CONNECTIVITY THRESHOLD AND RECOVERY TIME IN RANK-BASED MODELS FOR COMPLEX NETWORKS

PAWEŁ PRAŁAT

Abstract. The World Wide 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 study a generalized version of the protean graph (a random model of the 'web graph') in which the degree of a vertex depends on its age as well as its rank. Classic protean graphs can be seen as a special case of the rank-based approach where vertices are ranked according to age. Here, we investigate graph generation models based on other ranking schemes that also lead to graphs with a power law degree distribution. The main aim of this paper is to study the behaviour of the protean process 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, the book [1]). 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. [2] noticed that the distribution of degrees follows a power law: the fraction of vertices with degree k is proportional to $k^{-\gamma}$, where γ is a constant independent of the size of the network (more precisely, $\gamma \approx 2.1$ for in-degrees, $\gamma \approx 2.7$ 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 author of this paper introduced in [7] another random graph model of the undirected 'web graph': the protean graph

---

The author is supported by MITACS and NSERC. This work is part of the MITACS project Modelling and Mining of Networked Information Spaces (MoMiNIS)..

$\mathcal{P}_n(d,\eta)$, 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 [7] it is proved that the degrees of the $\mathcal{P}_n(d,\eta)$ are distributed according to the power law and the behaviour near the connectivity threshold is studied. The author of this paper showed also in [9] 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 $\Theta(\log n)$. (See also [11] where the growing protean graphs are studied.)

Classic protean graphs can be viewed as a special case of the rank-based approach where vertices are ranked according to age. The general approach was first proposed by Fortunato, Flammini and Menczer in [3], and the occurrence of a power law was postulated based on simulations (Janssen and the author of this paper provided rigorous proofs in [6]). In this approach, the vertices are ranked from 1 to n according to some ranking scheme (so the vertex with highest degree has rank 1, etc.), and the link probability of a given vertex is proportional to its rank, raised to the power $-\eta$ for some $\eta \in (0,1)$; we will refer to η as the *attachment strength*. (Negative powers are chosen since a low value for rank should result in a higher link probability.)

It has been shown that protean graphs with rank-based attachment lead to power law graphs (with the exponent $1+1/\eta$) for a variety of different ranking schemes [7], [9], [10]. These results suggest an explanation for the power law degree distribution often observed in real-life networks such as the web graph, protein interaction networks, and social networks. The growth of such networks can be seen as governed by a rank-based attachment scheme, based on a ranking scheme that can be derived from a number of different factors such as age, degree, or fitness. The exponent of the power law is independent of these factors, but is rather a consequence of the attachment strength. In addition, rank-based attachment accentuates the difference between higher ranked vertices: the difference in link probability between the vertices ranked 1 and 2 is much larger than that between the vertices ranked 100 and 101. This again corresponds to our intuition of what constitutes a credible mechanism for link attachment.

In this paper, we study a ranking scheme where an external prestige label for each vertex is given and vertices are ranked according to their prestige label. Another approach is to assign an initial rank to each vertex according to a given distribution. If the distribution is uniform, then the situation is very similar to the one described previously, and vertices with initial rank R exhibit behaviour as if they had received fitness R/n. We investigate how the threshold of connectivity is affected by the dependence structure of the protean graph, and characterize the limit distribution of the 'recovery time' for connectivity near the connectivity threshold.

Finally, let us mention that protean graphs are interesting not only as models of the web graphs, but they are also attractive from a theoretical point of view: they have a very rich dependence structure, and, unlike many other models of random graphs, $\mathcal{P}_n(d, \eta)$ can be viewed as the stationary distribution of the protean process.

2. Definitions

In this section, we formally define the graph generation model based on rank-based attachment. The model produces a sequence $\{G_t\}_{t=0}^{\infty} = \{(V_t, E_t)\}_{t=0}^{\infty}$ of undirected graphs on n vertices, where t denotes time. Our model has two fixed parameters: initial degree $d \in N$, and attachment strength $\eta \in (0, 1)$. At each time t, each vertex $v \in V_t$ has rank $r(v, t) \in [n]$ (we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$). In order to obtain a proper ranking, the rank function $r(\cdot, t) : V_t \to [n]$ is a bijection for all t, so every vertex has a unique rank. In agreement with the common use of the word 'rank', high rank refers to a vertex v for which $r(v, t)$ is small: the highest ranked vertex is ranked number one, so has rank equal to 1; the lowest ranked vertex has rank n. The initialization and update of the ranking is done according to a *ranking scheme*. Various ranking schemes can be considered; we first give the general model, and then list the ranking schemes.

