CS271 Randomness & Computation **Spring 2020**

Lecture 7: February 11

Instructor: Alistair Sinclair

7.1 The clique number of a random graph [cont.]

We begin by sketching a proof of the following theorem from last lecture:

Theorem 7.1 *For $G \in \mathcal{G}_{n,p}$ for any constant $p \in (0,1)$, the clique number of G is almost surely $\sim 2\log_{1/p}(n)$.*

Proof: We restrict attention to the case $p = 1/2$; the generalization to arbitrary p is straightforward. Throughout this proof, all logarithms are base 2.

Let X_k be the number of k-cliques in a graph G drawn from $\mathcal{G}_{n,\frac{1}{2}}$.

By linearity of expectation we have $g(k) := \mathrm{E}X_k = \binom{n}{k}2^{-\binom{k}{2}}$ and let us define $k_0(n)$ to be the largest value of k such that $g(k) \geq 1$.

An easy calculation (Exercise!) shows that $k_0(n) \sim 2\log n$. (To see this is plausible, note that for $k \ll n$, $g(k)$ is roughly $\frac{n^k}{k!}2^{-k^2/2} \sim 2^{k\log n - k^2/2 - k\log k}$.)

Now let c be some small integer constant (independent of n) to be decided later. In order to prove our claim, it is enough to show the following, for some constant c:

(1) for $k_1(n) = k_0(n) + c$: $\quad \Pr[X_{k_1(n)} > 0] \to 0$ as $n \to \infty$

(2) for $k_2(n) = k_0(n) - c$: $\quad \Pr[X_{k_2(n)} > 0] \to 1$ as $n \to \infty$

Note first that, if we work with expectations, then the above two claims are not hard to prove. Specifically, we can see that

(a) $\mathrm{E}X_{k_1(n)} \to 0$ as $n \to \infty$

(b) $\mathrm{E}X_{k_2(n)} \to \infty$ as $n \to \infty$

We will not rigorously prove these claims. However, they follow fairly easily by examining the rate of change of the function $g(k)$ near $k = k_0$. In particular, note that for $k \sim 2\log n$ we have

$$\frac{g(k+1)}{g(k)} = \frac{n-k}{k+1}2^{-k} \sim \frac{n}{2\log n}n^{-2} \to 0 \text{ as } n \to \infty.$$

This indicates that, as $n \to \infty$, in the neighborhood of $k_0(n)$ the graph of $g(k)$ decreases sharply, so for some constant c, $g(k_0(n)+c) \to 0$ as $n \to \infty$ and $g(k_0(n)-c) \to \infty$ as $n \to \infty$. The details are left as an Exercise. We illustrate the behavior of $g(k)$ near k_0 by depicting $\log g(k)$ in the neighborhood of k_0 (figure 7.1).

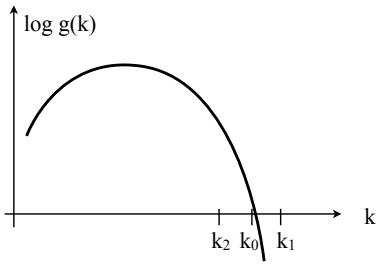

Figure 7.1: The descent of $\log g(k)$ near $k_0(n)$ for fixed n. Note that the domain of $g(k)$ is actually the integers, but we show an interpolated approximation for readability.

Now, we can go ahead and prove claims (1) and (2).

Claim (1) is almost immediate: since X_k is integer valued, and using Markov's inequality, we have

$$\Pr[X_{k_1} > 0] = \Pr[X_{k_1} \geq 1] \leq \mathrm{E}X_{k_1} \to 0 \quad \text{as } n \to \infty.$$

To prove Claim (2), we will use the second moment method to show that $\Pr[X_{k_2} = 0] \to 0$ as $n \to \infty$. Applying Chebyshev's inequality as we did for the case of 4-cliques earlier, we have

$$\Pr[X_{k_2} = 0] \leq \Pr[|X_{k_2} - \mathrm{E}X_{k_2}| \geq \mathrm{E}X_{k_2}] \leq \frac{\mathrm{Var}(X_{k_2})}{(\mathrm{E}X_{k_2})^2}.$$

