Crawling on simple models of web graphs

Colin Cooper
Department of Computer Science,
King's College,
University of London,
London WC2R 2LS, UK
ccooper@dcs.kcl.ac.uk

Alan Frieze*
Department of Mathematical Sciences,
Carnegie Mellon University,
Pittsburgh PA15213.
alan@random.math.cmu.edu

August 27, 2003

1 Introduction

We consider a simple model of an agent (which we call a spider) moving between the nodes of a randomly growing web graph. It is presumed that the agent examines the page content of the node for some specific topic. In our model the spider makes a random walk on the existing set of vertices. We compare the success of the spider on web graphs of two distinct types. For a random graph web graph model, in which new vertices join edges to existing vertices uniformly at random, the expected proportion of unvisited vertices tends to 0.57. For the comparable copy-based web graph model, in which new vertices join edges to existing vertices proportional to vertex degree, the expected proportion of unvisited vertices tends to 0.59.

A web graph is a sparse connected graph designed to capture some properties of the www. Studies of the graph structure of the www were made by [5] and [12] among others. There are many models of web graphs designed to capture the structure of the www found in the studies given above. For example see references [1], [2], [3], [4], [6], [7], [8], [9], [10], [11], [13], [14], [15], [16], [17], [18], [20], [21] and [22]. for various models.

In the simple models we consider, each new vertex directs m edges towards existing vertices, either randomly (random graph model) or according to the degree of existing vertices (copy model). Once a vertex has been added the direction of the edges is ignored.

There are several types of search which might be applied to the www. Complete searches of the web, usually in a breadth first manner, are carried out by search engines. Link and page data for visited pages is stored, and from the link data an undirected model of the www can be constructed. This

*Supported in part by NSF grant CCR-9818411.

model may be replaced when a new search is made at a future time period or may be continously updated by a continuously ongoing search. Such processes require considerable on-line and off-line memory.

Another possibility, requiring less memory, is a search by an agent (sniffer, spider) which examines the semantic content of nodes for some specific topic. This type of search can be made directly on the www or on a (continously updating) model of the www stored by a search engine. Typical search strategies might include: moving to a random neighbour (sampling pages for content), selecting a random neighbour of large degree (locating the hub/authority vertices of the search topic) or selecting a random neighbour of low degree (favouring the discovery of newer vertices during the search).

In this paper we consider the following abstract scenario. We have a sequence $(G(t), t = 1, 2, ...)$ of connected random graphs. The graph $G(t)$ is constructed from $G(t-1)$ by adding the vertex t, and m random edges from vertex t to $G(t-1)$. We refer to such graphs as web-graphs. See references [1], [2], [3], [5], [8], [9], [12], [16], [17], [18], [20] and [21] for various models of this and related types.

There is also a spider S walking randomly from vertex to vertex on the evolving graph $G(t)$.

The parameter ν_t we estimate is the expected number of vertices which have not been visited by the spider at step t, when t is large. This process is intended to model the success of a search-engine spider which is randomly crawling the world wide web looking for new web-pages.

To be more precise, we consider the following model for $G(t)$. Let $m \geq 1$ be a fixed integer. Let $[t] = \{1, ..., t\}$ and let $G(1) \subset G(2) \subset \cdots \subset G(t)$. Initially $G(1)$ consists of a single vertex 1 plus m loops. For $t \geq 2$, $G(t)$ is obtained from $G(t-1)$ by adding the vertex t and m randomly chosen edges $\{t, v_i\}, i = 1, 2, \ldots, m$, where

Model 1: The vertices $v_1, v_2, \ldots, v_m$ are chosen independently and uniformly with replacement from $[t-1]$.

Model 2 The vertices $v_1, v_2, \ldots, v_m$ are chosen proportional to their degree after step $t-1$. Thus if $d(v, \tau)$ denotes the degree of vertex v in $G(\tau)$ then for $v \in [t-1]$ and $i = 1, 2, \ldots, m$,

$$\mathbf{Pr}(v_i = v) = \frac{d(v, t-1)}{2m(t-1)}.$$

While vertex t is being added, the spider S is sitting at some vertex X_{t-1} of $G(t-1)$. After the addition of vertex t, and before the beginning of step $t+1$, the spider now makes a random walk of length ℓ, where ℓ is a fixed positive integer independent of t.

It seems unlikely that at time t, S will have visited every vertex. Let $\nu_{\ell,m}(t)$ denote the expected number of vertices not visited by S at the end of step t.

We will prove the following theorem:

Theorem 1. *In either model, if m is sufficiently large then, as $t \to \infty$,*

$$\nu_{\ell,m}(t) \sim \mathbf{E} \sum_{s=1}^{t} \prod_{\tau=s}^{t} \left(1 - \frac{d(s,\tau)}{2m\tau}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right)^{\ell}. \tag{1}$$

□

We have said that m is fixed. We however have to accept errors of order $1/m$ and so in our asympotics we let $t \to \infty$ first and then take m large.

Let

$$\eta_{\ell,m} = \lim_{t\to\infty} \frac{\nu_{\ell,m}(t)}{t}.$$

We will show that this gives the following limiting results for the models we consider.

Theorem 2. *Let* $\eta_\ell = \lim_{m\to\infty} \eta_{\ell,m}$, *then*

(a) *For Model 1,*

$$\eta_\ell = \sqrt{\frac{2}{\ell}} e^{(\ell+2)^2/(4\ell)} \int_{(\ell+2)/\sqrt{2\ell}}^{\infty} e^{-y^2/2}\, dy.$$

In particular, $\eta_1 = 0.57\cdots$, *and* $\eta_\ell \sim 2/\ell$ *as* $\ell \to \infty$.

