{-c}] = \Pr[\lfloor \log_2(1/\mathbf{H}) + \theta \rfloor = c].$$

Therefore we obtain the following corollary:

Corollary 2.6. *For a given n, let $\theta = \{\log_2 n\}$ and define*

$$\mathbf{h} = \lfloor \log_2(1/\mathbf{H}) + \theta \rfloor.$$

The variation distance between $\mathbf{k}$ and $\mathbf{h}$ is at most $\tilde{O}(1/n^{1/3})$.

The random variable $\log_2(1/\mathbf{H})$ has density

$$g(y) = (2C^{-1}\ln 2)2^{-y} \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2^{i-y}} \prod_{r=1}^{i-1} \frac{2}{2^r - 1},$$

and is plotted in Fig. 1.

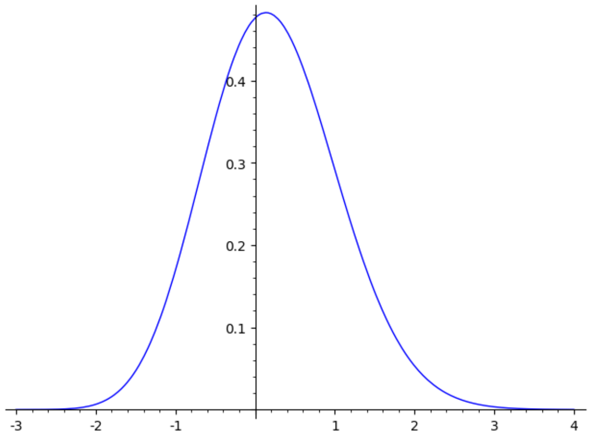

Figure 1: Density of $\log_2(1/\mathbf{H})$

3 Applications

Integrating the formula given in Lemma 2.4, we obtain the following estimate via Lemma 2.5:

$$Pr[\mathbf{k}=k] \approx C^{-1}\sum_{i=1}^{\infty}(-1)^{i-1}\left(e^{-n2^{i-k-1}} - e^{-n2^{i-k}}\right)\prod_{r=1}^{i-1}\frac{1}{2^r-1},$$

where the error is $O(k/2^k)$. If $k=\log_2 n + c$, then this becomes

$$Pr[\mathbf{k}=\log_2 n + c] \approx C^{-1}\sum_{i=1}^{\infty}(-1)^{i-1}\left(e^{-2^{i-c-1}} - e^{-2^{i-c}}\right)\prod_{r=1}^{i-1}\frac{1}{2^r-1}.$$

Using this, we can calculate the limiting distribution of $\mathbf{k}$, fixing $\{\log_2 n\}$. For example, if n is a power of 2 then we obtain the following limiting distribution:

c	$\lim \Pr[\mathbf{k}=\log_2 n + c]$
-4	0.000000389680708123307
-3	0.00116084271918975
-2	0.0610996920580558
-1	0.343335642221465
0	0.420730421531672
1	0.153255882765631
2	0.0194547690538043
3	0.000943671851018291
4	0.0000185343323798604
5	0.000000153237063593714

In this case, the expected deviation of $\mathbf{k}$ from $\log_2 n$ is -0.273947769982407, and the standard deviation of $\mathbf{k}$ is 0.763009254799132.

References

[GM75] G. R. Grimmett and C. J. H. McDiarmid. On colouring random graphs. *Math. Proc. Cambridge Philos. Soc.*, 77:313–324, 1975.