GROWING PROTEAN GRAPHS

PAWEŁ PRAŁAT AND NICHOLAS WORMALD

ABSTRACT. The web may be viewed as a graph each of whose vertices corresponds to a static HTML web page, and each of whose edges corresponds to a hyperlink from one web page to another. Recently there has been considerable interest in using random graphs to model complex real-world networks to gain an insight into their properties. In this paper we propose an extended version of a new random model of the web graph in which the degree of a vertex depends on its age. We use the differential equation method to obtain basic results on the probability of edges being present. From this we are able to characterize the degree sequence of the model and study its behaviour near the connectivity threshold.

1. INTRODUCTION

Recently many new random graphs models have been introduced and analyzed by certain common features observed in many large-scale real-world networks such as the 'web graph' (see, for instance a general survey [2]). The web may be viewed as a directed graph whose nodes correspond to static pages on the web, and whose arcs correspond to links between these pages.

One of the most characteristic features of this graph is its degree sequence. Broder et al. [3] noticed that the distribution of degrees follows a power law: the fraction of vertices with degree d is proportional to $d^{-\gamma}$, where γ is a constant independent of the size of the network (more precisely, $\gamma \sim 2.1$ for in-degrees, $\gamma \sim 2.7$ for out-degrees). These observations suggest that the web is not well modeled by traditional random graph models such as $G_{n,p}$ (see, for instance [5]).

Łuczak and the first author introduced in [6] another random graph model of the undirected 'web graphs', the protean graph $\mathcal{P}_n(d, \eta)$, which is controlled by two additional parameters ($d \in \mathbb{N}$ and $0 < \eta < 1$). The major feature of this model is that older vertices are preferred when joining a new vertex into the graph. In [6] it is proved that the degrees of the $\mathcal{P}_n(d, \eta)$ are distributed according to the power law. The first author showed also in [8] that the protean graph $\mathcal{P}_n(d, \eta)$ asymptotically almost surely (a.a.s.) has one giant component, containing a positive fraction of all vertices, whose diameter is equal to $\Theta(\log n)$.

Note that, unlike most of theoretical models of the internet graph, the number of vertices of the protean graph is large but fixed and does not grow during the protean process. One may view this as a weakness of the approach since the

Date: May 15, 2006.

1991 *Mathematics Subject Classification.* Primary: 05C80. Secondary: 05C07, 05C40.

Key words and phrases. random graphs, web graphs, protean graphs, degree sequence, connectivity, differential equations.

The second author was supported by the Canada Research Chairs Program and NSERC.

internet graph is, at least at this moment, rapidly expanding. In the present paper, the authors introduce another random graph model, a growing protean graph $\mathcal{P}_t(p, d, \eta)$, which is an extended version of standard protean graph controlled by an additional parameter p, $0.5 < p \leq 1$. This extension causes the number of vertices of $\mathcal{P}_t(p, d, \eta)$ to grow during the process.

In Section 4, we use the differential equation method to obtain a result similar to one in [6] for probability that a set of edges is present or absent in the graph, and then we use this result to derive degree distribution and connectivity properties of the growing protean graph, similar to those in [6] (see Section 5).

2. Definitions

A protean process, defined below, is a sequence $\{G_t\}_{t=0}^{\infty} = \{(V_t, E_t)\}_{t=0}^{\infty}$ of undirected graphs, where t denotes time. Our model has three fixed parameters: $0.5 < p \leq 1$, $d \in N$ and $0 < \eta < 1$. Let $G_0 = (V_0, E_0) = (\{v_1\}, \emptyset)$ be a fixed initial graph with a single vertex without edges. Let N_t be a random variable denoting the number of vertices *minus 1* at time t, i.e. $N_t = |V_t| - 1$. For $t > 0$ we form G_t from G_{t-1} according to the following rules:

- With probability p, add a new vertex $v = v_{N_{t-1}+1}$ together with d edges from v to existing vertices chosen randomly with weighted probabilities. The edges are added in d substeps. In each substep, one edge is added, and the vertex to join to is chosen as v_i with probability equal to $i^{-\eta} / \sum_{j=1}^{N_{t-1}+1} j^{-\eta}$.
- Otherwise, which occurs with probability $1 - p$, if $N_{t-1} = 0$ (G_{t-1} has a single vertex only) do nothing, whilst if $N_{t-1} > 0$, choose a random vertex $v_i, i \in [N_{t-1} + 1] = \{1, 2, \ldots, N_{t-1} + 1\}$, delete v_i together with all edges incident to it. Finally, relabel the remaining vertices preserving their order. Thus v_{j+1} becomes v_j for $i \leq j \leq N_{t-1}$.

$\mathcal{P}_t(p, d, \eta)$ denotes the protean graph G_t.

Our model allows loops and multiple edges; there seems no reason to exclude them. However, there will not in general be very many of these, so excluding them can be shown not to significantly affect our conclusions.

There is also some flexibility in the starting graph. We could alternatively start with any arbitrary graph G_0, provided its vertices are assigned distinct "ages". Since all our results are asymptotic, it is easy to see that the same results will follow; the influence of the initial graph diminishes over time. In particular, our starting point in some proofs, such as the first result in the following section, is the point at which the graph has grown suitably large but is otherwise arbitrary.