(b) *For Model 2*

$$\eta_\ell = e^\ell 2\ell^2 \int_\ell^\infty y^{-3} e^{-y}\, dy.$$

In particular, $\eta_1 = 0.59\cdots$, *and* $\eta_\ell \sim 2/\ell$ *as* $\ell \to \infty$.

□

Thus for large m, t and $\ell = 1$ it is slightly more difficult for the spider to crawl on a web-graph whose edges are generated by a copying process (Model 2) than on a uniform choice random graph (Model 1).

2 Proof of Theorem 1: The main ideas

We first consider the case where $\ell = 1$ and then generalise this case. When $\ell = 1$ the spider makes a random move to an adjacent vertex after vertex t has been added. The construction of $G(t)$ is really the construction of a digraph $D(t)$ where the direction of the arcs (x, y) satisfies $x > y$. The space $\mathcal{G}(t)$ of graphs $G(t)$ induces its measure from this space of digraphs.

Let $\Omega(t)$ denote the set of pairs $(G(t), W(t))$ where $G(t) \in \mathcal{G}(t)$ and $W(t)$ belongs to the set $\mathcal{W}_G(t)$ of t-step walks taken by the spider S which are compatible with the construction of G. Among other things, this means that the τ-th vertex of $G(t)$ visited by the walk must be in $[\tau]$.

The main idea of the proof is as follows. We fix a vertex s and estimate the probability that it is not visited by the end of step t. Thus for $s \le \tau \le t$ we define the events

$$\mathcal{A}_s(\tau) = \{\omega \in \Omega(t) : \text{ Vertex } s \text{ is not visited by } \mathsf{S} \text{ during the time interval } [s, \tau]\}.$$

Let

$$t_0 = t - 100(\ln t)^3.$$

It is convenient to condition on the sequence $d(s, \tau)$ for $\tau = s, s+1, \ldots, t_0$. Let $\boldsymbol{\theta} = (\theta_\tau : 1 \le \tau \le t_0)$ be integers satisfying

$$\theta_1 = \cdots = \theta_s = m \le \theta_\tau \le \theta_t \le \Delta_t^* = 10(\ln t)^5 \text{ and } \theta_{\tau+1} \le \theta_\tau + 5 \text{ for } \tau \le t_0 \tag{2}$$

and let $\boldsymbol{\Theta} = \{\boldsymbol{\theta} : \text{ (2) holds}\}$.

Let

$$\mathcal{D}(\boldsymbol{\theta}) = \{(G(t), W(t)) \in \Omega(t) : d(s, \tau) = \theta_\tau,\ s \le \tau \le t\},$$

and for some event $\boldsymbol{C}$ let $\mathbf{Pr}_{\boldsymbol{\theta}}(\boldsymbol{C}) = \mathbf{Pr}(\boldsymbol{C} \mid \mathcal{D}(\boldsymbol{\theta}))$ be the probability of the corresponding conditional event.

We will show

Lemma 1. *In both Model 1 and Model 2,*

$$\mathbf{Pr}(\bigcup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{D}(\boldsymbol{\theta})) = 1 - \widetilde{O}(t^{-3})^1.$$

$\square$

Let

$$\sigma_0 = \frac{\ln t}{100 \ln\ln t}$$

and let

$$B_t = \{s \in [t/\ln t, t] : \ s \text{ is within distance } \sigma_0 \text{ of a cycle of length at most } \sigma_0\}.$$

Let

$$\mathcal{G}_1(t) = \{G(t) : \ |B_t| \leq t^{7/8}\}.$$

We will prove that

Lemma 2. *If $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ then, in both Model 1 and Model 2,*

$$\mathbf{Pr}_{\boldsymbol{\theta}}(G(t) \notin \mathcal{G}_1(t)) = o(t^{-3}).$$

$\square$

We then prove

Lemma 3. *If $s \in [t/\ln t, t)$, $s \notin B_t$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then*

(a)

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\overline{\mathcal{A}}_s(t) \mid \mathcal{A}_s(t-1)) = \frac{\theta_{t_0}}{2mt_0}\left(1 + O\left(\tfrac{1}{m}\right)\right) + O(t^{-3})\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1}.$$

(b)

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(s)) = 1 - O(s^{-1}).$$

$\square$

(We condition on $\boldsymbol{\theta}$ in order to avoid some conditioning of the degree $d(s, t_0)$ due to assuming $\mathcal{A}_s(t_0)$.)

From this we prove Theorem 1 as follows: If $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $s \in [t/\ln t, t]$, $s \notin B_t$ then

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t)) = \left(1 - \frac{\theta_{t_0}}{2mt}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right)\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1)) + \widetilde{O}(t^{-3}).$$

We see then that if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $s \in [t/\ln t, t)$, $s \notin B_t$ and if $\tau_0 = \tau - 100(\ln t)^3$,

$$\begin{aligned}\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t)) &= \prod_{\tau=s+1}^{t}\left(1 - \frac{\theta_{\tau_0}}{2m\tau}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right) & (3)\\ &= \prod_{\tau=s+1}^{t}\left(1 - \frac{\theta_{\tau}}{2m\tau}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right) & (4)\end{aligned}$$

[1] The $\widetilde{O}$ notation ignores polylog factors.

Note that we can go from (3) to (4) because $\theta_\tau = \theta_{\tau_0}$ except for at most $100(\ln t)^3\Delta_t^*$ instances.

