Lecture 30: Concentration of the degree distribution of PA graphs

In this lecture we prove the first part of Lemma 29.3 from last lecture, using the Doob martingale version of the Azuma-Hoeffding inequality given in Lemma 29.5. First we recall the notation and definitions from last time. Let $(G_n)_{n\geq 1}$ be a sequence of preferential attachment graphs, for $1 \leq i \leq n$, let $D_n(i)$ be the degree of vertex i in G_n, and set

$$P_n(k) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[D_n(i) = k],$$

and

$$N_n(k) := \sum_{i=1}^{n} \mathbb{I}[D_n(i) = k];$$

the proportion and number of vertices with degree k in G_n.

Proof of Part 1 of Lemma 29.3. We want to show that there is a constant $c > 0$ such that

$$\mathbb{P}\Big(\max_{k\geq 1}\big|P_n(k) - \mathbb{E}\big[P_n(k)\big]\big| \geq c\sqrt{\log(n)/n}\Big) \xrightarrow[n\to\infty]{} 0,$$

which is the same as

$$\mathbb{P}\Big(\max_{k\geq 1}\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\Big) \xrightarrow[n\to\infty]{} 0. \tag{30.1}$$

To bound the probability on the left hand side, as usual, we first use a union bound to find

$$\begin{aligned}
&\mathbb{P}\Big(\max_{k\geq 1}\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\Big) \\
&\qquad = \mathbb{P}\Big(\cup_{k\geq 1}\big\{\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\big\}\Big) \\
&\qquad \leq \sum_{k=1}^{n-1} \mathbb{P}\Big(\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\Big),
\end{aligned} \tag{30.2}$$

where the sum is finite since $N_n(k) = \mathbb{E}[N_n(k)]$ for $k \geq n$. We claim that for $i = 1, \ldots, n$, we have

$$\big|\mathbb{E}[N_n(k)|G_i] - \mathbb{E}[N_n(k)|G_{i-1}]\big| \leq 2. \tag{30.3}$$

Taking this for granted, we can then apply Lemma 29.5 with $X = N_n(k)$, the G_i as the preferential attachment graph (note that G_{i-1} is part of G_i, so is a function of G_i) and $a_i = 2$, noting that $\mathbb{E}[N_n(k)|G_n] = N_n(k)$, to find

$$\begin{aligned}
\mathbb{P}\Big(\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\Big) &\leq 2\exp\{-c^2(n\log(n))/(8n)\} \\
&= 2\exp\{-c^2\log(n)/8\} \\
&= 2n^{-c^2/8}.
\end{aligned}$$

Plugging this in to (30.2), we find

$$\mathbb{P}\Big(\max_{k\geq 1}\big|N_n(k) - \mathbb{E}\big[N_n(k)\big]\big| \geq c\sqrt{n\log(n)}\Big) \leq 2n^{1-c^2/8},$$

which tends to zero as $n \to \infty$ for any $c > \sqrt{8}$.

To complete the proof, we only need to show (30.3). Before presenting the argument, let's think about how to understand $\mathbb{E}[N_n(k)|G_j]$. We think of G_j as fixed and known, say $G_j = g$, and then run the preferential attachment process started from $G_j = g$ for $n - j$ further steps to

obtain G_n. No matter what happens, G_n restricted to vertices $\{1,\dots,j\}$; denoted $G_n|_j$; equals g. The randomness of G_n given $G_j = g$ stems only from the preferential attachment steps taken after vertex j is added. Thus when we look at G_n given $G_{i-1} = g$, we first take a step to randomly generate G_i, whereas G_n given $G_i = h$ has a fixed value for this first step. Our strategy is to see how much $N_n(k)$ can change if at step i, we start from two different graphs h and h' (the latter standing for the random step after $G_{i-1} = g$), which satisfy $h|_{i-1} = h'|_{i-1} = g$ (this is because $G_i|_{i-1} = G_{i-1}$ always). We can make this more concrete in the following calculations. For any tree g on vertices $\{1,\dots,i-1\}$ with $\mathbb{P}(G_{i-1} = g) > 0$ and tree h on vertices $\{1,\dots,i\}$ satisfying $h|_{i-1} = g$, we have

$$\mathbb{E}[N_n(k)|G_{i-1} = g] = \sum_{h'|_{i-1}=g} \mathbb{E}[N_n(k)|G_i = h']\mathbb{P}(G_i = h'|G_{i-1} = g),$$
$$\mathbb{E}[N_n(k)|G_i = h] = \sum_{h'|_{i-1}=g} \mathbb{E}[N_n(k)|G_i = h]\mathbb{P}(G_i = h'|G_{i-1} = g),$$