Thus, it is enough to show that $\frac{\mathrm{Var}(X_{k_2})}{(\mathrm{E}X_{k_2})^2} \to 0$ as $n \to \infty$. For ease of notation, let us denote the random variable X_{k_2} by X and, for every subset S of the vertices of G of size k_2, let us denote by X_S the following random variable:

$$X_S = \begin{cases} 1, & \text{if } S \text{ is a clique;} \\ 0, & \text{if } S \text{ is not a clique.} \end{cases}$$

Clearly, $X = \sum_S X_S$. We now follow a similar path to our treatment of 4-cliques earlier, but in more generality and more systematically. First, we have

$$\mathrm{Var}(X) = \mathrm{Var}(\sum_S X_S) = \sum_S \mathrm{Var}(X_S) + \sum_{S \neq T} \mathrm{Cov}(X_S, X_T).$$

Writing $S \sim T$ to denote the fact that X_S, X_T are *not* independent, we have that $\mathrm{Cov}(X_S, X_T) = 0$ unless $X_S \sim X_T$. In our current application, $S \sim T$ iff $S \neq T$ and $|S \cap T| \geq 2$. Continuing:

$$\begin{aligned} \mathrm{Var}(X) &= \sum_S \mathrm{Var}(X_S) + \sum_{S \sim T} \mathrm{Cov}(X_S, X_T) \\ &\leq \sum_S \mathrm{E}(X_S^2) + \sum_{S \sim T} \mathrm{E}(X_S X_T) \\ &= \sum_S \mathrm{E}X_S + \sum_{S \sim T} \mathrm{E}(X_S X_T) \qquad \text{(since } X_S \text{ is 0/1 valued)} \\ &= \mathrm{E}X + \sum_{S \sim T} \mathrm{E}(X_S X_T). \end{aligned}$$

It follows that $\frac{\mathrm{Var}(X)}{(\mathrm{E}X)^2} \leq \frac{1}{\mathrm{E}X} + \frac{1}{(\mathrm{E}X)^2} \sum_{S \sim T} \mathrm{E}(X_S X_T)$. Thus, since $\mathrm{E}X \to \infty$, it is enough for us to prove that $\sum_{S \sim T} \mathrm{E}(X_S X_T) = o((\mathrm{E}X)^2)$.

We can go further than this using *symmetry* of the events X_S.[1] We have

$$\begin{aligned}
\sum_{S\sim T} \mathrm{E}(X_S X_T) &= \sum_{S\sim T} \Pr[X_S = 1 \wedge X_T = 1] \\
&= \sum_{S\sim T} \Pr[X_S = 1]\Pr[X_T = 1 | X_S = 1] \\
&= \sum_{S} \left(\Pr[X_S = 1] \sum_{T:T\sim S} \Pr[X_T = 1 | X_S = 1] \right) \\
&= \sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1] \sum_{S} \Pr[X_S = 1] \quad \text{(for any fixed } S_0\text{, by symmetry)} \\
&= \mathrm{E}X \sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1],
\end{aligned}$$

and so

$$\frac{\mathrm{Var}(X)}{(\mathrm{E}X)^2} \le \frac{1}{\mathrm{E}X} + \frac{1}{\mathrm{E}X} \sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1].$$

Thus, since $\mathrm{E}X \to \infty$, to conclude our proof it is enough to show that $\sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1] = o(\mathrm{E}X)$ for a fixed S_0. Note that the above argument is very general and applies in any symmetric setting.

Specializing now to our clique problem, and letting S_0 be any fixed set of vertices of size $k_2(n)$, we have

$$\sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1] = \sum_{i=2}^{k_2-1} \binom{k_2}{i} \binom{n-k_2}{k_2-i} 2^{-\left[\binom{k_2}{2} - \binom{i}{2}\right]},$$

where the sum on the right hand side is taken over $i = |T \cap S_0|$. The factor $\binom{k_2}{i}$ counts the number of ways that T can choose i vertices of S_0 to share; the factor $\binom{n-k_2}{k_2-i}$ counts the possible ways that T can choose its other vertices; and $2^{-\left[\binom{k_2}{2} - \binom{i}{2}\right]}$ is the probability that all edges in T that are not shared with S_0 appear.