Thus absorbing the cases where $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$ into the error term, (see Lemma 1), summing out the conditional probabilities over degree sequences, we get that for $s \in [t/\ln t, t)$, $s \notin B_t$

$$\begin{aligned}\mathbf{Pr}(\mathcal{A}_s(t)) &= \sum_{\boldsymbol{\theta}} \mathbf{Pr}(\mathcal{D}(\boldsymbol{\theta})) \prod_{\tau=s+1}^{t} \left(1 - \frac{\theta_\tau}{2m\tau}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right) \\ &= \mathbf{E} \prod_{\tau=s+1}^{t} \left(1 - \frac{d(s,\tau)}{2m\tau}\left(1 + O\left(\tfrac{1}{m}\right)\right)\right).\end{aligned}$$

Note that the contribution of $s \in [1, t/\ln t] \cup B_t$ to the expectation $\nu_{\ell,m}(t)$ can only be $o(t)$ and (1) follows. □

3 Proof of Theorem 1: The details

We emphasise that $s \geq t/\log t$ throughout and that m is a sufficiently large constant.

3.1 Proof of Lemma 1: Model 1

The degree $d(s,t)$ of vertex s in $G(t)$ is distributed as

$$m + B(m, (s+1)^{-1}) + \cdots + B(m, t^{-1}) \tag{5}$$

where the binomials $B(m, \cdot)$ are independent.

Lemma 4.

(a) $\mathbf{Pr}(\Delta(G(t)) \geq 2m\ln t) = \widetilde{O}(t^{-3})$
where $\Delta(G(t))$ is the maximum degree in $G(t)$.

(b) $\mathbf{Pr}(\exists \tau :\ d(s,\tau+1) - d(s,\tau) > 5) = \widetilde{O}(t^{-4})$.

Proof $\mathbf{E}\,(d(s,t)) = m(1 + H_t - H_s) \leq m(2 + \ln t/s)$ where $H_k = 1 + \frac{1}{2} + \cdots + \frac{1}{k}$. (a) now follows from Theorem 1 of Hoeffding [19]. (b) is easy, since $d(s,\tau+1) - d(s,\tau) = B(m, \tau^{-1})$. □

3.2 Proof of Lemma 1: Model 2

Lemma 5.

(a) $\mathbf{Pr}(d(s,t) \geq 10(\ln t)^5) = \widetilde{O}(t^{-3})$.

(b) $\mathbf{Pr}(\exists \tau :\ d(s,\tau+1) - d(s,\tau) > 5) = \widetilde{O}(t^{-3})$.

Proof (a) In order to get a crude upper bound on $d(s,t)$, we divide the interval $[s,t]$ into sub-intervals using the points (nearest to) $s, se^{1/8}, ...se^{r/8}, ..., se^{k/8}$. Here $se^{(k-1)/8} < t \leq \lceil se^{k/8} \rceil$, so that $k \leq 8\ln\ln t$, as $s \geq t/\ln t$.

Suppose that, at the start of $I_r = (\lceil se^{r/8} \rceil, \lceil se^{(r+1)/8} \rceil]$ we have an upper bound $d(r)$ on the degree of vertex s. We prove that if $d(r) \geq 10\ln t$ then $d(r+1) \leq 2d(r)$ with probability $1 - o(t^{-3})$.

Now as long as the degree of s is $\leq 2d(r)$, the number X_τ of edges acquired at step $\tau \in I_r$ is dominated by $B(m, d(r)/(m(\tau-1)))$, so that the number of edges gained during this time has expected value

$$\leq 2d(r) \ln e^{1/8} = \frac{d(r)}{4}.$$

Thus, by Chernoff bounds, provided $d(r) \geq 10 \ln t$,

$$\mathbf{Pr}(d(r+1) \geq 2d(r)) \leq \mathbf{Pr}\left(\sum_{\tau \in I_r} B(m, d(r)/(m(\tau-1))) \geq d(r)\right) \leq \left(\frac{e}{4}\right)^{d(r)} = o(t^{-3})$$

and thus $d(r+1) < 2d(r)$ with probability $1 - o(t^{-3})$. Choosing $d(0) = 10 \ln t$, we see that

$$d(s,t) < d(0)2^k \leq d(0)(\ln t)^4 = 10(\ln t)^5.$$

This proves (a). For (b) we use (a) and the fact that $d(s, \tau+1) - d(s, \tau)$ can then be dominated by $B(10(\ln t)^5, (2m\tau)^{-1})$. $\square$

3.3 Proof of Lemma 2: Model 1

Fix $t/\ln t \leq i_1 < \cdots < i_5 \leq t$ and let $I = \{i_1, \ldots, i_5\}$. We estimate $\mathbf{Pr}(I \subseteq B_t)$.

For each partition $\mathcal{P}$ of I into parts $A_1, \ldots, A_k$ we consider the event

$\mathcal{E}_\mathcal{P} = \{\exists$ small cycles $C_1, \ldots, C_k$ and paths $P_v, v \in I$ such that
(i) $|C_i|, |P_v| \leq \sigma_0$ for all i, v.
(ii) If $v \in A_i$ then P_v joins v to $C_i \cup \bigcup_{w \in A_i, w<v} P_w$,
(iii) P_v is edge disjoint from and shares one (endpoint) vertex with $C_i \cup \bigcup_{w \in A_i, w<v} P_w$,
(iv) The k collections $C_i, P_v, v \in A_i$ are pair-wise vertex disjoint. $\}$

Thus $\{I \subseteq B_t\} \subseteq \bigcup_\mathcal{P} \mathcal{E}_\mathcal{P}$ and

$$\mathbf{Pr}(\mathcal{E}_\mathcal{P}) \leq \sum_{\substack{C_1,\ldots,C_k \\ P_v, v \in I}} \prod_{(x,y) \in F^*} \frac{m}{\max\{x,y\}} \tag{6}$$

where F^* denotes the edge set of $\bigcup C_i \cup \bigcup P_v$. The term $\frac{m}{\max\{x,y\}}$ is a bound on the probability of the existence of edge (x, y) given the appearance or absence of other edges, not incident with $\max\{x, y\}$.

