

# Math 262 Lecture 10: 4/29/04

lecturer: Lincoln Lu  
scribe: Chris Calabro

May 23, 2004

## 1 Partial duplication model

The *partial duplication model* with parameters  $(p, t, t_0, G_{t_0})$  where  $p \in (0, 1)$ ,  $t, t_0 \in \mathbb{N}$ ,  $t \geq t_0$ , and  $G_{t_0}$  is an initial undirected graph is the probability distribution on undirected graphs  $G_t$  described by the following recursion:

If  $t = t_0$ , then  $G_t = G_{t_0}$ . Otherwise, select a graph  $G_{t-1} = (V_{t-1}, E_{t-1})$  from the partial duplication model with parameters  $(p, t-1, t_0, G_{t_0})$  and construct  $G_t$ , initially equal to  $G_{t-1}$ , as follows.

- Select a *model vertex*  $v$  from  $G_{t-1}$  uniformly at random.
- Add a *new vertex*  $u$  to  $G_t$ .
- For each neighbor  $w$  of  $v$  in  $G_{t-1}$ , add edge  $\{u, w\}$  to  $G_t$  with probability  $p$ .
- Add edge  $\{u, v\}$  to  $G_t$  with probability  $p$ .

The resulting graph is  $G_t = (V_t, E_t)$ . Of course we could view this as an iterative procedure as well.

Note that  $|V_t| = t - t_0 + |V_{t_0}|$ . For the remainder, let us choose  $t_0 = 1$  and  $G_{t_0}$  to be a graph with 1 vertex and no edges so that  $|V_t| = t$ . If we wanted to choose  $G_{t_0}$  to be the empty graph, we would have to change the description of the procedure slightly since the first step would have no model vertex.

Define  $\tau_t = |E_t|$ . Last time, we showed that for  $p < \frac{1}{2}$ ,

$$\forall k, \epsilon > 0 \Pr(\tau_t = \frac{pt}{1-2p} + O(t^{\frac{1}{2}+p+\epsilon})) \geq 1 - t^{-k},$$

although this looks suspicious since the left hand side does not seem to depend on  $k$  while the right hand side does.

It is left as an exercise to show that for  $p > \frac{1}{2}$ , w.h.p.  $\tau_t = O(t^{2p})$ . What about when  $p = \frac{1}{2}$ ?

## 2 Definition of a power law graph

The following discussion did not happen in class and is certainly not standard, but it seems necessary for what follows.

Before claiming that a graph is a “power law graph” we need to have a more rigorous definition of this term. It is sometimes said that a graph is a power law graph with exponent  $\beta > 1$  iff  $\exists c > 0 \forall k \geq 1$  the number of vertices of degree  $k$  is proportional to  $ck^{-\beta}$ . It is impossible for a finite graph to satisfy this definition literally, so we must introduce some notion of error.

We should introduce some notion of asymptotics so that we can say that the error shrinks as something else grows. Also we should introduce probability so that we can use this notion for random graphs.

It is not hard to show that for  $\beta \geq 4$  and  $n \geq 8$  every “power law graph” is disconnected if we force the number of vertices of degree  $k$  to be  $\lfloor ck^{-\beta}n \rfloor$  for all  $k \geq 1$ . So using the floor function does not introduce enough error. Also, for real applications, there are graphs that we want to call power law graphs for which the power law does not hold when  $k$  is too large or too small.

Let’s say we want the power law to hold for  $k \in [L, U]$ . We should choose  $L, U$  so that the power law holds for many of the vertices. Then, to motivate our choice of what we will call an acceptable amount of error, let’s choose as a goal proving that for exponent  $\beta \leq 1$ , there can not be a power law graph. Intuitively this should be because the  $p$ -series  $\sum_{k=1}^{\infty} k^{-p}$  diverges for  $p \leq 1$ .

With these goals in mind, let us say that a sequence of random graphs  $G_1, G_2, \dots$  satisfies the *power law* with exponent  $\beta > 1$  iff, letting  $f(k, t) =$  the number of vertices of degree  $k$  in  $G_t$ , there is a  $c > 0$  and functions  $L, U : \mathbb{N} \rightarrow \mathbb{N}, a : \mathbb{N} \rightarrow \mathbb{R}$  with  $1 \leq L(t) \leq U(t) \leq |G_t|, U \in \omega(1), \ln L, a \in o(\ln U)$  and

$$\lim_{t \rightarrow \infty} Pr(\exists k \in [L(t), U(t)] |f(k, t) - ck^{-\beta}|G_t| \geq a(t)) = 0. \quad (1)$$

