A NOTE ON THE DIAMETER OF PROTEAN GRAPHS

PAWEŁ PRAŁAT

Abstract. The web graph is a real-world self-organizing network whose vertices correspond to web pages, and whose edges correspond to links between pages. Many stochastic models for the web graph have been recently proposed, with the aim of reproducing one or more of its observed properties and parameters. Some of the most intensely studied parameters for the web graph are the degree distribution and diameter.

A recent stochastic model of the web graph is the protean graph $\mathcal{P}_n(d,\eta)$. In this model, vertices are renewed over time, and older vertices are more likely to receive edges than younger ones. While previous work on the model focussed on the power law degree distribution of protean graphs, in this note we study its diameter. Since the protean graphs may be disconnected, we focus on the diameter of the giant component. Our main result is that diameter of the giant component of $\mathcal{P}_n(d,\eta)$ is equal to $\Theta(\log n)$, which supports experimental data observed in the actual web graph.

1. Introduction

Several new random graphs models have been introduced and analyzed in recent years for certain features observed in large-scale real-world networks such as the web graph W (see for example, the survey [5]). The graph W has vertices representing web pages, and whose edges correspond to links between these pages. Many graphical parameters have been studied in W, and several random graph models for W have been introduced and rigorously analyzed. As described in [5], some of these parameters include: degree distribution, diameter and average distances, clustering, and the presence of many bipartite cliques.

The experimental results reported in [1, 6] provide strong evidence that the diameter of the web graph is about the logarithm of its order, indicating that the web forms a so-called *small-world network*; see [14]. Several models for W generate graphs with a comparable diameter (see Theorems 3, 4, and 8 in [5]).

In this note, we consider the diameter of protean web graph model, written $\mathcal{P}_n(d,\eta)$, that was first introduced in [12] (see also a growing model [13]). It seems that protean graphs become more and more interesting, both for math and CS community, as a model based on ranking of vertices (see results of simulations [8] and theoretical ones [11]). Note also that the definition of the protean process allows us to study *recovery time* (see [12] for definition and results for connectivity);

1991 *Mathematics Subject Classification.* Primary: 05C80. Secondary: 68R10, 94C15.

Key words and phrases. random graphs, web graphs, protean graph, giant component, diameter.

The author acknowledges partial support from NSERC.

an interesting and very important property which does not have its counterpart for the other models.

Both in experimental studies [9] and in theoretical analysis of preferential attachment models, there is shown to be a strong correlation between age and degree. This consideration led to the development of protean graphs, in [12, 13]. The principle of protean graphs is that "the old get richer", i.e. the link probability favours older nodes. In the protean graph model the link probability is not directly related to age, but rather to a ranking based on age: the oldest node has rank 1, etc. The link probability is proportional to the rank raised to the power $-\eta$, where η is a parameter of the model.

The reason for this choice is twofold. Firstly, rank-based models have very attractive properties. They generally lead to a power law degree distribution where the exponent of the power law can be controlled in a natural way by varying η [8]. They also capture the intuitive notion that the difference of being the oldest or second-oldest node matters more than that of being the last and second-last born. Secondly, in this model nodes are renewed constantly, so the ages of the nodes are hard to track. For example, the oldest and second oldest node can vary widely in age, and the normalizing factor, the sum of the ages of living nodes, is a random variable that can be hard to trace.