where the sum is over trees on vertices $\{1,\dots,i\}$ (plus the restriction), the first equation uses that $\{G_i = h', G_{i-1} = g\} = \{G_i = h'\}$ for $h'|_{i-1} = g$, and the second equation is because the probabilities sum to one. Taking the difference of these two equations and using the triangle inequality, we have

$$\begin{aligned}
&\big|\mathbb{E}[N_n(k)|G_i = h] - \mathbb{E}[N_n(k)|G_{i-1} = g]\big| \\
&\qquad \le \sum_{h'|_{i-1}=g} \big|\mathbb{E}[N_n(k)|G_i = h] - \mathbb{E}[N_n(k)|G_i = h']\big|\mathbb{P}(G_i = h'|G_{i-1} = g) \\
&\qquad \le \sup_{h'|_{i-1}=g} \big|\mathbb{E}[N_n(k)|G_i = h] - \mathbb{E}[N_n(k)|G_i = h']\big|,
\end{aligned}$$

and therefore

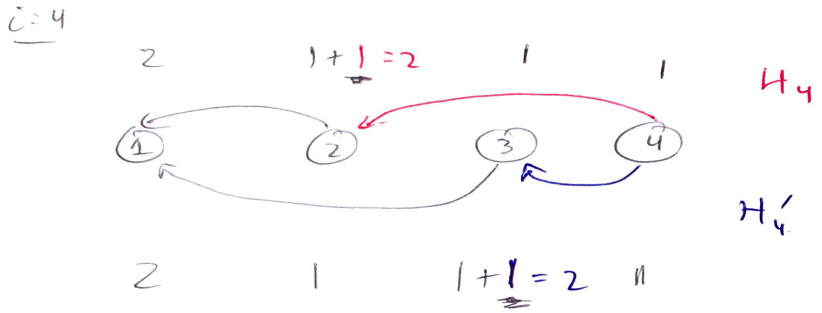
$$\big|\mathbb{E}[N_n(k)|G_i] - \mathbb{E}[N_n(k)|G_{i-1}]\big| \le \sup_{h'|_{i-1}=h|_{i-1}} \big|\mathbb{E}[N_n(k)|G_i = h] - \mathbb{E}[N_n(k)|G_i = h']\big|, \quad (30.4)$$

where the supremum is over trees h' and h with $\mathbb{P}(G_i = h') > 0$ and $\mathbb{P}(G_i = h) > 0$ (plus the restriction). To show (30.4) is no greater than two, we construct a coupling $(H_j, H'_j)_{j=i}^n$ of $\mathscr{L}\big((G_j)_{j=i}^n|G_i = h\big)$ and $\mathscr{L}\big((G_j)_{j=i}^n|G_i = h'\big)$ such that at most two vertices have degrees that don't agree between the two graphs, and hence that the number of vertices with degree k in the different processes can only differ by two.

We now define the coupling. Obviously $H_i = h$ and $H'_i = h'$. Assume now that $h \neq h'$. Then since $h|_{i-1} = h'|_{i-1}$, the graphs H_i and H'_i differ only in which of vertices $\{1,\dots,i-1\}$ the outgoing edge of vertex i is connected to. If we denote the degrees of vertex ℓ in H_j and H'_j by $E_j(\ell)$ and $E'_j(\ell)$, then there is $\ell_1 \neq \ell_2$ such that

$$E_i(\ell_1) = E'_i(\ell_1) + 1 \quad \text{and} \quad E'_i(\ell_2) = E_i(\ell_2) + 1,$$

and $E_i(\ell) = E'_i(\ell)$ for $\ell \notin \{\ell_1, \ell_2\}$. The following picture illustrates.



H_i and H'_i for $i = 4$. Black edges are in both H_4, H'_4, while red edges are in H_4 and blue edges in H'_4. The degrees of vertices only differ between the graphs at $\ell_1 = 2$ and $\ell_2 = 3$.