Thus

$$\mathbf{Pr}(\mathcal{E}_\mathcal{P}) \leq \prod_{r=1}^{5} \frac{m}{i_r} \sum_{\substack{3 \leq |C_i| \leq \sigma_0 \\ i=1,\ldots,k}} \sum_{\substack{0 \leq |P_v| \leq \sigma_0 \\ v \in I}} \prod_{v \in V^*} \frac{m}{v}$$

where V^* denotes the vertex set of $\bigcup C_i \cup \bigcup P_v$, less I,

$$\leq \left(\frac{m \ln t}{t}\right)^5 \sum_{\ell=1}^{10\sigma_0} \left(\sum_{v=1}^{t} \frac{m}{v}\right)^\ell$$
$$= o(t^{-4}).$$

Thus

$$\mathbf{E}\left(\binom{|B_t|}{5}\right) = o(t)$$

and

$$\mathbf{Pr}(|B_t| \geq t^{7/8}) \leq \frac{\mathbf{E}\left(\binom{|B_t|}{5}\right)}{\binom{t^{7/8}}{5}} = o(t^{-3}).$$

□

3.4 Proof of Lemma 2: Model 2

For this model we replace 5 by 10 and let $I = \{i_1, i_2, \dots, i_{10}\}$. Let

$$\mathcal{G}_2(t) = \{G(t) : \ d(s,t) \leq 2m\sqrt{t/s}(\ln t)^2 \text{ for all } 1 \leq s \leq t\}.$$

It is shown below, see Lemma 11a, that

$$\mathbf{Pr}(G(t) \in \mathcal{G}_2(t)) = 1 - \widetilde{O}(t^{-10}).$$

Therefore, we replace (6) by

$$\begin{aligned}
\mathbf{Pr}(\mathcal{E}_{\mathcal{P}} \mid G(t) \in \mathcal{G}_2(t)) &\leq \sum_{\substack{C_1,\dots,C_k \\ P_v, v \in I}} \prod_{(x,y)\in F^*} \frac{2m(\ln t)^2}{\max\{x,y\}} \sqrt{\frac{\max\{x,y\}}{\min\{x,y\}}} \frac{1}{\mathbf{Pr}(G(t) \in \mathcal{G}_2(t))} \\
&\leq 2 \sum_{\substack{C_1,\dots,C_k \\ P_v, v \in I}} \prod_{(x,y)\in F^*} \frac{2m(\ln t)^2}{x^{1/2} y^{1/2}}. \\
&\leq \left(\frac{2m(\ln t)^2}{t^{1/2}}\right)^{10} \sum_{\ell=1}^{20\sigma_0} \left(\sum_{v=1}^{t} \frac{2m(\ln t)^2}{v}\right)^{\ell} \\
&= \widetilde{O}(t^{-23/5}).
\end{aligned}$$

Thus we have

$$\mathbf{E}\left(\binom{|B_t|}{10}\right) = \widetilde{O}(t^{-10}) + \widetilde{O}(t^{10-23/5}) = \widetilde{O}(t^{27/5})$$

and

$$\mathbf{Pr}(|B_t| \geq t^{7/8}) \leq \frac{\mathbf{E}\left(\binom{|B_t|}{10}\right)}{\binom{t^{7/8}}{10}} = o(t^{-3}).$$

□

3.5 Proof of Lemma 3

3.5.1 Rapidly mixing walks

We now consider the random walk made by the spider S. A *random walk* on an fixed undirected graph G is a Markov chain (X_t), $X_t \in V$ associated to a particle that moves from vertex to vertex according to the following rule: The probability of a transition from vertex v, of degree d, to vertex

w is $1/d$ if v is adjacent to w, and 0 otherwise. Let π denote the steady state distribution of the random walk. The steady state probability $\pi_G(v)$ of the walk being at a vertex v is,

$$\pi_G(v) = \frac{d(v)}{d(G)}, \tag{7}$$

where $d(v)$ is the degree of v and $d(G)$ is the total degree (i.e. sum of the degrees) of the graph G.

We will need a finite time approximation of the probability distribution π_H pertaining to a random walk on a subgraph $H = G(t) - s$ of $G(t)$. We obtain this by considering the *mixing time* of the walk based on a conductance bound (11) of Jerrum and Sinclair [23].

Let s, t be fixed with $s \in \left[\frac{t}{\ln t}, t\right) \setminus B_t$. Let P denote the transition matrix of the random walk on H. Let $P^{i,\tau}$ denote the distribution of the τth step of a random walk on H which starts at vertex i. For $K \subset V(H) = [t] \setminus \{s\}$ let $\overline{K} = V(H) \setminus K$ and

$$\Phi_K = \frac{\sum_{i \in K, j \in \overline{K}} \pi_H(i) P(i,j)}{\pi_H(K)}.$$

It follows from (7) that

$$\Phi_K = \frac{e(K : \overline{K})}{d(K)}$$

where $e(K : \overline{K})$ is the number of edges from K to $\overline{K}$, and $d(K)$ is the total degree of vertices in set K.

The conductance of the walk is defined by

$$\Phi(s,t) = \min_{\pi_H(K) \leq 1/2} \Phi_K.$$

Let

$$\mathcal{G}_3(t) = \left\{ G(t) : \ \Phi(s,t) > \frac{1}{\ln t} \forall s \in [t/\ln t, t] \right\}.$$

Lemma 6. *If $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ then, in both Model 1 and Model 2,*

$$\mathbf{Pr}_{\boldsymbol{\theta}}(G(t) \notin \mathcal{G}_3(t)) = o(t^{-3}).$$

□

3.5.2 Proof of Lemma 6: Model 1

Since $d(K) \leq 2m|K| + e(K : \overline{K})$ it suffices to prove a high probability lower bound on $e(K : \overline{K})$, in both models.