Note that during the process, a vertex v_j "becomes" v_{j-1}. Since we want to track such changes for a particular vertex, we say that v_j has *label* j and regard the event of "becoming" v_{j-1} as a change of label only. So, when this occurs, v_j in G_{t-1} is the same vertex as v_{j-1} in G_t.

We say that an event holds *with extreme probability* (*wep*), if it holds with probability at least $1 - \exp(-\Theta(\log^2 t))$ as $t \to \infty$. More generally, an event holds *wep*$_s$ if it holds with probability at least $1 - \exp(-\Theta(\log^2 s))$ as $s \to \infty$. To combine this notion with other asymptotic notation such as $O()$ and $o()$, we follow the conventions in [10].

3. The growing protean process

We first show that the number of vertices of G_t is concentrated.

Lemma 3.1. *Let $p \in (0.5, 1]$, $d \in \mathbb{N}$, and $\eta \in (0, 1)$. Then* wep

$$N_t = N_0 + (2p-1)t + O\Big(\sqrt{t}\log t\Big)\,.$$

Proof. Let $\{Z_i\}$ be a sequence of t independent random variables each of which is equal to 1 with probability p and -1 with probability $1-p$. Then

$$N_t = N_0 + \sum_{i=1}^{t} Z_i + f\Big(N_0, \{Z_i\}\Big)\,,$$

where $f = f(N_0, \{Z_i\})$ is a deterministic function arising from the fact that a vertex is not deleted if N_i is about to drop to 0. Since f is nonnegative, the random variable N_t is stochastically bounded from below by $N_0 + \sum_{i=1}^{t} Z_i$. The lower tail of this variable has the sharp concentration claimed, by Chernoff's inequality (see, for instance Corollary 2.3 in [5]). Thus, for every $\varepsilon = \Theta\big(\log t/\sqrt{t}\big)$,

$$\begin{aligned}
&\mathbb{P}\Big(N_t < N_0 + (1-\varepsilon)(2p-1)t\Big)\\
&\qquad\le \mathbb{P}\Big(\sum_{i=1}^{t} Z_i < (1-\varepsilon)\mathbb{E}\sum_{i=1}^{t} Z_i\Big)\\
&\qquad\le 2\exp\Big(-\frac{\varepsilon^2}{3}\mathbb{E}\sum_{i=1}^{t} Z_i\Big) = \exp\Big(-\Theta(\log^2 t)\Big)\,.
\end{aligned}$$

For the upper tail, we note first (again using Chernoff) that *wep* the random variable $Z(k) = \sum_{i=1}^{k} Z_i$ is positive for every k in the range $t^{1/4} \le k \le t$. Hence *wep* $f < t^{1/4}$. The upper tail bound again follows from Chernoff's inequality. For every $\varepsilon = \Theta\big(\log t/\sqrt{t}\big)$

$$\begin{aligned}
&\mathbb{P}\Big(N_t > N_0 + (1+\varepsilon)(2p-1)t\Big)\\
&\qquad= \mathbb{P}\Big(\sum_{i=1}^{t} Z_i > (1+\varepsilon)\ \mathbb{E}\sum_{i=1}^{t} Z_i - f\Big)\\
&\qquad\le \mathbb{P}\Big(\sum_{i=1}^{t} Z_i > (1+\frac{\varepsilon}{2})\ \mathbb{E}\sum_{i=1}^{t} Z_i\Big)\\
&\qquad\le 2\exp\Big(-\frac{\varepsilon^2}{12}\mathbb{E}\sum_{i=1}^{t} Z_i\Big) = \exp\Big(-\Theta(\log^2 t)\Big)\,.
\end{aligned}$$

□

In the rest of this section, we will consider the growing protean process $\{G_t\}_{t=t_0}^{t_f}$ from a time t_0, conditional upon $G_{t_0} = G$ for some fixed graph G, and let n denote the number of vertices of G minus 1, that is, $n = N_{t_0} = |V_{t_0}| - 1$. For this section, we will consider the process only up to $t_f = t_0 + \lfloor cn^{2p}/\log^3 n\rfloor$, where $c > 0$ is

an arbitrary constant (included to make a nicer statement of Lemma 4.3). We desire effectively to assume that the vertex $v = v_{n+1} \in G_{t_0}$ survives until time t_f. Conditioning on this event, which we call $S(t_0, t_f)$, is equivalent to considering an altered process in which, for each step that deletes a vertex, the selection is made from the vertices other than v. Until further notice, we consider this altered process. We define the random variable J_t to be the number of vertices older than v in G_t. We need to show that J_t is sharply concentrated in the context of the conditional space under consideration. It is easy for completeness to treat N_t at the same time, even though Lemma 3.1 shows concentration of N_t in general.