Let $G_0 = (V_0, E_0)$ be any graph on n vertices and $r_0 = r(\cdot, 0) : V_0 \to [n]$ any initial rank function. (For random labeling scheme we take any function $l : V_0 \to (0, 1)$ and the initial rank function is a function of l; for degree scheme $r_0 = r_0(G_0)$.) For $t \geq 1$ we form G_t from G_{t-1} according to the following rules:

- Choose uniformly at random a vertex $u \in V_{t-1}$, then delete u together with all edges incident to it.
- Add a new vertex v_t together with d edges from v_t to existing vertices chosen randomly with weighted probabilities. The edges are added in d substeps. In each substep, one edge is added, and the probability that v is chosen as its endpoint (the link probability), equals
$$\frac{r(v, t-1)^{-\eta}}{\sum_{i=1}^{n} i^{-\eta}} = \frac{1-\eta}{n^{1-\eta} + O(1)} r(v, t-1)^{-\eta}.$$
- Update the ranking function $r(\cdot, t) : V_t \to [n]$ according to the ranking scheme.

Our model allows for 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 affect our conclusions in any significant way.

We now define the different ranking schemes.

- **Ranking by age**: The vertex added at time t obtains an initial rank n; its rank decreases by one each time a vertex with smaller rank is removed.
- **Ranking by inverse age**: The vertex added at time t obtains an initial rank 1; its rank increases by one each time a vertex with higher rank is removed.
- **Ranking by random labeling**: The vertex added at time t obtains a label $l(v_t) \in (0,1)$ chosen uniformly at random. Vertices are ranked according to their labels: if $l(v_i) < l(v_j)$, then $r(v_i,t) < r(v_j,t)$. Ties are broken by age.
- **Random ranking**: The vertex added at time t obtains an initial rank R_t which is randomly chosen from $[n]$ according to a prescribed distribution. Formally, let $F : [0,1] \to [0,1]$ be any cumulative distribution function. Then for all $k \in [t]$,

  $$\mathbb{P}(R_t \le k) = F(k/t).$$

- **Ranking by degree**: After each time step t, vertices are ranked according to their degrees in G_t, and ties are broken by age. Precisely, if $\deg(v_i,t) < \deg(v_j,t)$ then $r(v_i,t) < r(v_j,t)$, and if $\deg(v_i,t) = \deg(v_j,t)$ then $r(v_i,t) < r(v_j,t)$ if $i < j$.

The behaviour and state of a vertex clearly depends on its rank but also on its age relative to the ages of the other vertices. We use $a(\cdot,t)$ to denote the rank of the age of a vertex and $r(\cdot,t)$ for the ranking used in a given scheme.

We will use the stronger notion of *wep* in favour of the more commonly used a.a.s., since it simplifies some of our proofs. We say that an event holds *with extreme probability* (*wep*), if it holds with probability at least $1 - \exp(-\Theta(\log^2 n))$ as $n \to \infty$. Thus, if we consider a polynomial number of events that each holds *wep*, then *wep* all events hold. To combine this notion with asymptotic notations such as $O()$ and $o()$, we follow the conventions in [12].

In this paper, we focus on ranking by random labeling and random ranking with uniform distribution, that is, $F(x) = x$. For the random ranking scheme, it is enough to wait L steps for all vertices to be 'renewed' (for the random labeling case we have to wait twice as long: the first round is needed to have labels distributed uniformly at random;

during the second round the process 'forgets' about the initial graph). After all vertices have been renewed, the protean process is the Markov chain that is in the stationary distribution (that is, the distribution determined by G_t on the set of all ordered graphs on n vertices is identical for all t). By the coupon collector problem, a.a.s. $L = n(\log n + \omega(n))$ where $\omega(n)$ is any function tending to infinity with n (for random labeling scheme, clearly $L = 2n(\log n + \omega(n))$ a.a.s.). Furthermore, this distribution does not depend on the choice of G_0 and r_0. The random graph G_L corresponding to this distribution is called a protean graph $\mathcal{P}_n(d, \eta)$.