We use a simplified version of the model; a more general version with a detailed description may be found in [12]. There are infinitely many discrete time-steps. We begin at time 0 with any fixed graph G with vertex set $[n] = \{1, 2, \dots n\}$ and any permutation $\sigma : [n] \to [n]$. In each time-step $t \geq 1$ we pick uniformly at random one of the vertices j to be *renewed* and update a permutation σ, by moving j to the end of the permutation, to reflect the order in which vertices have been chosen. The vertex x for which $\sigma(x) = 1$ is the *oldest* one, while the currently chosen vertex j satisfies $\sigma(j) = n$. For a vertex v, $\sigma(v)$ is the *rank* of v. We then delete from G all edges incident to j and generate d new edges (one by one) incident to j. (The vertex j can be viewed as a node that establishes connections with existing nodes in the network.) In each of these d independent choices, each vertex v is chosen with probability proportional to $\sigma(v)^{-\eta}$. The latter condition is natural since old vertices of small ranks should be more attractive to new vertices. To simplify notation, we assume that the ranks of the vertices of the protean graph coincide with their labels; that is, σ is the identity permutation.

If each vertex of a graph is renewed at least once, the random graphs appearing over time during the protean process are identical random objects whose properties do not depend on the graph G and permutation σ we started with; more precisely, the protean process is in stationary distribution. The random graph corresponding to this distribution is a protean graph $\mathcal{P}_n(d, \eta)$. See [12] for additional details on protean graphs.

Our main goal is to prove that $\mathcal{P}_n(d, \eta)$ contains a giant component whose vertices comprise a positive fraction of all vertices, and whose diameter is equal to $\Theta(\log n)$. To simplify proofs, we assume that $d \geq 13$ and $0.58 \leq \eta \leq 0.92$ (Note that these ranges of the parameters are enough to model the power law degree distribution observed in W; see [12].) However, we conjecture the theorem holds

for a wider range of parameters d and η. The precise statement of our main result is as follows.

Theorem 1. *Let $d \geq 13$, $d \in \mathbb{N}$ and $0.58 \leq \eta \leq 0.92$. W.h.p. a protean graph $\mathcal{P}_n(d,\eta)$ has one giant component, containing a positive fraction of all vertices, whose diameter is equal to $\Theta(\log n)$. The remaining components have $O(\log n)$ vertices.*

We deduce this result from Theorems 6 and 7 proved below, and by using Lemma 2 proved in [12]. Throughout, we use the abbreviations w.h.p. to denote that a statement holds with probability tending to 1 as $n \to \infty$. If A is an event, then we denote $\mathbb{P}(A)$ for its probability; if X is a random variable, then we denote $\mathbb{E}A$ for its expectation.

2. Proof of Theorem 1

In this section we give a proof of Theorem 1, by first proving the upper and then lower bounds on the diameter. Before we begin, we state a technical lemma. *From now on we assume that $d \geq 13, d \in \mathbb{N}$ and $0.58 \leq \eta \leq 0.92$.*

The lemma states roughly that $\mathcal{P}_n(d,\eta)$ is, in a way, related to a random graph on the set of vertices $[n]$, in which a pair of two vertices i, j, $\log^3 n \leq i < j \leq n$, is adjacent with probability

$$p(i,j) = (1-\eta)\frac{d}{n}\Big(\frac{j}{i}\Big)^{\eta},$$

independently for each such pair. We prove that for the diameter of the protean graph studied in the note, this is indeed the case. However, since we claim nothing about edges between 'small vertices' i, $1 \leq i < \log^3 n$, we cannot show a general theorem that relates, say, monotone properties of our model with the one with independent edges (as is done, for instance, in [7]). For similar reasons we cannot use the general theory of inhomogeneous sparse random graphs [4]. Nonetheless, Lemma 2 is strong enough for our purposes.

Let

$$E_1, E_2 \subseteq \{\{i,j\} : \log^3 n < i < j \leq n\}, \quad E_1 \cap E_2 = \emptyset\,.$$

For every $i, j \in [n]$, $r = 1, 2$, let

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

$$w(i,j) = (1-\eta)\frac{1}{n}\Big(\frac{j}{i}\Big)^{\eta} = \big(1 + O\big(n^{\eta-1}\big)\big)\frac{(i\,n/j)^{-\eta}}{\sum_{s=1}^{n} s^{-\eta}} \tag{1}$$

and

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

For the proof of the following result, see [12].