Now let us see what goes wrong if we have  $\beta = 1$ . Note that

$$\ln n \leq \sum_{k=1}^n k^{-1} \leq 1 + \ln n, \quad (2)$$

which can be seen by integration. For  $t$  sufficiently large, the probability in (1) is  $\leq \frac{1}{2}$  and so with positive probability the total number of vertices in  $G_t$  is at least

$$\begin{aligned} \sum_{k=L(t)}^{U(t)} f(k, t) &\geq \sum_{k=L(t)}^{U(t)} (ck^{-1}|G_t| - a(t)) \\ &\geq c|G_t|(\ln U(t) - \ln L(t) - 1) - |G_t|a(t) \quad \text{by (2)} \\ &\in |G_t|\Omega(\ln U(t) - \ln L(t) - a(t)) \\ &\subseteq |G_t|\omega(1), \end{aligned}$$

which is absurd since  $G_t$  cannot have more vertices than  $|G_t|$ . A similar contradiction is reached if we consider  $\beta < 1$ .

A practical choice for  $L, U$ , suggested by Lincoln, would be  $L \in O(\ln t), U \in \Omega(t^{\frac{1}{\beta}})$  for the case that  $|G_t| = t$ .

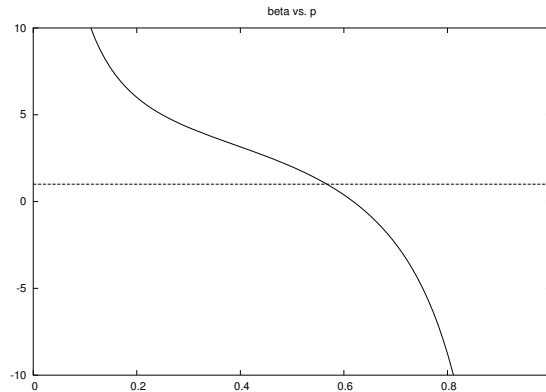
Another plausible definition would be to make demands on the *expected value* of  $f(k, t)$ . But then it is less clear what the global properties of a power law graph are.

### 3 Main theorem

Consider the equation

$$p\beta + p^{\beta-1} = 1 + p. \quad (3)$$

Elementary calculus shows that  $\exists p' \in (0, 1) \forall p \in (0, 1) - \{p'\}$ , (3) has exactly 2 solutions for  $\beta$ , and that (3) has exactly 1 solution for  $\beta$  at  $p = p'$ . One of the solutions is always  $\beta = 1$ . Define  $b(p)$  to be the other solution. At  $p = p'$ , the derivative of (3) with respect to  $\beta$  is 0 at  $\beta = 1$ , and so  $p'$  satisfies  $p' + \ln p' = 0$ , which implies  $p' \approx .567$ . The solutions to (3) are shown below.



In class we stated,

**Theorem 1.** *Almost surely, (i.e. with probability approaching 1 as  $t \rightarrow \infty$ ) the partial duplication model generates a power law graph with exponent  $\beta = b(p)$ .*

But this has problems. First, we need to restrict  $p$  to be  $< p'$  so that  $b(p) > 1$ . (The behavior of the graph when  $p \geq p'$  is discussed briefly in [1].) Second, the  $t$  is quantified too early. If we fix  $t$ , there is no notion of, “ $G_t$  is a power law graph.” To restore our asymptotics, we use the definition in the previous section and state

**Theorem 2.** *For  $0 < p < p'$ , where  $p' + \ln p' = 0$ , the partial duplication model generates a sequence of random graphs satisfying the power law with exponent  $\beta = b(p)$ .*

We will not prove this here but rather show that if

$$E(f(k, t)) = ck^{-\beta}t + h(t), \quad (4)$$

where  $h(t)$  satisfies

$$|h(t+1) - h(t)| \in o(1), \quad (5)$$

then  $\beta = b(p)$ .

(5) and the following lemma allow us to conclude that  $h \in o(t)$ . We will need (5), as the weaker hypothesis  $h \in o(t)$  will not suffice later.

**Lemma 3.** *Let  $d : \mathbb{N} \rightarrow \mathbb{R}, d \in o(1)$ , then  $\sum_{i=1}^t d(i) \in o(t)$ .*

*Proof.* Let  $m = \max_i |d(i)|$ . Let  $\epsilon > 0$  be given. Choose  $t_0$  so large that  $\forall i \geq t_0 |d(i)| \leq \frac{\epsilon}{2}$ . Then  $\forall t \geq \frac{2t_0m}{\epsilon}$

$$\begin{aligned} \left| \sum_{i=1}^t \frac{d(i)}{t} \right| &\leq \sum_{i=1}^{t_0-1} \frac{|d(i)|}{t} + \sum_{i=t_0}^t \frac{|d(i)|}{t} \\ &\leq \frac{t_0m}{t} + \frac{\epsilon}{2} \leq \epsilon. \end{aligned}$$

□

Also let us say that a function satisfying (5) is a  $o(1)$ -difference sequence.