If $n \cdot l(v_i) > \log^3 n$ in the random labeling scheme, then the Chernoff's inequality (see, for example, Theorem 2.8 in [5]) can be used to show that *wep* $r(v_i, t) = l(v_i)n + O(\sqrt{l(v_i)n} \log n) = l(v_i)n(1 + o(1))$ during the whole period of length $L = O(n \log n)$. If the rank of the new vertex v_i, $R_i = r(v_i, i)$, is chosen uniformly at random from $[n]$, we get similar behaviour to the random labeling case with a label equal to R_i/n. In [10] the supermartingale method of Pittel et al. [8], as described in [13, Corollary 4.1] has been used to show the following useful lemma:

Lemma 2.1 ([10])**.** *Suppose that vertex v obtained an initial rank $R \geq \sqrt{n} \log^2 n$. Then,* wep $r(v, t) = R + \sqrt{n} \log^{3/2} n = R(1 + o(1))$ *to the end of its life.*

Note that there is no difference between these two approaches from the point of view of this paper. Therefore, in the rest of the note, $\{G_t\}_{t=0}^{\infty}$ is assumed to be a graph sequence generated by the rank-based attachment model, with random ranking scheme with uniform distribution (less 'rounded' scheme). Parameters d and η are assumed to be the initial degree and attachment strength parameters of the model as defined above.

3. Threshold for connectivity

In this section we study the connectivity of $\mathcal{P}_n(d, \eta)$ to illustrate similarities and differences both in results and methods between protean graphs and the standard binomial random graph model $G_{n,p}$.

Let $\rho_n(d, \eta)$ denote the probability that $\mathcal{P}_n(d, \eta)$ is connected. Before we move to new results let us first discuss the simplest case $\eta = 0$. Then, all vertices have the same weight and, since the ranking scheme does not matter, the model is equivalent to the classic protean graph. The probability that two vertices are connected by an edge is given by

$$\bar{p}(i, j) = \hat{p}(n) = 1 - (1 - 1/n)^d = d/n + O(d^2/n^2)\,.$$

Thus, one should expect that the threshold function for connectivity is the same as in the binomial random graph model $G(n,\hat{p})$. Theorem 3.1 proved in [7] shows that it is roughly the case but the dependence structure of $\mathcal{P}_n(d,0)$ influences the second term of the threshold function.

Theorem 3.1 ([7])**.** *Let* $d=d(n)=\log n-\frac{1}{2}\log\log n+a(n)$, $a(n)=o(\log\log n)$. *Then*

$$\lim_{n\to\infty}\rho_n(d,0)=\begin{cases}1 & \text{if} \quad a(n)\to\infty\\ \exp\big(-\sqrt{\pi/2}e^{-a}\big) & \text{if} \quad a(n)\to a\\ 0 & \text{if} \quad a(n)\to-\infty\,.\end{cases}$$

In the case $\eta\in(0,1)$ the threshold for the connectivity is affected by a constant factor.

Theorem 3.2. *Let* $\eta\in(0,1)$, $d=d(n)=\frac{\log n}{1-\eta}-\frac{2\log\log n}{1-\eta}+a(n)$, $a(n)=o(\log\log n)$. *Then*

$$\lim_{n\to\infty}\rho_n(d,\eta)=\begin{cases}1 & \text{if} \quad a(n)\to\infty\\ \exp\Big(-\frac{1-\eta}{\eta}e^{-a(1-\eta)}\Big) & \text{if} \quad a(n)\to a\\ 0 & \text{if} \quad a(n)\to-\infty\,.\end{cases}$$

Proof. Let v_i denote a vertex with $a(v_i,n)=i=xn$ and $q^+(v_i)$ ($q^-(v_i)$) denote the probability that v_i has no neighbour u with $a(u,n)>i$ ($a(u,n)<i$, respectively). Suppose that v_i obtained an initial rank $R\ge\sqrt{n}\log^2 n$. Then using Lemma 2.1, the probability in question is equal to

$$\begin{aligned}q^+(v_i\mid R) &= \prod_{j=i+1}^{n}\left(1-\frac{1-\eta}{n^{1-\eta}}(R+O(\sqrt{n}\log^{3/2}n))^{-\eta}\right)^{d}\\ &= \exp\Big(-d(1-\eta)(1-x)(R/n+O(n^{-1/2}\log^{3/2}n))^{-\eta}\Big)\,.\end{aligned}$$