Lemma 2. *Let $\eta \in (0,1)$, d, E_1, E_2, $V_1(j)$, $w(i,j)$, $w_1(j)$ and $w_2(j)$ be defined as above, and let $|V_1(j)| \leq d$ for every $j \in [n]$. Let $P_n(E_1, E_2, d, \eta)$ denote the*

probability that all pairs from E_1 are edges of $\mathcal{P}_n(d,\eta)$, and no pair from E_2 is an edge of $\mathcal{P}_n(d,\eta)$. There are functions

$$f(d,n,\eta,E_1,E_2) = o(\exp(-\log^{3/2} n)) + \prod_{j=1}^{n}[1-(1+O(\log^{-1/2} n))(w_1(j)+w_2(j))]^{d-|V_1(j)|} \cdot d(d-1)\dots(d-|V_1(j)|+1)\prod_{i\in V_1(j)}(1+O(\log^{-1/2} n))w(i,j)$$

and

$$g(d,n,\eta,E_1,E_2) = o(\exp(-\log^{3/2} n)) + \prod_{j=1}^{n}(1-(1+O(\log^{-1/2} n))w_2(j))^{d-|V_1(j)|} \cdot d(d-1)\dots(d-|V_1(j)|+1)\prod_{i\in V_1(j)}(1+O(\log^{-1/2} n))w(i,j).$$

such that

$$f(d,n,\eta,E_1,E_2) \le P_n(E_1,E_2,d,\eta) \le g(d,n,\eta,E_1,E_2). \tag{2}$$

2.1. Upper bound on the diameter. We now show that a protean graph $\mathcal{P}_n(d,\eta)$ has one giant component, containing a positive fraction of all vertices, whose diameter is equal to $O(\log n)$, while the remaining components have $O(\log n)$ vertices. We reveal the component structure of $\mathcal{P}_n(d,\eta)$ step by step, using the *breadth-first search* (*BFS*) procedure or traversal. The main idea is to mark each vertex when we first visit it and keep track of what we have not completely explored. Each vertex will always be in one of the following three states: *undiscovered, discovered, or completely-explored*. For the BFS procedure we store the vertices in a *first in, first out queue*, written $\mathcal{Q}$; that is, we explore the oldest unexplored vertices first. We initialize the procedure by adding vertex v_0 we start with to $\mathcal{Q}$ and by changing its state from *undiscovered* to *discovered*. In each time-step k of the BFS process, we take a vertex v_k from $\mathcal{Q}$ (unless $\mathcal{Q}$ is empty), find all *undiscovered* neighbours of v_k, add them to $\mathcal{Q}$ and change their state to *discovered*. Finally, we mark v_k as *completely-explored*.

Let m_k denote the number of vertices that have already been discovered (both vertices being in *discovered* and *completely-explored* states). The *position* of a vertex is its rank in the last-renewed order. Note that the BFS process resembles a branching process [2]. In our case, the distribution of the number X_k of vertices we add to the queue $\mathcal{Q}$ in the k-th time-step, provided m_k of its elements have already been found, depends on the position of $v_1,\dots,v_{m_k}$ in the protean graph $\mathcal{P}_n(d,\eta)$, and m_k. In the branching process the distribution of the immediate offspring of a particle does not depend on the previous history of the process. Nonetheless, while $m_k < n^{2/3}$, one can show (see Theorem 3) that $\mathbb{P}(X_k \le 1) \le 1/3$. This means that the random variable X_k can be bounded from below by the independent random variable X with the following distribution

$$\begin{aligned} \mathbb{P}(X=0) &= 1/3\,, \\ \mathbb{P}(X=2) &= 2/3\,. \end{aligned} \tag{3}$$

Thus, the probability that the vertex is contained in a component of size at least $n^{2/3}$ is bounded from below by a probability that the branching process defined by a random variable X continues for a long time.