Note that the vector (N_t, J_t) is Markovian, that is, its distribution at time $t+1$ is determined by its value at time t and is independent of the earlier history. Also it is easy to see that for every $t_0 \le t \le t_f$, provided $N_t > 0$

$$\begin{aligned} \mathbb{E}(N_{t+1} - N_t | G_t) &= 2p - 1 \,, \\ \mathbb{E}(J_{t+1} - J_t | G_t) &= -(1-p)\frac{J_t}{N_t} \,. \end{aligned}$$

It provides some insight if we define real functions $z(x)$ and $y(x)$ to model the behaviour of the scaled functions $\frac{1}{n}N_{xn}$ and $\frac{1}{n}J_{xn}$ respectively. If we presume that the changes in the functions correspond to the expected changes of random variables, we obtain a system of differential equations

$$\begin{aligned} z'(x) &= 2p - 1 \,, \\ y'(x) &= -(1-p)\tfrac{y}{z} \,, \end{aligned}$$

with the initial conditions $z(t_0/n) = y(t_0/n) = 1$. The general solution of this system can be put in the form

$$\begin{aligned} (2p-1)\log y &= -(1-p)\log z + C_1 \,, \\ z &= (2p-1)x + C_2 \,. \end{aligned} \tag{1}$$

Defining

$$H(N_t, J_t) = (2p-1)\log\frac{J_t}{n} + (1-p)\log\frac{N_t}{n} \,, \tag{2}$$

in view of (1), the general solution of the scaled differential equation corresponds to the system of equations

$$\begin{aligned} H(N_t, J_t) &= C_1 \,, \\ N_t - (2p-1)t &= C_2 \,. \end{aligned}$$

This is a solution (taking $N_t = N(t)$ etc.) of the unscaled differential equations

$$\begin{aligned} N'(t) &= 2p - 1 \\ J'(t) &= -(1-p)\tfrac{J}{N}, \end{aligned} \tag{3}$$

where t is regarded as a real variable. Of course C_1 and C_2 are determined by the initial conditions. It should be emphasised that these differential equations are only suggested (at this stage). However, we will be able to show that J_t is well concentrated around the solution value $y(t/n)n$. For this we use the same supermartingale method as in Pittel et al. [7]. It is encapsulated by the following result [9, Corollary 4.1].

Lemma 3.2. *Let $G_0, G_1, \dots, G_t$ be a random process and X_i a random variable determined by $G_0, G_1, \dots, G_i$, $0 \le i \le t$. Suppose that for some real b and constants c_i, $\mathbb{E}(X_i - X_{i-1} | G_0, G_1, \dots, G_{i-1}) < b$ and $|X_i - X_{i-1} - b| \le c_i$ for $1 \le i \le t$. Then for all $\alpha > 0$,*

$$\mathbb{P}\big(\exists i \ (0 \le i \le t) : X_i - X_0 \ge ib + \alpha\big) \le \exp\Big(-\frac{\alpha^2}{2\sum c_j^2}\Big) .$$

We now come to the main result of this section. Let $I_{t_0,i,t}$ denote the label of the vertex in G_t that was v_i in the graph G_{t_0}, provided that vertex is still present in G_t. Note we can express J_t in terms of this notation, that is $J_t = I_{t_0,N_{t_0}+1,t} - 1$. When i and t_0 are understood, we abbreviate this to I_t.

Theorem 3.3. *Let $p \in (0.5, 1]$, $d \in \mathbb{N}$, $\eta \in (0,1)$, and for arbitrary t_0 let G_t, N_t, J_t be defined as above, and let $i \le N_{t_0}$. Condition on the events that $G_{t_0} = G$ for some fixed graph G and that $S(t_0, t_f)$ holds, and put $n = N_{t_0} = |V(G)| - 1$. Then* wep_n*, for every t in the range $t_0 \le t \le t_f$ we have*

$$N_t = n + (2p-1)(t - t_0) + O(n^p \log n), \tag{4}$$

$$J_t = n\Big(\frac{N_t}{n}\Big)^{\frac{p-1}{2p-1}}(1 + O(\log^{-1/2} n)) , \tag{5}$$

and, conditional upon the vertex $v_i \in G_{t_0}$ surviving until time t_f,

$$I_t = \frac{iJ_t}{n}(1 + O(\log^{-1/2} n)) \text{ or } \frac{iJ_t}{n} < \log^3 n . \tag{6}$$

Proof. In the first main part of the proof we show (5), and with almost no effort we obtain (4) at the same time. Alternatively, one can at the outset obtain (4) as follows. Note that for every $t_0 \le t \le t_0 + n^p$ equation (4) holds (deterministically). We observe that Lemma 3.1 applies for the growing protean process starting from an arbitrary initial graph with $N_0 + 1$ vertices. This implies immediately that wep_{t-t_0}

$$N_t = n + (2p-1)(t - t_0) + O\big(\sqrt{t - t_0}\log(t - t_0)\big)$$

holds for every $t_0 \le t \le t_f$. So, wep_n for every $t_0 + n^p < t \le t_f$ (4) holds.