Since R is taken uniformly at random from $[n]$, we get

$$\begin{aligned}q^+(v_i) &= \int_0^1\exp\Big(-d(1-\eta)(1-x)(l+O(n^{-1/2}\log^{3/2}n))^{-\eta}\Big)\,dl\\ &= \int_0^{n^{-1/4}}\exp\Big(-d(1-\eta)(1-x)(l+O(n^{-1/2}\log^{3/2}n))^{-\eta}\Big)\,dl\\ &+ \int_{n^{-1/4}}^1\exp\big(-d(1-\eta)(1-x)l^{-\eta}(1+O(n^{-1/5}))\big)\,dl\\ &= (1+o(1))\int_0^1\exp\big(-d(1-\eta)(1-x)l^{-\eta}(1+O(n^{-1/5}))\big)\,dl.\end{aligned}$$

(Note that we cannot control vertices with very small initial ranks but this does not cause a problem since for those vertices the probability of being isolated is negligible.) Now putting $A = d(1-\eta)(1-x)(1+O(n^{-1/5}))$ and then $u = Al^{-\eta}$ we obtain

$$q^+(v_i) \;=\; (1+o(1))\frac{A^{1/\eta}}{\eta}\int_A^\infty e^{-u}u^{-1-1/\eta}du \;=\; (1+o(1))\frac{A^{1/\eta}}{\eta}\Gamma(-1/\eta, A),$$

where $\Gamma(\cdot,\cdot)$ denotes the upper incomplete gamma function. Using an asymptotic formula for the gamma function (see, for example, [4]) we get

$$q^+(v_i) \;=\; (1+o(1))\frac{A^{1/\eta}}{\eta}e^{-A}A^{-1/\eta-1} \;=\; (1+o(1))\frac{\exp\left(-d(1-\eta)(1-x)\right)}{\eta d(1-\eta)(1-x)}.$$

(Note that an error term of $(1+O(n^{-1/5}))$ in the exponent is absorbed in $(1+o(1))$.)

In order to calculate $q^-(v_i)$ we use the fact that vertex v_i generated d edges at the time it was born but exactly $n-i$ vertices that were already in the graph at that time have been deleted. Since vertices are being removed uniformly at random we get

$$q^-(v_i) = (1+o(1))\frac{\binom{n-d}{(1-x)n-d}}{\binom{n}{(1-x)n}} = (1+o(1))(1-x)^d.$$

Therefore, for the expectation of the number Y_n of isolated vertices in $\mathcal{P}_n(d,\eta)$ we have

$$\begin{aligned}\mathbb{E}Y_n &= (1+o(1))n\int_0^1 q^-(v_{xn})q^+(v_{xn})dx \\ &= (1+o(1))\frac{n}{d(1-\eta)\eta}\int_0^1 (1-x)^{d-1}\exp\left(-d(1-\eta)(1-x)\right)dx.\end{aligned}$$

Substituting $u = d(1-\eta)(1-x)$ we get

$$\begin{aligned}\mathbb{E}Y_n &= (1+o(1))\frac{n}{[d(1-\eta)]^{d+1}\eta}\int_0^{d(1-\eta)} u^{d-1}e^{-u}dx \\ &= (1+o(1))\frac{n}{[d(1-\eta)]^{d+1}\eta}\;\gamma(d, d(1-\eta)),\end{aligned}$$

where $\gamma(\cdot,\cdot)$ denotes the lower incomplete gamma function. Using the following asymptotic expansion for the incomplete gamma function (so the error of truncation at N terms is of order at most the $(N+1)$st term)

$$\gamma(a,x) = -(1+o(1))x^a e^{-x}\sum_{k=0}^{\infty}\frac{(-a)^k b_k(\lambda)}{(x-a)^{2k+1}},$$

where $x = \lambda a$ and a goes to infinity, $0 < \lambda < 1$; the $b_k(\lambda)$'s satisfy $b_0 = 1, b_1 = \lambda, b_2 = \lambda(2\lambda+1)$ and $b_k = \lambda(1-\lambda)b'_{k-1} + (2k-1)\lambda b_{k-1}$ (see, for example, Section 8.11(iii) in [14]) we obtain

$$\begin{aligned}
\mathbb{E}Y_n &= (1+o(1))\frac{n}{[d(1-\eta)]^{d+1}\eta}\frac{-[d(1-\eta)]^d e^{-d(1-\eta)}}{d(1-\eta)-d} \\
&= (1+o(1))\frac{n}{d^2\eta(1-\eta)}e^{-d(1-\eta)} \\
&= (1+o(1))\frac{1-\eta}{\eta}e^{-a(1-\eta)}.
\end{aligned}$$