The following lemma gives some plausibility to the hypothesis (4).

**Lemma 4.** *Suppose that  $G_1, G_2, \dots$  satisfies the power law with  $|G_t| = t$  and let*

$$Pr(\exists k \in [L(t), U(t)] |f(k, t) - ck^{-\beta}t| \geq a(t)) \leq \epsilon(t), \quad (6)$$

where  $\epsilon \in o(1)$ . Then  $|E(f(k, t)) - ck^{-\beta}t|$  is bounded above by an  $o(1)$ -difference sequence. Also we may weaken the hypothesis slightly so that we only assume  $a \in o(t)$ .

*Proof.* We claim that generality is not lost by assuming  $a(t), t\epsilon(t)$  are  $o(1)$ -difference sequences. To see this, first note that if  $\frac{a(t)}{t} \in o(1)$  and  $a$  is nonincreasing, then  $a(t)$  is an  $o(1)$ -difference sequence. We can easily replace  $\frac{a(t)}{t}, \epsilon(t)$  by monotone vanishing  $\frac{a'(t)}{t}, \epsilon'(t)$  that upper bound  $\frac{a(t)}{t}, \epsilon(t)$  and still have (1) hold. So assume  $a(t), t\epsilon(t)$  are  $o(1)$ -difference sequences.

Fix  $k, t$  and let  $I = [ck^{-\beta}t - a(t), ck^{-\beta}t + a(t)]$ . Then (6) implies

$$\begin{aligned} E(f(k, t)) &= \sum_{z \in I} zPr(f(k, t) = z) + \sum_{z \notin I} zPr(f(k, t) = z) \\ &\leq ck^{-\beta}t + a(t) + t\epsilon(t) \\ E(f(k, t)) &\geq (ck^{-\beta}t - a(t))(1 - \epsilon(t)) \geq ck^{-\beta}t - c\epsilon(t) - a(t), \end{aligned}$$

which proves the lemma. □

Note that  $|E(f(k, t)) - ck^{-\beta}t|$  being bounded above by an  $o(1)$ -difference sequence does not directly imply (4); and so, although (4) is plausible given that  $G_1, G_2, \dots$  satisfies the power law, we must continue to explicitly assume it. (It may be possible to use lemma 1 (“a useful lemma”) from the April 8, 2004 lecture notes to remove this assumption, but I don’t see how.)

Let us now develop a recurrence for  $E(f(k, t))$ . Let  $v$  be a vertex of degree  $k$  at time  $t$ . With probability  $\frac{(k+1)p}{t}$ ,  $v$  will have degree  $k+1$  at time  $t+1$  by choosing  $v$  or any of its  $k$  neighbors as the model vertex and then adding the appropriate edge. The same reasoning shows that a vertex  $v$  of degree  $k-1$  at time  $t$  will have degree  $k$  at time  $t+1$  with probability  $\frac{kp}{t}$ . Also the new vertex  $u$  will have degree  $k$  at time  $t+1$  with probability  $\sum_{j=k-1}^{t-1} \frac{f(j, t)}{t} \binom{j+1}{k} p^k (1-p)^{j+1-k}$ . Combining these observations gives the recurrence <sup>1</sup>

$$E(f(k, t+1)) = E(f(k, t)) \left(1 - \frac{(k+1)p}{t}\right) + E(f(k-1, t)) \frac{kp}{t} + \sum_{j=k}^t \frac{E(f(j-1, t))}{t} \binom{j}{k} p^k (1-p)^{j-k}. \quad (7)$$

Using (4) and simplifying gives

$$(1 + (k+1)p)k^{-\beta} = (k-1)^{-\beta}kp + \sum_{j=k}^t (j-1)^{-\beta} \binom{j}{k} p^k (1-p)^{j-k} + c^{-1} \left( -h(t+1) + h(t) - \frac{h(t)p}{t} + \sum_{j=k}^t \frac{h(t)}{t} \binom{j}{k} p^k (1-p)^{j-k} \right). \quad (8)$$

Now we will do something amazing. We can extend the definition of the binomial coefficient  $\binom{n}{k}$  to allow  $n$  to be an arbitrary complex number while  $k$  is a nonnegative integer as follows.

$$\binom{n}{k} = \frac{\prod_{i=0}^{k-1} (n-i)}{k!}$$

It is easy to see that

$$\binom{n}{k} = (-1)^k \binom{k-n-1}{k}. \quad (9)$$

**Theorem 5 (binomial).** <sup>2</sup> Let  $x, y, r \in \mathbb{C}$  with  $|\frac{x}{y}| < 1$ . Then

$$(x+y)^r = \sum_{k \geq 0} \binom{r}{k} x^k y^{r-k}.$$

*This also holds when  $r$  is a nonnegative integer and  $x, y$  are arbitrary complex numbers.*

<sup>1</sup>(7) disagrees with both the recurrence given in class and that found in [1], although it will not matter substantially after approximations are applied.