Let $\mathbf{w}_t = (N_t, J_t)$, and consider the sequence of random variables

$$\{X_t\}_{t=t_0}^{t_f} = \{H(\mathbf{w}_t)\}_{t=t_0}^{t_f} ,$$

where the function H is defined in (2), and the stopping time

$$T = \min\{t \ge t_0 : J_t < n^p/2 \vee N_t < n/2 \vee t = t_f\} .$$

(A stopping time is any random variable T with values in $\{0, 1, \dots\} \cup \{\infty\}$ such that it is determined whether $T = \hat{t}$ for any time $\hat{t}$ from knowledge of the process up to and including time $\hat{t}$.) Note that the second-order partial derivatives of H with respect to N_t and J_t are $O(1/N_t^2 + 1/J_t^2) = O(1/J_t^2) = O(n^{-2p})$, provided $T > t$. Therefore, with $i \wedge T$ denoting $\min\{i, T\}$, we have

$$\begin{aligned} &H(\mathbf{w}_{(t+1)\wedge T}) - H(\mathbf{w}_{t\wedge T}) \\ &\quad = (\mathbf{w}_{(t+1)\wedge T} - \mathbf{w}_{t\wedge T}) \cdot \ \mathrm{grad} \ \ H(\mathbf{w}_{t\wedge T}) + O(n^{-2p}) . \end{aligned} \tag{7}$$

Recall that $H(\mathbf{w})$ is constant along every trajectory $\mathbf{w}$ of the unscaled differential equations (3). So, taking the expectation of (7) conditional on $G_{t\wedge T}$, we obtain

$$\mathbb{E}(H(\mathbf{w}_{(t+1)\wedge T}) - H(\mathbf{w}_{t\wedge T})|G_{t\wedge T}) = O(n^{-2p}) .$$

Also from (7), noting that

$$\text{grad} \;\; H(\mathbf{w}_t) = (O(1/N_t), O(1/J_t)) ,$$

and from the fact that N_t and J_t change by at most 1 per step of the process, we also have

$$\begin{aligned} &|H(\mathbf{w}_{(t+1)\wedge T}) - H(\mathbf{w}_{t\wedge T})| \\ &\quad = O(1/N_{t\wedge T}) + O(1/J_{t\wedge T}) + O(n^{-2p}) = O(n^{-p}) . \end{aligned}$$

Now we can apply Lemma 3.2 to the sequence $\{H(\mathbf{w}_{t\wedge T})\}_{t=t_0}^{t_f}$, and symmetrically to $\{-H(\mathbf{w}_{t\wedge T})\}_{t=t_0}^{t_f}$, with $\alpha = \log^{-1/2} n$, $b = O(n^{-2p})$ and $c_j = O(n^{-p})$, to show that *wep*$_n$

$$|H(\mathbf{w}_{t\wedge T}) - H(\mathbf{w}_{t_0})| = O(\log^{-1/2} n).$$

As $H(\mathbf{w}_{t_0}) = 0$, this implies from the definition (2) of the function H, that *wep*$_n$ equation (5) holds for every $t_0 \le t \le T$. By the same type of argument, but much simpler, we immediately obtain (4) *wep*$_n$ for t up to T.

To complete the first part of the proof we need to show that *wep*$_n$, $T = t_f$. The events asserted by (4) and (5) hold with this probability up until time T, as shown above, and the conjunction of these events implies that $J_t > n^p/2$ and $N_t > n/2$ for n sufficiently large, $t_0 \le t \le T$. It follows that $T = t_f$ *wep*$_n$. Together with the conclusion above, this completes the proof of the claim on the distribution of N_t and J_t.

We now turn to the claim on the distribution of the random variable I_t. It is easy to observe that, conditional upon v_i surviving until time t_f, I_t follows the hypergeometric distribution with parameters $n-2$, $k-1$ and $J_t - 2$, that is, conditional upon it surviving,

$$\mathbb{P}(I_t = k \mid J_t) = \frac{\binom{i-1}{k-1}\binom{n-i-1}{J_t-k-1}}{\binom{n-2}{J_t-2}} ,$$

and thus

$$\mathbb{E}(I_t \mid J_t) = \frac{(i-1)(J_t-2)}{n-2} = \frac{iJ_t}{n}(1 + O(n^{p-1})) .$$

We can apply a well known bound for the tail of the hypergeometric distribution (see, for instance, Theorem 2.10 in [5]) to show that the random variable I_t is sharply concentrated around its mean. Indeed, working in the conditional space under consideration (that the vertex survives and that J_t is given) using the fact that $\mathbb{E}I_t \ge \log^3 n$, we obtain

$$\mathbb{P}\left(|I_t - \mathbb{E}I_t| > \frac{\mathbb{E}I_t}{\log^{1/2} n}\right) \le 2\exp\Big(-\frac{\mathbb{E}I_t}{3\log n}\Big) = \exp\big(-\Omega(\log^2 n)\big) ,$$

which is the assertion required for (6). □

It is straightforward to obtain results like those in Theorem 3.3 but with much smaller error bounds than $O(\log^{-1/2} n)$, at the expense of reducing the value of t_f. One could then apply the lemma to successive intervals of time, tracking the progress of vertices in the later intervals using (6). However to obtain the main corollaries in later sections, this is not required. We do however need to convert the theorem to a form that does not require conditioning on G_{t_0}, as follows.