Theorem 3. *Let $k \in \mathbb{N}$, $v_k \in [n]$, $m_k < n^{2/3}$ and let X_k be the random variable defined as above. Then*

$$\mathbb{P}(X_k \leq 1) \leq 1/3 \,. \tag{4}$$

Proof. Note first that in order to estimate the random variable X_k one should condition on the entire detailed history of the BFS exploration. Unfortunately, we cannot use Lemma 2 directly to evaluate a conditional probability; the lemma should be applied twice, with the set of edges found so far, which has size up to $n^{2/3}$, but then the error term is too large. However, since we refresh vertices uniformly at random, it is known that with probability $1 - o(\exp(-\log^{3/2} n))$ for every $i, j, \log^3 n \leq i < j \leq n$, the rank of i at the moment when j is refreshed for the last time is well concentrated around its mean (see proof of Lemma 2 in [12]). Thus, (2) holds for conditional probability as well.

Denote the parent of vertex v_k (in the BFS tree) by $p[v_k]$. Observe that w.h.p. vertex v_k, at the moment when it is renewed for the last time, has not chosen a neighbour, except its parent $p[v_k]$, from the set of m_k vertices that have already been discovered. Indeed, the probability that v_k has chosen a neighbour from any set of $m_k < n^{2/3}$ vertices is bounded from above by

$$\big(1 + o(1)\big) d \Big(\sum_{i=1}^{n^{2/3}} i^{-\eta} \Big) \Big/ \Big(\sum_{i=1}^{n} i^{-\eta} \Big) = \big(1 + o(1)\big) d n^{-(1-\eta)/3} \,.$$

We first consider the probability that the random variable X_k is equal to zero. This probability conditioning on the event that $p[v_k] < v_k$ is larger than an analogous probability conditioning on the event that $p[v_k] > v_k$. Note that we cannot apply Lemma 2 for early vertices, but we can easily show that for any $v_k < \log^3 n$, the probability that $X_k = 0$ is less than or equal to an analogous probability for vertex $\lceil \log^3 n \rceil$. Then, using notation as in (1), by Lemma 2 we have that

$$\begin{aligned} &\mathbb{P}(X_k = 0) \qquad\qquad (5) \\ &\leq \big(1 + o(1)\big) \left(1 - \sum_{i=1, i \neq p[v_k]}^{v_k - 1} w(i, v_k) \right)^{d-1} \prod_{j=v_k+1}^{n} \big(1 - w(v_k, j)\big)^d \\ &= \big(1 + o(1)\big) \Big(1 - \frac{v_k}{n}\Big)^{d-1} \exp\left(- \big(1 + o(1)\big) d \, \frac{1 - \eta}{1 + \eta} \Big(\Big(\frac{v_k}{n}\Big)^{-\eta} - \frac{v_k}{n} \Big) \right) . \end{aligned}$$

Using similar arguments and calculation as in (5), we can prove the following inequality

$$\begin{aligned}&\mathbb{P}(X_k = 1) \qquad (6)\\ &\leq \big(1+o(1)\big)\Big(1-\frac{v_k}{n}\Big)^{d-1}\exp\left(-\big(1+o(1)\big)d\,\frac{1-\eta}{1+\eta}\Big(\Big(\frac{v_k}{n}\Big)^{-\eta}-\frac{v_k}{n}\Big)\right)\\ &\quad\cdot d\left(\frac{v_k}{n}\Big(1-\frac{v_k}{n}\Big)\cdot\frac{1-\eta}{1+\eta}\Big(\Big(\frac{v_k}{n}\Big)^{-\eta}-\frac{v_k}{n}\Big)\right).\end{aligned}$$

From (5) and (6), by considering cases for the parameters d and η, we may derive (4). (We omit this tedious though straightforward argument.) □

Theorem 3 states that random variables X_k are bounded from below by random variables $\bar{X}_k$, where $\bar{X}_k$ are independently and identically distributed random variables with distribution X defined in (3). Because the expected value of X is equal to $4/3$, one should expect that the BFS process, starting from a given vertex v, discovers a component of size at least $n^{2/3}$.