One can also check that, for a given integer $r \geq 2$, the rth factorial moment of Y_n tends to $\left(\frac{1-\eta}{\eta}e^{-a(1-\eta)}\right)^r$, so the random variable Y_n tends to a Poisson distribution and, in particular, the probability that $\mathcal{P}_n(d,\eta)$ contains no isolated vertex tends to $\exp\left(-\frac{1-\eta}{\eta}e^{-a(1-\eta)}\right)$ as n goes to infinity.

Not surprisingly, the threshold for disappearing isolated vertices is also the threshold for connectivity. Therefore, in order to finish the proof it is enough to show that if, say, $d(n) = \frac{\log n}{1-\eta} - \frac{3\log\log n}{1-\eta}$, the protean graph consists of one giant component and, perhaps, some number of isolated vertices.

It is clear that at most $2k/\sqrt{d}$ vertices from a spanning tree of a component of size k have degree more than $\sqrt{d}$. Hence, we can estimate the probability that the vertices from a tree have no neighbours outside this component by

$$\left(1-(1+o(1))\frac{1-\eta}{n}\right)^{d(k-2k/\sqrt{d})(n-k)} = \exp\left(-(1+o(1))d(1-\eta)k\left(1-\frac{k}{n}\right)\right)$$

(note that the probability that there is an edge between v_i and v_j $(i<j)$ is minimized if v_i had rank n when v_j was introduced). The probability that $\mathcal{P}_n(d,\eta)$ contains a component of size k, where $2 \leq k \leq (1-\eta)n/4$,

is bounded from above by

$$
\begin{aligned}
&\sum_{k=2}^{(1-\eta)n/4} \binom{n}{k} k^{k-2} \exp\left(-(1+o(1))d(1-\eta)k\left(1-\frac{k}{n}\right)\right)\left((1+o(1))\frac{d}{n^{1-\eta}}\right)^{k-1} \\
&\qquad \leq \sum_{k=2}^{(1-\eta)n/4} \left(\frac{ne}{k}\right)^{k} k^{k-2} \exp\left(-(1+o(1))\left(d(1-\eta)k\left(1-\frac{k}{n}\right)+(1-\eta)(k-1)\log n\right)\right) \\
&\qquad \leq \sum_{k=2}^{(1-\eta)n/4} \exp\left(-(1+o(1))\left(\left(1-\frac{1-\eta}{4}\right)k+(1-\eta)(k-1)-k\right)\log n\right) \\
&\qquad \leq \sum_{k=2}^{(1-\eta)n/4} \exp\left(-(1+o(1))\left(\frac{3(1-\eta)}{4}k-(1-\eta)\right)\log n\right) \\
&\qquad \leq n^{-(1+o(1))(1-\eta)/2},
\end{aligned}
$$

and tends to zero as $n \to \infty$. (Note that the probability that there is an edge between v_i and v_j $(i < j)$ is maximized if v_i had rank 1 when v_j was introduced.) It is also clear that there are no two components each containing a positive fraction of all vertices. Indeed, the expected number of pairs of vertex sets, each of size $(1-\eta)n/4$, with no edge between them is bounded from above by

$$
\binom{n}{(1-\eta)n/4}^{2}\left(1-(1+o(1))\frac{1-\eta}{n}\right)^{d((1-\eta)n/4)^2} = \exp(O(n)-\Omega(n\log n)) = o(1).
$$

Thus, by the Markov's inequality, a.a.s. the protean graph consists of a giant component and some number of isolated vertices, which completes the proof of the theorem. □