Corollary 3.4. *Let $p \in (0.5, 1]$, $d \in \mathbb{N}$, $\eta \in (0, 1)$, for arbitrary t_0 define J_t and I_t as above, and define $t_F = t_0 + \lfloor ct_0^{2p}/\log^3 t_0 \rfloor$, where $c > 0$ is an arbitrary constant. Let $D(i, t_0, t)$ denote the event that either $i > N_{t_0} + 1$ or the vertex of label i in G_{t_0} is not still present in G_t. Then* wep_{t_0}*, for every t in the range $t_0 \le t \le t_F$ we have*

$$N_t = (2p-1)t + O(t_0^p \log t_0), \tag{8}$$

$$D(N_{t_0}+1, t_0, t) \text{ or } J_t = N_{t_0}\Big(\frac{N_t}{N_{t_0}}\Big)^{\frac{p-1}{2p-1}}(1 + O(\log^{-1/2} t_0)), \tag{9}$$

and for all $i > 0$

$$D(i, t_0, t) \text{ or } I_t = \frac{iJ_t}{N_{t_0}}(1 + O(\log^{-1/2} t_0)) \text{ or } \frac{iJ_t}{N_{t_0}} < \log^3 t_0\,. \tag{10}$$

Proof. Lemma 3.1 shows that $N_{t_0} = \Theta(t_0)$ *wep*$_{t_0}$. Then, conditioning on the event that $N_{t_0} = \Theta(t_0)$, (4) implies that (8) holds *wep*$_{t_0}$. It then holds *wep*$_{t_0}$ without the conditioning since $N_{t_0} = \Theta(t_0)$ *wep*$_{t_0}$. We obtain (9) and (10) similarly; replacing the conditioning in Theorem 3.3 by the disjunction with the event $D(i, t_0, t)$ merely weakens the result. □

4. Basic lemma for edge probabilities

In this section we introduce the main tool that allows us easily to compute the probability of some events in the protean graphs, Lemma 4.3. This shows a relationship between $\mathcal{P}_t(p, d, \eta)$ and the random graph $G(n, q)$ on the set of vertices $[n] = \{1, 2, \ldots, n\}$ (where $n = N_t + 1$), in which a pair of vertices i, j, $1 \le i < j \le n$, are adjacent with probability

$$q = q(i, j) = d(1-\eta)n^{(p-1)/p} j^{\eta + (1-2p)/p} i^{-\eta}$$

independently for each such pair. Of course the protean graph $\mathcal{P}_t(p, d, \eta)$ has a very rich dependence structure, so it only shares some properties with $G(n, q)$.

First we consider a generalization of a well known 'balls into bins' model, which will be useful to prove Lemma 4.3. Suppose that we sequentially put d balls into m bins by placing each ball into a bin independently and the probability that we choose a bin k, $1 \le k \le m$, is equal to ρ_k, where $\sum_{i=1}^m \rho_k = 1$. Let $S_1, S_2 \subseteq [m]$, $S_1 \cap S_2 = \emptyset$, $|S_1| \le d$, and let $p(S_1, S_2)$ denote the probability that every bin from the set S_1 has at least one ball, and bins from the set S_2 have no balls. In the following, we use the notation $[x]_k = x(x-1)\cdots(x-k+1)$.

The following fact was used in [6].

Fact 4.1. *Using the notation above, we have*

$$p(S_1, S_2) \geq \Big(1 - \sum_{j \in S_1 \cup S_2} \rho_j\Big)^{d-|S_1|} [d]_{|S_1|} \prod_{i \in S_1} \rho_i \ ,$$

and

$$p(S_1, S_2) \leq \Big(1 - \sum_{j \in S_2} \rho_j\Big)^{d-|S_1|} [d]_{|S_1|} \prod_{i \in S_1} \rho_i \ .$$

The proof is simple: in the first inequality $p(S_1, S_2)$ is estimated by the probability that each bin from S_1 contains precisely one ball; in the second, some configurations are counted more than once.

Note that we may consider the process as two separate processes. The first process adds and deletes vertices and decides what the vertex sets are for all graphs G_t. Let us call this the *vertex process*. The second process (*edge process*) then decides which pairs of vertices are adjacent by using the rules of the growing protean process at each time t, for the vertex added. We will consider the vertex process first, and when we have enough facts about it at our fingertips we will consider the edge process. Note that Corollary 3.4 only really describes the vertex process.

Before stating the main results of this section, we define $n = n(t)$ to be the deterministic function of t that approximates the number of vertices in $\mathcal{P}_t(p, d, \eta)$, that is,

$$n = n(t) = (2p-1)t \, . \tag{11}$$

Given n, p, d, and η, define

$$u(j) = j^{(2p-1)/p} n^{(1-p)/p}$$

and

$$w(i,j) = (1-\eta)(j/i)^{\eta}/u(j) = (1-\eta) n^{(p-1)/p} j^{\eta+(1-2p)/p} i^{-\eta}.$$

Note that

$$w(i,j) = \big(1 + O\big(u(j)^{\eta-1}\big)\big) \frac{(iu(j)/j)^{-\eta}}{\sum_{s=1}^{u(j)} s^{-\eta}} \ .$$

First we need to "invert" Corollary 3.4 to obtain a statement which gives information about the vertex process for many times t_0 earlier than t.