Theorem 4. *Consider the BFS traversal of a protean graph $\mathcal{P}_n(d,\eta)$, starting from a given vertex $v \in [n]$. The probability that the BFS process discovers a component of size at least $n^{2/3}$ is not smaller than $1/2$.*

Proof. Let X be a random variable defined in (3). A basic fact about branching process (see [2] or any textbook of probability theory) states that if $\mathbb{E}X > 1$, then with positive probability the process will continue forever. More precisely, let $f_X : [0,1] \to \mathbb{R}$ denote the probability-generating function of X, defined as $f_X(x) = \sum_{i\geq 0} x^i \mathbb{P}(X=i) = \frac{1}{3} + \frac{2}{3}x^2$. If $\mathbb{E}X = 4/3 > 1$ and $\mathbb{P}(X=0) = 1/3 > 0$, then the probability of extinction of the branching process is equal to x_0, where x_0 is the unique solution of the equation $f_X(x) = x$ that belongs to the interval $(0,1)$. In our case, this root is equal to $1/2$, which, based on Theorem 3, completes the proof of the theorem. □

The next theorem states that a BFS process dies out quickly (thereby discovering a component of size at most $150\log n$), or finds a component of size at least $n^{2/3}$. Recall that m_k denotes the number of vertices that have been discovered in k steps of the process (both vertices being in *discovered* and *completely-explored* states). Note also that in time-step k number of vertices being in *completely-explored* states is equal exactly to k.

Theorem 5. *Consider the BFS traversal of a protean graph $\mathcal{P}_n(d,\eta)$, starting from a given vertex $v \in [n]$. In each time-step k of the process, the following inequality holds*

$$\mathbb{P}\Big(m_k \leq \frac{7}{6}k \;\text{ and }\; 150\log n \leq m_k \leq n^{2/3}\Big) < o(n^{-2}).$$

Proof. It is straightforward to see that $1 + \sum_{i=1}^{k-1} X_i = m_k$. Let $X'_1, X'_2, \ldots, X'_{k-1}$ be an independent random variables with the distribution defined by (3). Using Theorem 3 it follows that the sequences $\{X_i\}_{i=1}^{k-1}$ and $\{X'_i\}_{i=1}^{k-1}$ can be coupled so

that $X_i \geq X_i'$ holds until either the BFS exploration dies out (that is, $m_k = k$) or $m_k \geq n^{2/3}$. Hence the probability we would like to estimate is less than or equal to the probability that $1 + \sum_{i=1}^{k-1} X_i' \leq 7k/6$. But

$$\mathbb{E}\Big(1 + \sum_{i=1}^{k-1} X_i'\Big) = 1 + (k-1)\mathbb{E}X_1' = \frac{4}{3}k - \frac{1}{3}\,,$$

and we can use the well-known method, going back at least to [3] (see also [10] for more details), of applying Markov's inequality to $\mathbb{E}\exp\left(u\sum_{i=1}^{k-1} X_i'\right)$ to show that for large k we have a good concentration. Since we only need to consider $k \geq 900\log n/7$, the assertion holds. □

We now prove the main result of this subsection.

Theorem 6. *W.h.p. a protean graph $\mathcal{P}_n(d,\eta)$ has one giant component containing a positive fraction of all vertices, whose diameter is equal to $O(\log n)$. The remaining components have $O(\log n)$ vertices.*

Proof. By Theorem 4 we conclude the existence of a component of size at least $n^{2/3}$. That there are no components of size l, $150\log n < l < n^{2/3}$, follows from Theorem 5. Thus, in order to prove the theorem, we show that $\mathcal{P}_n(d,\eta)$ has exactly one giant component containing a positive fraction of all vertices of small diameter.