4. Recovery time

In this section we would like to come back to the protean process $\{G_t\}_{t=0}^{\infty} = \{\mathcal{P}_n^t(d,n)\}_{t=0}^{\infty}$ and study an interesting (from both theoretical and application point of view) property which does not have its counterpart for the classic random graph process $\{G(n,p)\}_{0\leq p\leq 1}$. Let $\mathcal{A}$ be a graph property such that $\mathcal{A}$ holds for $\mathcal{P}_n(d,\eta)$ a.a.s. but for $\tau(\mathcal{A})$, defined as

$$
\tau(A) = \min\{t : \mathcal{P}_n^t(d,\eta) \text{ has not } \mathcal{A}\}\,,
$$

we have $\mathbb{P}(\tau(\mathcal{A}) < \infty) = 1$, that is, with probability one at some stage of the protean process $\{\mathcal{P}_n^t(d,n)\}_{t=0}^{\infty}$ the property $\mathcal{A}$ disappears for some time. Then, the recovery time $\mathrm{rec}(\mathcal{A})$ for property $\mathcal{A}$ is defined as

$$
\mathrm{rec}(\mathcal{A}) = \min\{t > \tau(\mathcal{A}) : \mathcal{P}_n^t(d,\eta) \text{ has } \mathcal{A}\} - \tau(\mathcal{A})\,,
$$

that is, $\mathrm{rec}(\mathcal{A})$ tells us how long it takes for the protean process to regain a typical property $\mathcal{A}$. Note that since $\mathcal{A}$ holds a.a.s., and a.a.s. after $O(n \log n)$ steps each vertex of $\mathcal{P}_n(d,\eta)$ is renewed at least once, $\mathrm{rec}(\mathcal{A}) = O(n \log n)$ a.a.s. However, typically, the recovery time is smaller than the above universal upper bound implied by the coupon collector problem. The following theorem estimates $\mathrm{rec}(\mathcal{C})$, the recovery time for connectivity.

Theorem 4.1. *Let $\eta \in (0,1)$ and $d = \frac{a}{1-\eta} \log n$, where $a > 1$. Then*

$$\mathrm{rec}(\mathcal{C}) \cdot \frac{a \log n}{n} \xrightarrow{D} Z\,,$$

where the random variable Z has the exponential distribution, that is, for every $z \geq 0$, $\mathbb{P}(Z \geq z) = e^{-z}$.

Proof. The main part of the proof is to show that a.a.s. at time $\tau(\mathcal{C})$, the protean graph consists of a giant component and a single isolated vertex v of the rank $w = (1+o(1))n$ (note that such a rank maximizes the probability of being isolated). Then, in order to finish the proof it will be enough to show that a.a.s. graph becomes connected again when a new vertex creates an edge to v.

Let us focus on any period of $n \log^2 n$ steps of the protean process. The probabilities that during that time in the process we get

- an isolated vertex of the rank w, where $(w/n)^{-\eta} \leq 1+\varepsilon$,
- an isolated vertex of the rank w, where $(w/n)^{-\eta} > 1+\varepsilon$,
- a component of size k, $2 \leq k \leq 2n/3$,

we denote by $\rho_1(\varepsilon)$, $\rho_2(\varepsilon)$, and ρ_3, respectively. To estimate these probabilities, let us first compute the probability $\rho(i,j,t)$ that a vertex $v_i = v_{xn}$ becomes isolated at time t due to the fact that in this step we chose the only neighbour v_j of v_i in the preceding graph to be deleted. Let w_i and w_j denote the ranks in $\mathcal{P}_n^{t-1}(d,\eta)$ of v_i and v_j, respectively. Then, arguing as in the proof of Theorem 3.2, we may estimate $\rho(i,j,t)$ by

$$(1+o(1))\frac{1}{n} \cdot d \frac{1-\eta}{n^{1-\eta}} (w_i + O(n^{1/2} \log^{3/2} n))^{-\eta} \\ \cdot (1-x)^d \exp\left(-d(1-\eta)(1-x) \left(\frac{w_i}{n} + O(n^{-1/2} \log^{3/2} n) \right)^{-\eta} \right) \tag{1}$$

for $i < j$, and

$$(1+o(1))\frac{1}{n}\cdot d\frac{1-\eta}{n^{1-\eta}}(w_j + O(n^{1/2}\log^{3/2} n))^{-\eta} \\ \cdot(1-x)^{d-1}\exp\left(-d(1-\eta)(1-x)\left(\frac{w_i}{n} + O(n^{-1/2}\log^{3/2} n)\right)^{-\eta}\right) \tag{2}$$

for $i > j$.