<sup>2</sup>can be found in [2]

The conclusion should be familiar, but the weakness of the hypothesis is astounding.

We continue with (8) by bounding the expression in the second line.

$$\begin{aligned}
\sum_{j=k}^t \binom{j}{k} (1-p)^{j-k} &\leq \sum_{j=k}^{\infty} \binom{j}{j-k} (1-p)^{j-k} \\
&= \sum_{j=0}^{\infty} \binom{j+k}{j} (1-p)^j \\
&= \sum_{j=0}^{\infty} \binom{-k-1}{j} (p-1)^j \quad \text{by (9)} \\
&= p^{-k-1} \quad \text{by thm 5,}
\end{aligned}$$

and so

$$\sum_{j=k}^t \binom{j}{k} p^k (1-p)^{j-k} \leq p^{-1},$$

which does not depend on  $k$ . Using this fact and (5), we see that the second line in (8) is  $o(1)$ , with a constant factor depending on  $p, c$  but not  $k$  and shrinking to 0 as  $t \rightarrow \infty$ .

The summation in the top line of (8) is

$$k^{-\beta} \sum_{j=k}^t \left(\frac{k}{j}\right)^{\beta} \left(1 + \frac{1}{j-1}\right)^{\beta} \binom{j}{j-k} p^k (1-p)^{j-k}. \quad (10)$$

To go further, we will employ some approximations. Recall that the gamma function is

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

The Stirling approximation gives

$$\Gamma(x) = \sqrt{\frac{2\pi}{x}} \left(\frac{x}{e}\right)^x \left(1 + \frac{1}{12x} + O\left(\frac{1}{x^2}\right)\right),$$

which implies that for fixed  $\alpha$

$$\frac{\Gamma(x-\alpha)}{\Gamma(x)} = \left(1 + O\left(\frac{1}{x}\right)\right) x^{-\alpha}$$

and for integer  $0 \leq l \leq x$

$$\left(1 - \frac{l}{x}\right)^{\alpha} = \left(1 + O\left(\frac{1}{x-l}\right)\right) \frac{\binom{x-\alpha}{l}}{\binom{x}{l}}. \quad (11)$$

The details are a bit messy, but not hard.

Using (11) with  $x = j, l = j - k, \alpha = \beta$  in (10) gives

$$\left(1 + O\left(\frac{1}{k}\right)\right) k^{-\beta} \sum_{j=k}^t \binom{j-\beta}{j-k} p^k (1-p)^{j-k}. \quad (12)$$

Now (9) and theorem 5 gives

$$\begin{aligned} \sum_{j=k}^{\infty} \binom{j-\beta}{j-k} (1-p)^{j-k} &= \sum_{j=0}^{\infty} \binom{j+k-\beta}{j} (1-p)^j \\ &= \sum_{j=0}^{\infty} \binom{\beta-k-1}{j} (p-1)^j = p^{\beta-k-1}. \end{aligned} \quad (13)$$

Combining (12) and (13) shows that the summation in the top line of (8) is

$$\left(1 + O\left(\frac{1}{k}\right)\right) k^{-\beta} p^{\beta-1},$$

and so we have

$$(1 + (k+1)p)k^{-\beta} = (k-1)^{-\beta} k p + \left(1 + O\left(\frac{1}{k}\right)\right) k^{-\beta} p^{\beta-1} + o(1),$$

which, after multiplying through by  $k^\beta$ , simplifies to

$$1 + p = \beta p + p^{\beta-1} + O\left(\frac{1}{k}\right) + o(1)k^\beta.$$

This needs to be interpreted carefully: for large  $k$  and much larger  $t$  (how large  $t$  needs to be grows as  $k$  grows), the equation (3) approximately holds and so  $E(f(k, t))$  is approximately  $ck^{b(p)}t + o(t)$ . For small  $k$  or for  $t$  too small, we conclude nothing. But since we assumed (4) for  $k \leq U(t)$  and  $U(t) \in \omega(1)$ , the only possible choice for  $\beta$  is  $b(p)$ .

## References

- [1] F. Chung, L. Lu, T. Dewey, D. Galas, *Duplication Models for Biological Networks*, Journal of Computational Biology, pp. 667-687, vol 10, num 5, 2003
- [2] R. Graham, D. Knuth, O. Patashnik, *Concrete Mathematics, A Foundation for Computer Science, 2nd Ed.*, Addison Wesley, 1994