Consider a pair of vertices v' and v'' which belong to components of size at least $n^{2/3}$. We determine the probability that the pair belong to different components. We run the BFS process of identifying vertices of the component containing v'. We stop the process when the number of discovered vertices is equal to $n^{2/3}$. According to Theorem 5, at the end of this procedure we are left with some set V' of vertices of the component containing v', such that at least $\frac{1}{7}n^{2/3}$ vertices from V' are in *discovered* states (vertices from set $\hat{V}'$ stored in the queue $\mathcal{Q}$); that is, we do not check out all their incident edges. We next run a similar process starting at the vertex v''. Then, either we join v'' to some of the vertices which belong to V', or end up with some set of vertices V'' of the component containing v'', among which at least $\frac{1}{7}n^{2/3}$ vertices from set $\hat{V}''$ have not been completely explored yet. Now, one can point out two subsets $\bar{V}' \subset \hat{V}'$ and $\bar{V}'' \subset \hat{V}''$, each containing $\frac{1}{14}n^{2/3}$ vertices, such that for every pair of vertices $i \in \bar{V}'$ and $j \in \bar{V}''$ $i < j$ (or for every pair of vertices $i \in \bar{V}'$ and $j \in \bar{V}''$ $i > j$). The probability that there are no edges between vertices of $\bar{V}'$ and $\bar{V}''$ is bounded from above by

$$\Big(1 - \frac{n^{2/3}}{14}\frac{1-\eta}{n}\Big)^{\frac{n^{2/3}}{14}} = o(n^{-2})\,.$$

Hence, the probability that $\mathcal{P}_n(d,\eta)$ contains two vertices v' and v'' which belong to two different components both of size at least $n^{2/3}$ tends to 0 as $n \to \infty$.

Thus, we have shown that w.h.p. the vertices of $\mathcal{P}_n(d,\eta)$ can be divided into two classes: "small" ones, which belong to components of size at most $O(\log n)$, and "large" ones, contained in one large component of size at least $n^{2/3}$. Observe that from Theorem 4 it follows that the probability that a vertex is small is bounded

from above by 1/2. Hence the expectation of the number Y of small vertices is smaller than $n/2$. Finally, estimating the variance and using Chebyshev's inequality we find that w.h.p. the giant component of the protean graph contains at least $(1-o(1))n/2$ vertices.

To complete the proof, we need to estimate the diameter of the giant component. Theorem 5 states that, when we discover more than $150\log n$ of vertices, the BFS process spreads quickly. More precisely, if we denote by D_k the number of vertices at distance at most k, then w.h.p. $D_{k+1} \geq \frac{7}{6}D_k$. Then the diameter of the graph induced by the set V' is bounded from above by $150\log n + \log_{\frac{7}{6}} n^{2/3} = O(\log n)$, which implies, according to the fact we have just proved, the diameter of the giant component is equal to $O(\log n)$. □

2.2. Lower bound on the diameter. An *isolated path* P is an induced path whose vertices are joined to no other vertices except ones in P, with the exception of exactly one of its endpoints. To prove that the diameter of the giant component of a protean graph $\mathcal{P}_n(d,\eta)$ is equal to $\Omega(\log n)$, we show that w.h.p. there is an isolated path of length $\Theta(\log n)$ whose first vertex is connected to the giant component. An isolated path P is *special* in $\mathcal{P}_n(d,\eta)$ if

(1) P has length $k = k(n) = \frac{\log n}{4d - 2\log(1-\eta)}$,
(2) the first vertex x_1 of P belonging to interval $[1, n/2)$ is connected to the giant component, and
(3) all vertices of P different than x_1 belong to $[n/2, 3n/4]$.

Let Y be random variable denoting the number of special paths in $\mathcal{P}_n(d,\eta)$. The following theorem establishes the lower bound in Theorem 1, and hence, finishes its proof.