Let $\varepsilon > 0$ be a positive constant. Let us denote by $\mathbf{A}_t(i)$ an event that a vertex v_i of the rank w_i becomes isolated at step t of the process and $(w_i/n)^{-\eta} \le 1 + \varepsilon/4$; moreover, let $\mathbf{A}_t = \bigcup_{i=1}^n \mathbf{A}_t(i)$. Events $\mathbf{B}_t(i)$ and $\mathbf{B}(i)$ are defined in a similar way, but this time we would like to have $(w_i/n)^{-\eta} > 1 + \varepsilon$. From (1) and (2) we get

$$\begin{aligned} \mathbb{P}(\mathbf{A}_t(i)) &\ge n^{-1+o(1)}(1+x)^d \exp\left(-d(1-\eta)(1-x)(1+\varepsilon/4)\right) \\ \mathbb{P}(\mathbf{A}_t(i)) &\le n^{-1+o(1)}(1+x)^d \exp\left(-d(1-\eta)(1-x)\right), \end{aligned}$$

while

$$\mathbb{P}(\mathbf{B}_t(i)) \le n^{-1+o(1)}(1+x)^d \exp\left(-d(1-\eta)(1-x)(1+\varepsilon)\right).$$

Using the same argument as in the proof of Theorem 3.2, we get

$$\begin{aligned} \mathbb{P}(\mathbf{B}_t) &\le \sum_{i=1}^n \mathbb{P}(\mathbf{B}_t(i)) \\ &\le n^{o(1)} \exp\left(-d(1-\eta)(1+\varepsilon)\right) \\ &\le n^{-(1+o(1))a(1+\varepsilon)}. \end{aligned} \tag{3}$$

In order to estimate the probability that $\mathbf{A}_t$ holds, we can bound the probability $\rho(i, i', j, t)$ that v_i and $v_{i'}$ become isolated at time t because the only their neighbour v_j is removed from the graph. It is clear (and so is omitted) that for $i \ne i'$ the events $\mathbf{A}_t(i)$ and $\mathbf{A}_t(i')$ are, in a way, 'weakly dependent', that is,

$$\mathbb{P}(\mathbf{A}_t(i) \cap \mathbf{A}_t(i')) = \mathbb{P}(\mathbf{A}_t(i))\mathbb{P}(\mathbf{A}_t(i'))n^{o(1)}.$$

Thus, Bonferroni's inequality gives

$$\begin{aligned} \mathbb{P}(\mathbf{A}_t) &= \mathbb{P}\Big(\bigcup_{i=1}^n \mathbf{A}_t(i)\Big) \\ &\ge \sum_{i=1}^n \mathbb{P}(\mathbf{A}_t(i)) - \sum_{1\le i<i'\le n} \mathbb{P}(\mathbf{A}_t(i) \cap \mathbf{A}_t(i')) \\ &\ge n^{o(1)} \exp\left(-d(1-\eta)(1+\varepsilon/3)\right) \\ &\ge n^{-a(1+2\varepsilon/5)}. \end{aligned}$$

From (3) we get immediately

$$\rho_2(\varepsilon) \le \sum_{t=1}^{n\log^2 n} \mathbb{P}(\mathbf{B}_t) \le n^{1-(1+o(1))a(1+\varepsilon)}. \tag{4}$$

Creating an isolated vertex at time t_1 affects the probability of creating another isolated vertex at time t_2 ($t_1 < t_2$). But, since ranks are well concentrated by the Chernoff's bound, it can be shown that

$$\mathbb{P}(\mathbf{A}_{t_1} \cap \mathbf{A}_{t_2}) = \mathbb{P}(\mathbf{A}_{t_1})\mathbb{P}(\mathbf{A}_{t_2})n^{o(1)}\ ,$$

Using Bonferroni's inequality one more time, we get

$$\rho_1(\varepsilon) \ge \mathbb{P}\Big(\bigcup_{t=1}^{n\log^2 n} \mathbf{A}_t \Big) \ge n^{1-a(1+\varepsilon/2)}. \tag{5}$$

Moreover, it can also be proved that

$$\rho_3 \le n^{1+o(1)}[\mathbb{P}(\mathbf{A}_t)]^2 \le \rho_2(\varepsilon) \tag{6}$$

(since the argument is fairly standard we omit details).