Lemma 4.2. *Let $j_0 = \sqrt{t} \log^{3/(4p-2)} t$. Then* wep *for every i and j with $2\log^3 t < i < j \leq N_t + 1$ and $j > j_0$, the vertex with label j at time t was added at time*

$$\hat{t} = \frac{j^{(2p-1)/p} N_t{}^{(1-p)/p}}{2p-1} \big(1 + O(\log^{-1/2} t)\big).$$

Furthermore, if we let $\hat{i}$ denote the label in $G_{\hat{t}}$ of the vertex of label i in G_t, then wep

$$\hat{i} = \frac{iN_{\hat{t}}}{j} \big(1 + O(\log^{-1/2} t)\big) = \frac{iu(j)}{j} \big(1 + O(\log^{-1/2} t)\big) \ .$$

Proof. Put $t_I = c' t^{1/(2p)} \log^{3/(2p)} t$ for some $c' > 0$. Then for $t_I \le t_0 \le t$, wep_{t_0} is equivalent to *wep*. Also, for any such t_0, we may apply Corollary 3.4 since, for appropriate c in that corollary, $t < t_F$. Since a polynomial number of statements holding individually *wep* also hold jointly *wep*, we deduce that *wep*, (8–10) hold simultaneously for all t_0 in the range $t_I \le t_0 \le t$. By the same argument and Lemma 3.1, we have that *wep*

$$N_{t_0} = (2p-1)t_0 + O\Big(\sqrt{t_0}\log t_0\Big) \tag{12}$$

for all t_0 in the same range.

For some fixed $C > 0$, define $t_1 = \lfloor t_1^* \rfloor$ where

$$t_1^* = \frac{j^{(2p-1)/p} N_t^{(1-p)/p}\Big(1 - C\log^{-1/2} t\Big)}{2p-1}\,.$$

In view of the conclusions above, the following statements hold *wep*. From (12), for every t_3, $t_I \le t_3 \le t_1$

$$\begin{aligned}
N_{t_3} &= (2p-1)t_3\big(1 + O(t_3^{-1/2}\log t_3)\big)\\
&\le (2p-1)t_1^*\big(1 + O(t_3^{-1/2}\log t_3)\big)\\
&= j^{(2p-1)/p} N_t^{(1-p)/p}\Big(1 - C\log^{-1/2} t\Big)\big(1 + O(t_I^{-1/2}\log t_I)\big)\\
&\le j^{(2p-1)/p} N_t^{(1-p)/p}\Big(1 - c''C\log^{-1/2} t_I\Big)\\
&\le j^{(2p-1)/p} N_t^{(1-p)/p}\Big(1 - c''C\log^{-1/2} t_3\Big)\,,
\end{aligned}$$

where $c'' > 0$ is a constant, since $t = \Theta(t_I^{2p}/\log^3 t_I)$. Now from the statement above using (9), for sufficiently large C, all vertices added at any time in the interval $[t_I, t_1]$ have label strictly less than j if they survive until time t. This statement holds *wep*.

Note that for small enough c', t_I can be made an arbitrarily small fraction of t_1. So for any time $\hat{t}$, $t^{1/2} < \hat{t} < t_I$, we may apply the same argument but reducing the value of t to a smaller value, t', and if convenient reducing j to a smaller value, j', to deduce that *wep* the vertex added at time $\hat{t}$ has label strictly less than j at time t', if it survives until then. Since the label of a vertex cannot increase as the process continues, this is also true at time t. We may thus extend the interval to encompass all vertices added in the time interval $[t^{1/2}, t_1]$. Of course, vertices added before this interval have label less than $t^{1/2} < j_0 \le j$. So for sufficiently large C, *wep* all vertices added at any time before t_1 have label strictly less than j if they survive until time t.

Similarly if we define $t_2 = \lceil t_2^* \rceil$ where

$$t_2^* = \frac{j^{(2p-1)/p} N_t^{(1-p)/p}\Big(1 + C\log^{-1/2} t\Big)}{2p-1}\,.$$

Then *wep* all vertices added at any time in the interval $[t_2, t]$ have label strictly greater than j at time t, if they survive until then.

We deduce from these conclusions that *wep* for all $j \in [j_0, N_t]$, the vertex with label j at time t was added at some time between $t_1 = t_1(j)$ and $t_2 = t_2(j)$. In view of Corollary 3.4, we may approximate N_t by n, and this gives the first statement in the lemma.

Again from the above observations, the number of vertices at such a time $\hat{t}$, $t_1 \le \hat{t} \le t_2$, is *wep* equal to (recalling the definitions (11) etc.)

$$\begin{aligned} N_{\hat{t}} &= \hat{t}(2p-1)\big(1 + O(\hat{t}^{-1/2}\log\hat{t})\big) \\ &= j^{(2p-1)/p} n^{(1-p)/p}\big(1 + O(\log^{-1/2} t)\big) \\ &= u(j)\big(1 + O(\log^{-1/2} t)\big). \end{aligned}$$

Taking $t_0 \in [t_1, t_2]$, as mentioned above, we may assume that (10) holds *wep*. Hence, using a sandwiching argument as above (but for positions rather than times), if we let $\hat{i}$ denote the label in G_{t_0} of the vertex of label i in G_t, then *wep*

$$\hat{i} = \frac{iN_{t_0}}{j}\big(1 + O(\log^{-1/2} t)\big) = \frac{iu(j)}{j}\big(1 + O(\log^{-1/2} t)\big)\,.$$

(Note in particular the condition $2\log^3 t \le i$ ensures that the condition $\frac{\hat{i}J_t}{N_{t_0}} < \log^3 t_0$ *wep* does not hold.) This gives the second assertion of the lemma. □

We will use the following lemma to estimate the probability that pairs of vertices are adjacent in G_t, and others are not.