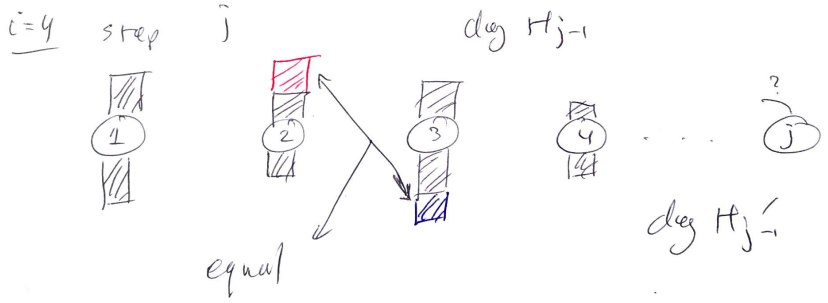
To generate (H_{i+1}, H'_{i+1}) given (H_i, H'_i), we add vertex $(i+1)$ and connect it to one of vertices $\{1, \dots, i\}$ in each graph, with probability proportional to degree. But the degrees only differ at vertices ℓ_1 and ℓ_2, and by the same amount at both positions, so we can couple as follows. For $j \geq (i+1)$, denote the vertex label that the outgoing edge of vertex j connects to in H_j and H'_j by I_j and I'_j.

$$\mathbb{P}\big(I_{i+1} = \ell, I'_{i+1} = \ell' | (H_i, H'_i)\big) = \frac{1}{2(i-1)} \times \begin{cases} E_i(\ell), & \ell = \ell',\ \ell \notin \{\ell_1, \ell_2\}, \\ E'_i(\ell_1), & \ell = \ell' = \ell_1, \\ E_i(\ell_2), & \ell = \ell' = \ell_2, \\ 1, & \ell = \ell_1, \ell' = \ell_2. \end{cases}$$

Now note in any of the events above, $E_{i+1}(\ell) = E'_{i+1}(\ell)$ for $\ell \notin \{\ell_1, \ell_2\}$, $E'_{i+1}(\ell_1) < E_{i+1}(\ell_1)$, $E_{i+1}(\ell_2) < E'_{i+1}(\ell_2)$, and $E_{i+1}(\ell_1) - E'_{i+1}(\ell_1) = E'_{i+1}(\ell_2) - E_{i+1}(\ell_2)$. Continuing inductively, we have the general update rule

$$\mathbb{P}\big(I_j = \ell, I'_j = \ell' | (H_{j-1}, H'_{j-1})\big) = \frac{1}{2(j-2)} \times \begin{cases} E_{j-1}(\ell), & \ell = \ell',\ \ell \notin \{\ell_1, \ell_2\}, \\ E'_{j-1}(\ell_1), & \ell = \ell' = \ell_1, \\ E_{j-1}(\ell_2), & \ell = \ell' = \ell_2, \\ E_{j-1}(\ell_1) - E'_{j-}(\ell_1), & \ell = \ell_1, \ell' = \ell_2. \end{cases}$$

The following picture illustrates the coupling.



Heights of bars indicate degrees of vertices in H_{j-1}, H'_{j-1}. Black bars on the top and bottom of a vertex are the same size and make up the total degree for vertices not equal to $2, 3$. The red bar indicates the additional edges received to vertex 2 in H_{j-1}, coming from the contribution of the extra "degree" in H_i and is of size $E_{j-1}(2) - E'_{j-1}(2)$. The blue bar indicates the analogous quantity for H'_i at vertex 3 and is of the same size, $E'_{j-1}(3) - E_{j-1}(3)$.

It is straightforward to see, by checking transition probabilities, that this coupling satisfies $\mathscr{L}(H_n) = \mathscr{L}(G_n | H_i = h)$ and $\mathscr{L}(H'_n) = \mathscr{L}(G_n | H_i = h')$. Moreover, for every $j = i, \dots, n$, $E_j(\ell) = E'_j(\ell)$ if and only if $\ell \notin \{\ell_1, \ell_2\}$, so

$$\left| \sum_{\ell=1}^{n} \mathbb{I}[E_n(\ell) = k] - \sum_{\ell=1}^{n} \mathbb{I}[E'_n(\ell) = k] \right| \leq \sum_{\ell=1}^{n} \mathbb{I}[E_n(\ell) \neq E'_n(\ell)] = 2. \tag{30.5}$$

Since

$$\mathscr{L}(N_n(k) | G_i = h) = \mathscr{L}\left(\sum_{\ell=1}^{n} \mathbb{I}[E_n(\ell) = k] \right),$$
$$\mathscr{L}(N_n(k) | G_i = h') = \mathscr{L}\left(\sum_{\ell=1}^{n} \mathbb{I}[E'_n(\ell) = k] \right),$$

the bounds (30.5) and (30.4) establish (30.3) □