This leads to the following:

$$\begin{aligned}
\frac{\sum_{T:T\sim S_0} \Pr[X_T = 1 | X_{S_0} = 1]}{\mathrm{E}X} &= \frac{\sum_{i=2}^{k_2-1} \binom{k_2}{i}\binom{n-k_2}{k_2-i} 2^{-\left[\binom{k_2}{2} - \binom{i}{2}\right]}}{\binom{n}{k_2} 2^{-\binom{k_2}{2}}} \\
&= \sum_{i=2}^{k_2-1} \frac{\binom{k_2}{i}\binom{n-k_2}{k_2-i}}{\binom{n}{k_2}} 2^{\binom{i}{2}} \\
&= \sum_{i=2}^{k_2-1} f(i), \qquad \text{where } f(i) := \frac{\binom{k_2}{i}\binom{n-k_2}{k_2-i}}{\binom{n}{k_2}} 2^{\binom{i}{2}} \\
&\le k_2 \cdot \max_{2\le i\le k_2-1} \{f(i)\}.
\end{aligned}$$

Now it can be shown that, provided the constant c is chosen large enough, $f(i)$ is maximized at $i = 2$. (This is an optional exercise for those who enjoy combinatorial analysis!) Its value there is

$$f(2) = \frac{\binom{k_2}{2}\binom{n-k_2}{k_2-2} 2^{\binom{2}{2}}}{\binom{n}{k_2}} \sim \frac{k_2^2 \cdot k_2^2}{n^2} \quad (\text{since } k_2 \ll \sqrt{n}).$$

Since $k_2 \sim 2\log n$, we have $k_2 f(2) \sim k_2^5/n^2 \to 0$ as $n \to \infty$, as required. This concludes the proof of Claim (2) and the proof of the theorem. ■

[1] i.e., there is an automorphism of the probability space that maps any X_S onto any other

7.2 Random k-SAT

Background and Motivation Determining whether a k-CNF formula is satisfiable is an NP-complete problem. However, the problem may be easier to solve for *random* formulae. This is because of the empirical observation that a random formula with a certain density of clauses is either almost surely satisfiable or almost surely unsatisfiable depending on whether the density is below or above a certain critical value. There are ties between this threshold phenomenon and so-called "spin glasses" in statistical physics; indeed, some people believe that this connection may one day unlock the P vs NP question.

Model We now define a model. $\varphi_k(n, rn)$ is a k-CNF formula over n variables generated by choosing rn clauses independently and uniformly at random. What exactly do we mean by choosing the clauses randomly? The two standard possibilities are (i) to choose the k literals in the clause independently and u.a.r. from the $2n$ possibilities (so that repetitions of variables are allowed); or (ii) to choose the clause u.a.r. from the set of all $2^k\binom{n}{k}$ "valid" k-clauses (i.e., without repeated variables). It turns out that most results hold for either model, so we will use the most convenient implementation in each application.

The above model has just one parameter, r, which we can think of as the *density* of the formula. Obviously, as r increases there are more constraints so it becomes less likely that $\varphi_k(n, rn)$ is satisfiable.

Conjecture 7.2 *For every $k \geq 2$, there exists a threshold value $r_k^* \in \mathbb{R}$ such that*

$$r > r_k^* \Rightarrow \Pr[\varphi_k(n, rn) \text{ is satisfiable}] \to 0 \text{ as } n \to \infty$$

$$r < r_k^* \Rightarrow \Pr[\varphi_k(n, rn) \text{ is satisfiable}] \to 1 \text{ as } n \to \infty$$

This conjecture has been a major open problem for at least 30 years. The first landmark result was the following, which resolves the case $k = 2$.

Theorem 7.3 [CR92, FdlV92, G96] *The threshold r_2^* exists and is equal to 1.*

Some years later, a weaker form of the conjecture was proved by Friedgut, who showed the existence of a *non-uniform* threshold for every k:

Theorem 7.4 [F99] *For every $k \geq 2$, there exists a sequence $\{r_k(n)\}$ such that $\forall \epsilon > 0$*

$$r > (1+\epsilon) r_k(n) \Rightarrow \Pr[\varphi_k(n, rn) \text{ is satisfiable}] \to 0 \text{ as } n \to \infty$$

$$r < (1-\epsilon) r_k(n) \Rightarrow \Pr[\varphi_k(n, rn) \text{ is satisfiable}] \to 1 \text{ as } n \to \infty.$$

The crucial difference between Theorem 7.4 and Conjecture 7.2 is that in the conjecture r_k^* is a constant, while in the theorem $r_k(n)$ may depend on n. Despite this weakness, Friedgut's result was a surprising and significant step.