Lemma 7.

$$\mathbf{Pr}_{\boldsymbol{\theta}}\left(\Phi(s,t) \leq \frac{1}{200} \right) = o(t^{-3}).$$

Proof For $K \subseteq [t]$ let $d(K,t) = \sum_{s \in K} d(s,t)$. Then

$$\mathbf{Pr}(\exists K \subseteq [t] : \ |K| \geq 3t/4 \text{ and } d(K,t) \leq (1.1)mt) = o(e^{-cmt}) \tag{8}$$

for some absolute constant $c > 0$.

To see this let $K \subseteq [t]$ with $|K| = k = 3t/4$. Then

$$\mathbf{E}\left(d(K,t)\right) \geq \mathbf{E}\left(d([t-k+1,t],t)\right) = mk + m \sum_{s=t-k+1}^{t} \frac{s-(t-k)}{s}$$
$$\geq 2mk - m(t-k)\ln(t/(t-k)) = \left(\frac{3}{2} - \frac{1}{4}\ln 4\right) mt \geq (1.15)mt.$$

Applying Theorem 1 of Hoeffding we see that

$$\mathbf{Pr}(\exists K \subseteq [t] :\ |K| \geq 3t/4 \text{ and } d(K,t) \leq (1.1)mt) \leq \binom{t}{3t/4} e^{-c'mt}$$

for some absolute constant $c' > 0$. This completes the proof of (8)

Now for $K, L \subset [t] \setminus \{s\}$, let $e(K : L)$ denote the number of edges of $G(t)$ which have one end in K and the other end in L (we only use this definition for $L = \overline{K} = [t] \setminus (K \cup \{s\})$ and $L = K$).

It follows from (8) that with probability $1 - o(t^{-3})$

$$\Phi(s,t) \geq \min_{\pi(K) \leq 1/2} \frac{e(K:\overline{K}) - 5|K|}{m|K| + e(K:\overline{K})} \geq \min_{|K| \leq 3t/4} \frac{e(K:\overline{K}) - 5|K|}{m|K| + e(K:\overline{K})}. \tag{9}$$

($e(K:\overline{K}) - 5|K|$ bounds the number of $K:\overline{K}$ edges in $H_s(t)$ and then observe that the degree sum of K is at most $m|K| + e(K:\overline{K})$.)

We prove the following high probability lower bound on $e(K:\overline{K})$. Together with (9) this proves the lemma.

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K :\ e(K:\overline{K}) \leq m|K|/150) = o(t^{-3}). \tag{10}$$

Suppose $K \subset [t]$, $k = |K|$ and $Y_K = e(K:\overline{K})$. Let $\kappa = \frac{1}{2}\sqrt{kt}$ and $K_- = K \cap [\kappa]$ and $K_+ = K \setminus K_-$.

Case 1: $|K_-| \geq 3k/7$.

$$\mathbf{E}_{\boldsymbol{\theta}}\left(Y_K\right) \geq \sum_{\tau=\kappa}^{t-4k/7-1} \frac{3(m-5)k/7}{\tau + 4k/7} \geq \frac{3(m-5)k}{7} \ln\left(\frac{t-1}{\kappa + 4k/7}\right).$$

Explanation: Consider the $\geq t - \kappa - 4k/7 - 1$ vertices of $[t] - [\kappa] - \{s\} - K$. Each chooses at least $m-5$ random neighbours from lower numbered neighbours (plus themselves) and the sum minimises the expected number of these choices in K_-. The 5 comes from $\theta_{\tau+1} - \theta_t \leq 5$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Applying Theorem 1 of [19] we obtain

$$\mathbf{Pr}_{\boldsymbol{\theta}}(Y_K \leq \mathbf{E}_{\boldsymbol{\theta}}\left(Y_K\right)/2) \leq \exp\left\{-\frac{1}{8}\frac{3mk}{7}\ln\left(\frac{t-1}{\kappa+4k/7}\right)\right\} = \left(\frac{\kappa+4k/7}{t-1}\right)^{3mk/56}.$$

So,

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K :\ |K_-| \geq 3k/7,\ |K| \leq 3t/4 \text{ and } Y_K \leq \mathbf{E}\left(Y_K\right)/2) \leq$$
$$\sum_{k=1}^{3t/4} \binom{t}{k} \left(\frac{\kappa+4k/7}{t-1}\right)^{3mk/56} \leq \sum_{k=1}^{3t/4} \left(\frac{te}{k}\left(\frac{\kappa+4k/7}{t-1}\right)^{3m/56}\right)^k \leq$$
$$\sum_{k=1}^{3t/4} \left(\frac{3t}{k}\left(\sqrt{\frac{k}{t}}\left(\frac{1}{2} + \frac{4}{7}\sqrt{\frac{3}{4}}\right)\right)^{3m/56}\right)^k = o(t^{-3}).$$

This yields (10) for this case.