Lemma 4.3. *Let* $0.5 < p \le 1$, $d \in \mathbb{N}$, $0 < \eta < 1$,

$$E_1, E_2 \subseteq \{\{v_i, v_j\} : 2\log^3 t < i < j \le N_t + 1 \text{ and } j \ge j_0\}, \quad E_1 \cap E_2 = \emptyset\,.$$

For every $i, j \in [N_t + 1]$, $r = 1, 2$, *let*

$$V_r(j) = \{i : i < j \text{ and } \{v_i, v_j\} \in E_r\}\,,$$

$$w_r(j) = \sum_{i \in V_r(j)} w(i, j)\,,$$

and assume that $|V_1(j)| \le d$ *for every* $j \in [n]$.

Let $P_t(E_1, E_2, p, d, \eta)$ *denote the probability that all pairs from* E_1 *are edges of* $\mathcal{P}_t(p, d, \eta)$, *and no pair from* E_2 *is an edge of* $\mathcal{P}_t(p, d, \eta)$. *Then*

$$\begin{aligned} P_t(E_1, E_2, p, d, \eta) \le\ & o(\exp(-\log^{3/2} t)) \\ & + \prod_{j=1}^{n} [1 - (1 + O(\log^{-1/2} t))w_2(j)]^{d - |V_1(j)|} \\ & \quad \times [d]_{|V_1(j)|} \prod_{i \in V_1(j)} (1 + O(\log^{-1/2} t))w(i, j), \end{aligned}$$

and

$$P_t(E_1, E_2, p, d, \eta) \geq o(\exp(-\log^{3/2} t)) + \prod_{j=1}^{n} [1 - (1 + O(\log^{-1/2} t))(w_1(j) + w_2(j))]^{d-|V_1(j)|} \times [d]_{|V_1(j)|} \prod_{i \in V_1(j)} (1 + O(\log^{-1/2} t)) w(i, j).$$

Proof. For this we need to consider the edge process as defined at the start of this proof. To obtain conclusions *wep*, we may condition on any of the times t_0 that vertex of label j in G_t was added in the vertex process, with $t_1(j) \leq t_0 \leq t_2(j)$. The assertion then follows from Fact 4.1, Lemma 4.2 and the definition of the edge process. □

From the above lemma it follows that the behaviour of the protean graph $\mathcal{P}_t(p, d, \eta)$ is related to that of random graph with vertex set $[n]$ in which two vertices i, j, $2\log^3 t \leq i < j \leq n$, $j_0 \leq j$, are adjacent with probability

$$p(i, j) = dw(i, j) = d(1 - \eta) n^{(p-1)/p} j^{\eta + (1-2p)/p} i^{-\eta},$$

independently for each such pair.

Indeed, if $|V_1(j)| = o(d)$ for every $j \in [n]$, then Lemma 4.3 gives

$$\begin{aligned} P_t(E_1, E_2, p, d, \eta) &\sim \prod_{j=1}^{n} \Big(1 - \sum_{i \in V_2(j)} w(i, j)\Big)^d d^{|V_1(j)|} \prod_{i \in V_1(j)} w(i, j) \\ &= (1 + o(1)) \exp\Big(-\sum_{\{i,j\} \in E_2} p(i, j)\Big) \prod_{\{i,j\} \in E_1} p(i, j), \end{aligned}$$

whereas if we consider a graph with independent edges, the probability that an analogous event holds is equal to

$$\prod_{\{i,j\} \in E_2} \Big(1 - p(i, j)\Big) \prod_{\{i,j\} \in E_1} p(i, j) = (1 + o(1)) \exp\Big(-\sum_{\{i,j\} \in E_2} p(i, j)\Big) \prod_{\{i,j\} \in E_1} p(i, j).$$

Finally, note that the situation with these results is similar to that in [6]: since we claim nothing about edges between 'small' vertices i, $1 \leq i < 2\log^3 t$, it is difficult to obtain a general theorem which relates properties of our model to the model with independent edges (as is done, for instance, for a different model by Chung and Lu [4]). For the same reason we cannot use the general theory of Bollobás, Janson and Riordan [1] of inhomogeneous sparse random graphs. Nonetheless, as in [6], our Lemma 4.3 is strong enough to show that many properties of the independent model which, roughly speaking, do not depend on the behaviour of the first $2\log^3 t$ vertices, hold also for the growing protean graph. We discuss some examples in the next section.

5. Degrees of vertices, and connectivity

In this section we study the shape of the degree sequence of $\mathcal{P}_t(p,d,\eta)$, and its connectivity. The proofs are virtually the same as for the corresponding results in [6], but with our new Lemma 4.3 in place of Lemma 3.5 of that paper. We begin with the expected degree of vertex v_i. Recall that $j_0 = j_0(t) = \sqrt{t}\log^{3/(4p-2)} t$ and $n = n(t) = (2p-1)t$.