Following a long sequence of developments by several authors, a major breakthrough was achieved rather recently in a technical tour-de-force by Ding, Sly and Sun:

Theorem 7.5 [DSS14] *For all sufficiently large k, the threshold r_k^* exists.*

The actual value of the threshold is also given in [DSS14] as the minimum of an explicit function. As had been previously established by Coja-Oghlan and Panagiotou [CP16], the asymptotic form of r_k^* is

$$r_k^* = 2^k \ln 2 - \tfrac{1}{2}(1 + \ln 2) - \varepsilon_k,$$

where $\varepsilon_k \to 0$ as $k \to \infty$.

For small values of $k > 2$, the conjecture remains open. There has been much work establishing bounds on r_k^* (assuming it exists) for the important case of $k = 3$ (3-SAT). For example, it is known that $3.52 < r_3^* < 4.49$, and based on large-scale experiments and heuristic arguments, r_3^* is believed to be close to 4.2. The upper bound here is due to [DKMP09], and the lower bound to [KKL03,HS03].

Upper Bounds on r_k^*. A more or less tight upper bound on r_k^* can be found by a simple application of the first moment method. Let $X =$ the number of satisfying assignments of $\varphi_k(r, rn)$. Then $X = \sum_\sigma X_\sigma$ where

$$X_\sigma = \begin{cases} 1, & \text{if assignment } \sigma \text{ satisfies } \varphi \\ 0, & \text{if assignment } \sigma \text{ does not satisfy } \varphi \end{cases}$$

Assuming the model in which every literal is independent, we clearly have $\mathrm{E}[X_\sigma] = (1-2^{-k})^{rn}$, and therefore $\mathrm{E}[X] = 2^n(1-2^{-k})^{rn}$. Now it is easy to see that, if $r \geq 2^k \ln 2$, $\mathrm{E}[X] \to 0$ as $n \to \infty$. By Markov's inequality, this immediately yields:

Corollary 7.6 $r_k^* > 2^k \ln 2$.

Lower Bounds on r_k^*. A natural way to prove lower bounds is via algorithms: if a certain algorithm finds a satisfying assignment with high probability in some range of r, then obviously a satisfying assignment must exist in this range. However, bounds found in this way (see, e.g., [CF90]) usually have the form $r_k^* \geq c\frac{2^k}{k}$ for some constant c, which is a factor $O(k)$ off from the upper bound. An important step was the introduction of *non-algorithmic* proofs based on the second moment method [AM02,AP04] to yield the much better bound

$$r_k^* \geq 2^k \ln 2 - k \qquad \text{for all } k \geq 3.$$

This result already bounds the value of r_k^* up to a linear term in k.

The works [CP16,DSS14] mentioned earlier use more sophisticated non-algorithmic ideas inspired by statistical physics to obtain the final tight bound. The gap between algorithmic and non-algorithmic lower bounds suggests that there exist two different thresholds: the first separating those densities for which it is easy to *find* a satisfying assignment w.h.p. (and therefore one necessarily exists) and those for which a satisfying assignment exists w.h.p. but is hard to find; the second threshold separating the latter densities from those for which no satisfying assignment exists w.h.p.

In the next lecture we will sketch the second moment approach to this lower bound. For now, we return to the proof of theorem 7.3.

Proof of Theorem 7.3: We begin with the lower bound, namely that if $r \leq 1-\epsilon$ then, with high probability, a random 2-SAT formula with $m = rn \leq (1-\epsilon)n$ clauses is satisfiable. Let us recall the well-known fact (easy **exercise**) that a 2-SAT formula is satisfiable if and only if its *graph of implications* does not contain a (directed) cycle that contains both a variable and its negation. (The graph of implications contains a vertex for every literal and, for every clause $(l_1 \vee l_2)$, the edges $(\bar{l}_1, l_2)$ and $(\bar{l}_2, l_1)$. For example, the edges induced by clause $(x_1 \vee x_2)$ are as shown in Figure 7.2.)

We will actually work with a slightly stronger sufficient condition which is easier to quantify, namely: the formula is satisfiable if it does not contain a *bicycle*, which is defined as follows.

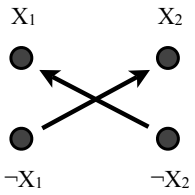

Figure 7.2: The edges induced by clause $(x_1 \vee x_2)$.