Theorem 7. (1) $\mathbb{E}Y \geq n^{1/2}$.
(2) *W.h.p.,* $Y \geq 1$.

Proof. For item (1), let $x_1 \in [1, \frac{1}{2}n)$ and $x_i \in [\frac{1}{2}n, \frac{3}{4}n]$ for every $2 \leq i \leq k+1$. Let $B(x_1, x_2, \ldots, x_{k+1})$ denote the event that a protean graph $\mathcal{P}_n(d,\eta)$ contains an isolated path $(x_1, x_2, \ldots, x_{k+1})$, and let $C(x_1)$ denote the event that vertex x_1 belongs to the giant component. Finally, let

$$A(x_1, x_2, \ldots, x_{k+1}) = B(x_1, x_2, \ldots, x_{k+1}) \cap C(x_1) \; .$$

We can use Lemma 2 and calculation similar to (5) to show that the probability that $x_i \in [\frac{1}{2}n, \frac{3}{4}n)$ has no neighbours (excluding vertices x_{i-1} and x_{i+1}) can be bounded from below by

$$\begin{aligned} \big(1+o(1)\big)\Big(1-\frac{x_i}{n}\Big)^d \exp\left(-\big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big(\Big(\frac{x_i}{n}\Big)^{-\eta} - \frac{x_i}{n}\Big)\right) \\ \geq \Big(\frac{1}{4}\Big)^d \exp\left(-d\frac{1-\eta}{1+\eta}\Big(\Big(\frac{1}{2}\Big)^{-\eta} - \frac{1}{2}\Big)\right) \geq \Big(\frac{1}{4\sqrt{e}}\Big)^d . \end{aligned}$$

Although the existence of an isolated path affects the probability that x_1 is connected to the giant component, this influence is not strong (note, that $k = O(\log n)$ and $x_2 > x_1$). Then one can use the argument used in the proof of Theorem 4

to show that path $(x_1, x_2, \ldots, x_{k+1})$ is connected to the giant component with probability at least $(1 + o(1))/2 > 1/3$. Thus, the following inequality holds

$$\mathbb{P}(A(x_1, x_2, \ldots, x_{k+1})) \geq \big(1 + o(1)\big)\frac{1}{3}\Big(\frac{1-\eta}{n}d\Big)^k \Big(\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big)^k .$$

Let $Y(x_1, x_2, \ldots, x_{k+1})$ be the indicator variable of the event $A(x_1, x_2, \ldots, x_{k+1})$. Then

$$Y = \sum_{1 \leq x_1 < 1/2n} \sum_{1/2n \leq x_2, \ldots, x_{k+1} \leq 3/4n} Y(x_1, x_2, \ldots, x_{k+1}) .$$

Hence,

$$\begin{aligned} \mathbb{E}Y &= \sum_{1 \leq x_1 < 1/2n} \sum_{1/2n \leq x_2, \ldots, x_{k+1} \leq 3/4n} \mathbb{P}(A(x_1, x_2, \ldots, x_{k+1})) \\ &\geq \frac{n}{2}\binom{n/4}{k} k! \big(1 + o(1)\big)\frac{1}{3}\Big(\frac{1-\eta}{n}d\Big)^k \Big[\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \\ &\geq n\Big[(1-\eta)\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \geq n^{1/2} \end{aligned}$$

which proves item (1) of the theorem.

For (2), we prove next that Y is concentrated around its mean; more precisely, $\mathrm{Var}Y = o\big((\mathbb{E}Y)^2\big)$, where $\mathrm{Var}Y$ is the variance of Y. By Chebyshev's inequality, w.h.p. $Y \geq 1$, which proves item (2) and the theorem.