Case 2: $|K_-| \leq 3k/7$.
Assume first that $k \geq 1000$. Now let Z_K denote the number of edges from the set W of $\lceil k/15 \rceil$ lowest numbered vertices of K_+ which have their lower numbered endpoints also in K. Z_K is dominated by $B(m\lceil k/15\rceil, \sqrt{k/t})$ since there are at most $3k/7 + \lceil k/15 \rceil \leq k/2$ vertices of K below any vertex w of W and there are at least κ vertices in all below such a w. We use $Y_K = e(K : \overline{K}) \geq M - Z_K$ where $M = (m-5)\lceil k/15 \rceil$. For $|K| \leq t/1000$ we write

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K:\ 1000 \leq |K| \leq t/1000,\ Z_K \geq M/2) \leq \sum_{k=1000}^{t/1000} \binom{t}{k} 2^M \left(\frac{k}{t}\right)^{M/2} \leq$$
$$\sum_{k=1000}^{t/1000} \left(\frac{te}{k}\left(\frac{4k}{t}\right)^{(m-5)/30}\right)^k = o(t^{-3}).$$

For $|K| > t/1000$ we use Chernoff bounds and write, for some absolute positive constant $c > 0$

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K:\ t/1000 \leq |K| \leq 3t/4,\ Z_K \geq 9M/10) \leq \sum_{k=t/1000}^{3t/4} \binom{t}{k} e^{-cM} = o(t^{-3}).$$

For $|K| \leq 1000$ we can write

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K:\ e(K,K) \geq 3mk/4) \leq \sum_{k=1}^{1000} \binom{t}{k} \binom{mk}{3mk/28} \left(\frac{1000}{t^{1/2}}\right)^{3mk/28} = o(t^{-3}).$$

Note that if $e(K,K) \geq 3mk/4$ then at least $3mk/4 - 3mk/7$ of these edges must have one end in K_+.

This completes the proof of (10). $\square$

3.5.3 Proof of Lemma 6: Model 2

Lemma 8. *There is an absolute constant $\xi > 0$ such that*

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K \subseteq [t] - \{s\}, |K| \geq (1-\xi)t:\ d(K,t) \leq (1+\xi)mt) = o(t^{-3}).$$

Proof Let ζ be a small positive constant and divide $[t]$ into approximately $1/\zeta$ consecutive intervals $I_1, I_2, \ldots$ of size $\lceil \zeta t \rceil$ plus an interval of $t - \lfloor 1/\zeta \rfloor \lceil \zeta t \rceil$. We put a high probability bound on the total degree $d(I_1, t)$. Now consider the random variables $\beta_k, k = 1, 2, \ldots$ where $\beta_k = d(I_1, k\lceil \zeta t \rceil)/(m\lceil \zeta t \rceil)$. Now $\beta_1 = 2$ and conditional on the value of β_k

$$(\beta_{k+1} - \beta_k)m\lceil \zeta t \rceil \text{ is dominated by } B\left(m\lceil \zeta t \rceil, \frac{\beta_k + 1}{2k}\right)$$

It follows that we can find an absolute constant $c > 0$ such that

$$\mathbf{Pr}_{\boldsymbol{\theta}}\left(\beta_{k+1} \leq \beta_k\left(1 + \frac{3}{4k}\right)\right) \leq e^{-cm\zeta t}.$$

So, with probability $1 - O(e^{-cm\zeta t})$ we find that

$$d(I_1, t) \leq 2m\lceil \zeta t \rceil \prod_{k=1}^{\lceil 1/\zeta \rceil} \left(1 + \frac{3}{4k}\right) \leq 2m\lceil \zeta t \rceil \times e^{3/4}\lceil 1/\zeta \rceil^{3/4} \leq 6m\zeta^{1/4}t,$$

for small enough ζ.

Now $d([\lceil \zeta t \rceil], t)$ dominates $d(L, t)$ for any set L of size $\lceil \zeta t \rceil$. So, if $m > 1/(c\zeta)$ then the probability there is a set of size $\lceil \zeta t \rceil$ which has degree exceeding $6m\zeta^{1/4}$ is exponentially small ($\leq \binom{t}{\lceil \zeta t \rceil} e^{-t}$)). In which case, every set K of size at least $t - \lceil \zeta t \rceil$ has total degree $d(K, t) \geq 2mt - 6m\zeta^{1/4}t$ and the lemma follows by taking ζ sufficiently small. □

Lemma 9. *If m is sufficiently large then*

$$\mathbf{Pr}_{\boldsymbol{\theta}}\left(\Phi(t) < \frac{1}{\ln t}\right) = O(t^{-3}).$$

Proof For $K \subseteq [t]$, $|K| = k$ we say K is *small* if $\ln t \leq k \leq ct$ and K is large otherwise, where $c = e^{-8}$.

3.5.4 Case of K small

Let $K_- = K \cap [\sqrt{kt}]$ and let $K_+ = K \setminus K_-$.

Case of $|K_-| \geq k/2$.
Let $X_t = X_t(K_-)$ be the number of those edges directed into K_- from vertices created after time $\sqrt{kt}$. The number of such edges generated at step $\tau \geq \sqrt{kt}$ dominates $B(m-5, mq/(2m\tau))$, independently of any previous step. (Here $q = |K_-|$). Thus

$$\mathbf{E}\left(X_t\right) \geq \sum_{\tau > \sqrt{kt}}^{t} \frac{(m-5)q}{2\tau} = \frac{(m-5)q}{4} \ln \frac{t}{k}(1+o(1)).$$

Hence

$$\mathbf{Pr}\left(X_t \leq \frac{mq}{6} \ln \frac{t}{k}\right) \leq \exp\left(-\frac{mq}{73} \ln \frac{t}{k}\right).$$

Thus

$$\begin{aligned} \mathbf{Pr}\left(\exists K_- : X_t(K_-) \leq \frac{mq}{6} \ln \frac{t}{k}\right) &\leq \binom{\sqrt{kt}}{q} \exp\left(-\frac{mq}{73} \ln \frac{t}{k}\right) \\ &\leq \exp\left(-q\left(\frac{m}{73} \ln \frac{t}{k} - \ln\left(2e\sqrt{\frac{t}{k}}\right)\right)\right) \\ &\leq t^{-4} \end{aligned}$$

provided m is sufficiently large. Thus **whp** the set K_- has at least $\frac{mk}{12} \ln t/k$ edges directed into it, of which at most $mk/2$ are incident with K_+. This completes the analysis of this case.