Now, let us consider first $n^{a(1+3\varepsilon/4)}\log^2 n$ steps of the protean process. From (4), (5) and (6) it follows that if the graph becomes disconnected during this period, then a.a.s. it is due to the appearance of a single isolated vertex of rank w with $(w/n)^{-\eta} \le 1+\varepsilon$. We will show that this is indeed the case, but in order to do that we split the time interval into a number of smaller subintervals to avoid dependent events.

Let $\mathbf{D}_k$, $k = 0, 1, \dots, k_0$, where $k_0 = n^{a(1+3\varepsilon/4)-1}/3$, be an event that between time-step $2kn\log^2 n$ and time-step $(2k+1)n\log^2 n$ an isolated vertex of the rank w appears with $(w/n)^{-\eta} \le 1+\varepsilon$. Let $\mathbf{F}$ be an event that every vertex was at least one time renewed in the time period $((2k-1)n\log^2 n, 2kn\log^2 n))$, for each $k = 1, \dots, k_0$. By the coupon collector problem, $\mathbf{F}$ holds *wep*. Moreover, $\mathbb{P}(\mathbf{D}_k) = \rho_1(\varepsilon)$ and, conditioned on $\mathbf{F}$, all events $\mathbf{D}_k$'s are independent. Thus, since $k_0\rho_1(\varepsilon)$ tends to infinity as $n \to \infty$, a.a.s. at least one of $\mathbf{D}_k$'s holds by the Chernoff's bound. Consequently, a.a.s. $\tau(\mathcal{C}) = n^{a(1+o(1))}$ and at the time $\tau(\mathcal{C})$, the protean graph consists of a giant component and a single isolated vertex v_{i_0} of rank $(1+o(1))n$.

The rest of the proof is straightforward. Let us consider the first $O(n/\log n)$ steps after the moment when the graph became disconnected. The probability that we renew vertex v_{i_0} at that time tends to zero as $n \to \infty$ and, by the argument similar to one we used to estimate $\rho_1(\varepsilon)$, $\rho_2(\varepsilon)$, ρ_3 above, so is the probability that we create an additional small component. Thus, the graph becomes connected if one of the renewed vertices will choose v_{i_0} as a neighbour. Since the rank of v_{i_0} can change only slightly during $O(n/\log n)$ steps, the probability that for some $z \geq 0$,

$$\mathrm{rec}(\mathcal{C}) \geq z \frac{n}{a \log n} = z \frac{n}{(1-\eta)d}\,,$$

is given by

$$\left[1 - (1+o(1))(1-\eta)\frac{d}{n^{1-\eta}} w^{-\eta}\right]^{z \frac{n}{(1-\eta)d}} = (1+o(1))e^{-z},$$

and the assertion follows. □

References

[1] A. Bonato, *A Course on the Web Graph*, American Mathematical Society, Providence, Rhode Island, 2008.

[2] 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.

[3] S. Fortunato, A. Flammini, and F. Menczer, *Scale-free network growth by ranking*, Phys. Rev. Lett. **96**(21): 218701 (2006).

[4] W. Gautschi, *The incomplete gamma function since Tricomi.* In Tricomi's Ideas and Contemporary Applied Mathematics, Atti dei Convegni Lincei, n. **147**, Accademia Nazionale dei Lincei, Roma, 1998, pp. 203–237.

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

[6] J. Janssen and P. Prałat, *Rank-based attachment leads to power law graphs*, preprint.

[7] T. Łuczak and P. Prałat, *Protean graphs*, Internet Mathematics **3** (2006), 21–40.

[8] 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.

[9] P. Prałat, *A note on the diameter of protean graphs*, Discrete Mathematics **308** (2008), 3399–3406.

[10] P. Prałat, *Protean graphs with a variety of ranking schemes*, Proceedings of the 2nd Annual International Conference on Combinatorial Optimization and Applications (COCOA'08), Lecture Notes in Computer Science, Springer, 2008, 149–159.

[11] P. Prałat and N. Wormald, *Growing protean graphs*, Internet Mathematics, accepted, 13pp.

[12] 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.

[13] 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.

[14] The NIST Digital Library of Mathematical Functions, `http://dlmf.nist.gov/`

Department of Mathematics and Statistics, Dalhousie University, Halifax NS, Canada
E-mail address: `pralat@mathstat.dal.ca`