Definition 7.7 *For $k \geq 2$, a* bicycle *is a path $(u, w_1, w_2, \ldots, w_k, v)$ in the graph of implications, where the w_i's are literals on* distinct *variables and $u, v \in \{w_i, \bar{w}_i : 1 \leq i \leq k\}$.*

Exercise: Verify (carefully!) that if ϕ does not contain a bicycle then ϕ is satisfiable. Note that the variables appearing among the w_i must be distinct!

Therefore we may bound the probability that ϕ is not satisfiable as follows:

$$\Pr[\varphi \text{ not satisfiable}] \leq \Pr[\varphi \text{ contains a bicycle}] \leq \sum_{k=2}^{n} n^k 2^k (2k)^2 m^{k+1} \left(\frac{1}{4\binom{n}{2}}\right)^{k+1}$$

Here we are summing over the possible sizes k of the bicycle; factor n^k is an upper bound on the number of ways to choose k distinct variables for the bicycle, 2^k accounts for the sign of each variable in the bicycle, $(2k)^2$ is the number of ways to choose u and v, m^{k+1} is an upper bound on the number of ways to choose which of the m clauses of the formula induce the $k+1$ edges of the bicycle, and $\left(\frac{1}{4\binom{n}{2}}\right)^{k+1}$ is the probability that the chosen clauses induce the corresponding edges of the bicycle. (Here we are working with the model which chooses valid clauses u.a.r.) Now, the right hand side of the previous inequality becomes

$$\sum_{k=2}^{n} n^k 2^k (2k)^2 m^{k+1} \left(\frac{1}{4\binom{n}{2}}\right)^{k+1} = \frac{2}{n} \sum_{k=2}^{n} k^2 \left(\frac{m}{n-1}\right)^{k+1} \to 0 \text{ as } n \to \infty$$

The last expression tends to zero because the factor outside the summation goes to 0 and the summation is bounded by a constant. (This latter fact follows because $\frac{m}{n-1} = \frac{(1-\epsilon)n}{n-1}$ is less than 1, so the terms of the sum form a decreasing geometric series weighted by a polynomial factor.)

This concludes the proof of the first part (lower bound) of Theorem 7.3. The upper bound, which uses the second moment method, will be proved in the next lecture. ■

References

[AM02] D. Achlioptas and C. Moore, "The asymptotic order of the random k-SAT threshold," *Proceedings of the 43rd IEEE FOCS*, 2002, pp. 779–788.

[AP04] D. Achlioptas and Y. Peres, "The threshold for random k-SAT is $2^k \log 2 - O(k)$," *Journal of the American Mathematical Society* **17** (2004), pp. 947–973.

[CF90] M.-T. Chao and J. Franco, "Probabilistic analysis of a generalization of the unit-clause literal selection heuristics for the k-satisfiability problem," *Information Science* **51** (1990), pp. 289–314.

[CR92] V. Chvátal and B. Reed, "Mick gets some (the odds are on his side)," *Proceedings of the 33rd IEEE FOCS*, 1992, pp. 620–627.

[CP16] A. Coja-Oghlan and K. Panagiotou, "The asymptotic k-SAT threshold," *Advances in Mathematics* **288** (2016), pp. 985–1068.

[DKMP09] J. Diaz, L. Kirousis, D. Mitsche and X. Perez-Gimenez, "On the satisfiability threshold of formulas with three literals per clause," *Theoretical Computer Science* **410** (2009), pp. 2920–2934.

[DSS14] J. Ding, A. Sly and N. Sun. "Proof of the satisfiability conjecture for large k," ArXiv preprint arXiv:1411.0650 [math.PR], 2014.

[FdlV92] W. Fernandez de la Vega, "On random 2-SAT," Manuscript, 1992.

[F99] E. Friedgut, "Necessary and sufficient conditions for sharp thresholds of graph properties," *Journal of the American Mathematical Society* **12** (1999), pp. 1017–1054.

[G96] A. Goerdt, "A threshold for unsatisfiability," *Journal of Computer & System Sciences* **53** (1996), pp. 469–486.

[HS03] M. Hajiaghayi and G.B. Sorkin, "The satisfiability threshold for random 3-SAT is at least 3.52," submitted, 2003.

[KKL03] A. Kaporis, L.M. Kirousis and E. Lalas, "Selecting complementary pairs of literals," *Proc. LICS'03 Workshop on Typical Case Complexity and Phase Transitions*, June 2003.