Case of $|K_+| \geq k/2$. We consider the evolution of the set $K_+ = \{u_1, u_2, \ldots, u_r\}$ from step $T = \sqrt{kt}$ onwards. Assume that at the final step t there are δk edges directed into K from $\overline{K}$. We can assume w.l.o.g. that $\delta \leq m/10$, for otherwise there is nothing to prove.

The number Y_{j+1} of $K : K$ edges generated by vertex u_{j+1} is a Binomial random variable with expectation at most

$$\mu_{j+1} = m\frac{2mk + \delta k}{2mt_{j+1}}.$$

The numerator in the above fraction is a bound on the total degree of K.

If $Z = Z(K_+) = \sum_{j=1}^{r} Y_j$ then

$$\begin{aligned}
\mathbf{E}(Z) &\leq \frac{2mk+\delta k}{2}\left(\frac{1}{t_1}+\cdots+\frac{1}{t_r}\right) \\
&\leq \frac{2mk+\delta k}{2}\frac{r}{\sqrt{kt}} \\
&\leq 1.05\,\frac{mkr}{\sqrt{kt}}.
\end{aligned}$$

Thus, for $\alpha > 0$,

$$\begin{aligned}
\mathbf{Pr}\left(\exists K_+ :\ Z(K_+) \geq \alpha k\right) &\leq \sum_{r=k/2}^{k}\binom{t}{r}\left(\frac{e\times 1.05\times kmr}{\sqrt{kt}\times\alpha k}\right)^{\alpha k} \\
&\leq k\left(\left(\frac{3mk^{1/2}}{\alpha t^{1/2}}\right)^{\alpha}\frac{te}{k}\right)^{k} \\
&\leq t^{-4}
\end{aligned}$$

if $\alpha = m/4$, $k \leq ct$ and m is sufficiently large. We have therefore proved that for small values of k there are at least $mk/2 - mk/4$ out-edges generated by K_+ not incident with K on the condition that $\delta \leq m/10$, completing the analysis of this case.

3.5.5 Case of K large

Let $T = t/2$ and let $ct \leq |K|, |\overline{K}| \leq (1-\xi)t$ where ξ is as in Lemma 8. Let $M = [T]$ and $N = [T+1, t]$. Let $K_- = K \cap M$, $K_+ = K \cap N$, $q = |K_-|$ and $r = |K_+|$. We calculate the expected number of edges $\mu(K_-, K_+)$ of $L = (K_+ \times (M \setminus K_-)) \cup ((N \setminus K_+) \times K_-)$ generated at steps τ, $T \leq \tau \leq t$ which are directed into K. At step τ the number of such edges falling in L is an independent random variable with distribution dominating

$$1_{\tau\in N\setminus K_+} B\left(m-5, \frac{mq}{2m\tau}\right) + 1_{\tau\in K_+} B\left(m-5, \frac{(T-q)m}{2m\tau}\right).$$

Thus

$$\begin{aligned}
\mu(K_-, K_+) &\geq \frac{(m-5)q}{2}\sum_{\tau\in N\setminus K_+}\frac{1}{\tau} + \frac{(m-5)(T-q)}{2}\sum_{\tau\in K_+}\frac{1}{\tau} \\
&= \frac{m-5}{2}\left((k-r)\sum_{\tau\in N\setminus K_+}\frac{1}{\tau} + (T-(k-r))\sum_{\tau\in K_+}\frac{1}{\tau}\right).
\end{aligned}$$

Let $\mu(k) = \min_{K_-,K_+} \mu(K_-, K_+)$. Then 'somewhat crudely'

$$\begin{aligned}
\sum_{\tau\in N\setminus K_+}\frac{1}{\tau} &\geq \ln\frac{t}{T+r} \\
\sum_{\tau\in K_+}\frac{1}{\tau} &\geq \ln\frac{t}{t-r}.
\end{aligned}$$

Thus

$$\mu(k) \geq \frac{m-5}{2}\left((k-r)\ln\frac{2t}{t+2r} + \left(\frac{t}{2}-(k-r)\right)\ln\frac{t}{t-r}\right).$$

Putting $k = \kappa t$ and $r = \rho t$ we see that

$$\mu(k) \geq \frac{(m-5)t}{2} g(\kappa, \rho)$$

where

$$g(\kappa, \rho) = (\kappa - \rho) \ln \frac{2}{1+2\rho} + \left(\tfrac{1}{2} - \kappa + \rho\right) \ln \frac{1}{1-\rho}.$$

We put a lower bound on g:

$$\rho \leq \frac{\xi}{2} \text{ implies } \kappa - \rho \geq \frac{\xi}{2} \text{ and so } g(\kappa, \rho) \geq \frac{\xi}{2} \ln \frac{2}{1+\xi}.$$

So we can assume that $\rho \geq \xi/2$. Then

$$\begin{array}{rcl} \kappa - \rho \leq \dfrac{1-\xi}{2} & \text{implies} & g(\kappa, \rho) \geq \dfrac{\xi}{2} \ln \dfrac{2}{2-\xi}. \\ \kappa - \rho > \dfrac{1-\xi}{2} \text{ and } \rho \leq \dfrac{1-\xi}{2} & \text{implies} & g(\kappa, \rho) \geq \dfrac{1-\xi}{2} \ln \dfrac{2}{2-\xi}. \\ \kappa - \rho > \dfrac{1-\xi}{2} \text{ and } \rho > \dfrac{1-\xi}{2} & \text{implies} & \kappa > 1 - \xi. \end{array}$$

We deduce that within our range of interest,

$$\mu(k) \geq \eta m t$$

for some absolute constant η.