Theorem 5.1. *Let $0.5 < p \le 1$, $d = o\big(t^{(1-\eta)/3}\big)$ and $0 < \eta < 1$. Then the expected degree of a vertex $v_{i=i(t)}$ is given by*

$$\mathbb{E}d(v_i) \sim d\frac{1-\eta}{(1-p)/p+\eta}\Big(\Big(\frac{n}{i}\Big)^{\eta} + \frac{(1-2p)/p+2\eta}{1-\eta}\Big(\frac{i}{n}\Big)^{(1-p)/p}\Big)$$

for $j_0 < i \le N_t + 1$ and

$$\mathbb{E}d(v_i) \sim d\frac{1-\eta}{(1-p)/p+\eta}\Big(\frac{n}{i}\Big)^{\eta}$$

for $2\log^3 t < i \le j_0$. Moreover, the expected number of edges in the protean graph $\mathcal{P}_t(p,d,\eta)$ is equal to $(1+o(1))pdn$.

Note that for small $i = o(n)$, the expected degree of the vertex v_i is dominated by the factor $d\,\frac{1-\eta}{(1-p)/p+\eta}\,\big(\frac{n}{i}\big)^{\eta}$. Consequently, the degrees of the protean graph $\mathcal{P}_t(p,d,\eta)$, are distributed according to a power law. More specifically, let $Z_k = Z_k(n;p;d;\eta)$ denote the number of vertices of degree k in $\mathcal{P}_t(p,d,\eta)$ and $Z_{\ge k} = \sum_{\ell\ge k} Z_\ell$. Here and below *a.a.s.* means 'with probability tending to 1 as $n\to\infty$'.

Theorem 5.2. *Let $0.5 < p \le 1$, $d \in \mathbb{N}$, $0 < \eta < 1$, $k = k(n) \ge \log^2 n$, and $d = o(k)$. Then a.a.s.*

$$Z_{\ge k} = (1+o(1))n\Big(\frac{1-\eta}{(1-p)/p+\eta}\cdot\frac{d}{k}\Big)^{1/\eta} + O(\log^3 n)\,.$$

As with the non-growing protean graph in [6], we may attune the parameters of this model to obtain roughly the same degree distribution as the (undirected) web graph.

We next consider connectivity of $\mathcal{P}_t(p,d,\eta)$. Let $\rho_t(p,d,\eta)$ denote the probability that $\mathcal{P}_t(p,d,\eta)$ is connected.

Theorem 5.3. *Let $0.5 < p \le 1$, $0 < \eta < 1$ and $d = d(n) = a\log n$, where a is a positive constant. Then*

$$\lim_{t\to\infty}\rho_t(p,d,\eta) = \begin{cases} 1 & \text{if} \quad a > 1/g(x_0) \\ 0 & \text{if} \quad a < 1/g(x_0)\,, \end{cases}$$

where

$$g(x) = \frac{1-\eta}{(1-p)/p+\eta}(x^{-\eta} - x^{(1-p)/p}) - \log(1-x^{(1-p)/p})\,,$$

and $x_0 = x_0(p,\eta)$ is a value of x which minimizes function $g(x)$ in the interval $(0,1)$.

We observe that, as for the model in [6], *a.a.s.* near the threshold all isolated vertices have labels $(1+o(1))x_0(p,\eta)\,n$. The probability of being isolated is greatest for the vertices of medium labels since they have lost their 'old' neighbours which have already been deleted, yet they are not old enough to attract the 'new' ones.

References

[1] B. Bollobás, S. Janson and O. Riordan, *The phase transition in inhomogeneous random graphs*, Tech. Report 2005:18, Uppsala.

[2] A. Bonato, *A survey of web graph models*, Proceedings of Combinatorial and Algorithm Aspects of Networking, 2004.

[3] A. Broder, R. Kumar, F. Maghoul, P. Rahaghavan, S. Rajagopalan, R. State, A. Tomkins and J. Wiener, *Graph structure in the web*, Proc. 9th International World-Wide Web Conference (WWW), 2000, pp. 309–320.

[4] F. Chung and L. Lu, *Coupling Online and Offline Analyses for Random Power Law Graphs*, Internet Mathematics **1** (2004), 409–461.

[5] S. Janson, T. Łuczak and A. Ruciński, "Random Graphs", Wiley, New York, 2000.

[6] T. Łuczak and P. Prałat, *Protean graphs*, Internet mathematics, accepted, 20pp.

[7] B. Pittel, J. Spencer and N. Wormald, Sudden emergence of a giant k-core in a random graph, *J. Combinatorial Theory, Series B* **67** (1996), 111–151.

[8] P. Prałat, *Protean graphs - giant component and its diameter*, Discussiones Mathematicae Graph Theory, submitted, 13pp.

[9] N. Wormald, *The differential equation method for random graph processes and greedy algorithms* in Lectures on Approximation and Randomized Algorithms, eds. M. Karoński and H. J. Prömel, PWN, Warsaw, pp. 73-155, 1999.

[10] N.C. Wormald, Random graphs and asymptotics. Section 8.2 in *Handbook of Graph Theory*, J.L. Gross and J. Yellen (eds), pp. 817–836. CRC, Boca Raton, 2004.

Department of Combinatorics and Optimization, University of Waterloo, Waterloo ON, Canada N2L 3G1 and Adam Mickiewicz University, Wieniawskiego 1, 61-712 Poznań, Poland

E-mail address: ppralat@math.uwaterloo.ca

Department of Combinatorics and Optimization, University of Waterloo, Waterloo ON, Canada N2L 3G1

E-mail address: nwormald@math.uwaterloo.ca