Let us consider two paths

$$\hat{x} = (x_1, x_2, \ldots, x_{k+1}), \quad \text{and} \quad \hat{y} = (y_1, y_2, \ldots, y_{k+1}).$$

These paths are vertex-disjoint or have exactly one common vertex, that is, $z = x_1 = y_1$. (We consider such a pairs of paths only since these paths can occur simultaneously. The contribution to the covariance from pairs of events which cannot both occur is negative, and so do not affect our calculations.) Note that the existence of one path affects the probability that the second path exists, but one can use Lemma 2 to show that this influence is not strong (even when paths have one common vertex z, because $z < x_2$ and $z < y_2$). Then we get

$$\begin{aligned} \mathbb{P}(A(\hat{x}) \cap A(\hat{y})) &= (1 + o(1))\mathbb{P}(B(\hat{x}))\mathbb{P}(B(\hat{y}))\mathbb{P}(C(x_1))\mathbb{P}(C(y_1)) \\ &= (1 + o(1))\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y})) , \end{aligned}$$

when $\hat{x}$ and $\hat{y}$ are disjoint and

$$\begin{aligned} \mathbb{P}(A(\hat{x}) \cap A(\hat{y})) &= (1 + o(1))\mathbb{P}(B(\hat{x}))\mathbb{P}(B(\hat{y}))\mathbb{P}(C(z)) \\ &= O(1)\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y})) , \end{aligned}$$

when paths have one common vertex $z = x_1 = x_2$. Let $\mathrm{Cov}(Z_1, Z_2)$ be the covariance of the variables Z_1, Z_2. Then

$$\begin{aligned}\sum_{\hat{x},\hat{y}} \mathrm{Cov}(Y(\hat{x}), Y(\hat{y})) &= \sum_{\hat{x},\hat{y}} \Big[\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) - \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\Big] \\ &= \sum_{\hat{x}} \Big[\mathbb{P}(A(\hat{x})) \sum_{\hat{y}, y_1 \neq x_1} o(1)\mathbb{P}(A(\hat{y})) + \sum_{\hat{y}, y_1 = x_1} O(1)\mathbb{P}(A(\hat{y}))\Big] \\ &= o\Big(\sum_{\hat{x}} \mathbb{P}(A(\hat{x})) \sum_{\hat{y}} \mathbb{P}(A(\hat{y}))\Big) = o\Big((\mathbb{E}Y)^2\Big)\ . \end{aligned} \tag{7}$$

Because the random variables $Y(\hat{x})$ have values either 0 and 1, it is straightforward to see that

$$\sum_{\hat{x}} \mathrm{Var}Y(\hat{x}) \leq \sum_{\hat{x}} \mathbb{E}Y(\hat{x})^2 = \sum_{\hat{x}} \mathbb{E}Y(\hat{x}) = \mathbb{E}Y\ . \tag{8}$$

From (7) and (8), we obtain that $\mathrm{Var}Y = o\big((\mathbb{E}Y)^2\big)$. □

References

[1] R. Albert, H. Jeong and A. Barabási, *Diameter of the World-Wide Web*, Nature **401** (1999), 130-131.

[2] K.B. Athreya and P.E. Ney, *Branching processes*, Die Grundlehren der mathematischen Wissenschaften, Band 196, Springer, Berlin (1972).

[3] S. Bernstein, *On a modification of Chebyshev's inequality and of the error formula of Laplace*, Ann. Sci. Inst. Sav. Ukraine, Sect. Math. **1**, 38-49 (Russian).

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

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

[6] 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.

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

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

[9] B.A. Huberman, L.A. Adamic, *Growth dynamics of the world-wide web*, Nature **401** (1999) 131.

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

[11] J. Janssen and P. Prałat, *Rank-based attachment leads to power law graphs*, Internet Mathematics, submitted.

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

[13] P. Prałat and N. Wormald, *Growing Protean Graphs*, Internet mathematics, accepted.

[14] D.J. Watts, *Small Worlds*, Princeton University Press, Princeton, 1999.

Department of Mathematics and Statistics, Dalhousie University, Halifax NS, Canada B3H 3J5

E-mail address: pralat@mathstat.dal.ca