Let Z be the number of edges generated within L, so that Z counts a subset of the edges between K and $\overline{K}$. Then

$$\begin{aligned} \mathbf{Pr}\left(\exists K_-, K_+ \subseteq N :\ Z \leq \frac{1}{2}\eta m t\right) &\leq 2^t e^{-\eta m t/8} \\ &\leq e^{-\eta m t/10}. \end{aligned}$$

Recall that m is sufficiently large. This completes the proof of the lemma, except for very small sets K.

For sets K of size $s \leq \ln t$ we note that, as $G(t)$ is connected, the conductance Φ_K is always $\Omega(1/|K|)$. □

3.6 Proof of Lemma 3: Continued

Define

$$\mathcal{G}(t) = \begin{cases} \mathcal{G}_1(t) \cap \mathcal{G}_3(t) & \text{Model 1} \\ \mathcal{G}_1(t) \cap \mathcal{G}_2(t) \cap \mathcal{G}_3(t) & \text{Model 2} \end{cases}.$$

We apply the main result of [23].

$$|P^{i,\tau}(v) - \pi_H(v)| \leq \left(1 - \frac{\Phi^2}{2}\right)^{\tau} \frac{\pi_H(v)}{\pi_{\min}} \tag{11}$$

where $\pi_{\min} = \min_v \pi_H(v)$.

Using (11) and Lemma 6 we see that with $\mu_0 = 10(\ln t)^3$, **whp**

$$|P^{i,\mu_0}(v) - \pi_H(v)| = O(t^{-4}) \qquad \forall v \in [t] \setminus \{s\}. \tag{12}$$

We are glossing over one technical point here. Strictly speaking, (11) only holds for Markov chains in which $P(x,x) \geq 1/2$ for all states x. To get round this one usually makes the walk flip a fair coin and stay put if the coin comes up heads. In our case we also omit to add a new vertex if the coin is heads. So what we have been describing is the outcome, ignoring those times when the coin flip is heads.

For the moment, we fix some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and assume that $t/\ln t \leq s < t$.

Now by definition $t_0 = t - 10\mu_0$ and we define

$$\begin{aligned} I &= [t_0+1, t-1] \\ J &= \{\sigma \in I: \exists \tau \in I \text{ such that } X_\tau = \sigma\} \\ \mathcal{E}_0 &= \{X_\tau \neq s,\, \tau \in I\} \\ \mathcal{E}_1 &= \{\exists i, j \in J: X_i = j \text{ and } j \text{ has a neighbour in } \{X_\sigma:\ \sigma \in [t_0, i-2]\}\} \\ \mathcal{F}_k &= \{|J| = k\} \qquad k \geq 0 \\ \mathcal{F}_{\geq k} &= \{|J| \geq k\} \qquad k \geq 0 \end{aligned}$$

and write

$$\begin{aligned} &\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid \mathcal{A}_s(t-1)) = \\ &\sum_{\substack{G \in \mathcal{G}(t_0) \\ w \in [t_0] \setminus \{s\}}} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0, \mathcal{A}_s(t_0)) \mathbf{Pr}_{\boldsymbol{\theta}}(X_{t_0} = w, G(t_0) = G \mid \mathcal{A}_s(t-1)) + \\ &\qquad \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s, G(t_0) \notin \mathcal{G}(t_0) \mid \mathcal{A}_s(t-1)). \end{aligned} \tag{13}$$

It follows from Lemma 6 that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s, G(t_0) \notin \mathcal{G}(t_0) \mid \mathcal{A}_s(t-1)) = o(t^{-3} \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1}). \tag{14}$$

To deal with the rest of (13) we write

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0, \mathcal{A}_s(t_0)) &= \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) \\ = \sum_{k=0}^{1} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_k) &\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_k \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) \\ + \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_{\geq 2}) &\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_{\geq 2} \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0). \end{aligned} \tag{15}$$

Given $w, G(t_0)$, conditioning on $\mathcal{B} \stackrel{def}{=} \mathcal{E}_0 \cap \mathcal{F}_0$ is "almost" equivalent to S doing a random walk on $G(t_0) - \{s\}$ starting at w. In fact we get

Lemma 10.

$$\begin{aligned} &\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{B}) = \\ &\qquad \frac{\theta_{t_0}}{2mt_0} \left(1 + \frac{1}{\theta_{t_0}} \sum_{y \in N(s,t_0)} \mathbf{E}_{\boldsymbol{\theta}} \left(\frac{d(y,t_0) - d(y,t)}{d(y,t)}\right)\right) \left(1 + O\left(\frac{1}{m}\right)\right). \end{aligned} \tag{16}$$

where $N(s,t_0)$ denotes the set of neighbours of s in $G(t_0)$.

Proof We emphasize that throughout the proof of this lemma, a graph $G \in \mathcal{G}(t_0)$ is fixed as well as $X_{t_0} = w$. All probabilities are conditional on this, even if not stated explicitly. The only randomness in the graph $G(t)$ itself is due to new vertices.

Let

$$\mathcal{M} = \{\nexists v \in [t_0], v \neq s: \quad v \text{ has more than five neighbours in } I\}.$$

Then in both models

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\overline{\mathcal{M}} \mid G(t_0) = G) \leq |I|^6 \left(\frac{2mt_0^{1/2}(\ln t_0)^2}{mt_0} \right)^6 = \widetilde{O}(t^{-3}). \tag{17}$$

Fix $y \in N(s, t_0)$ and let $\mathcal{W}_k(y)$ denote the set of walks in $H = G(t_0) - s$ which start at w, finish at y, are of length $\lambda_0 = t - t_0 = 100(\ln t)^3$ and which *leave* N^* exactly k times where N^* is the (random) set of neighbours of $I \cup \{s\}$ in $G(t_0)$. Let $\mathcal{W}_k = \bigcup_y \mathcal{W}_k(y)$ and let $W = (w_0, w_1, \ldots, w_{\lambda_0}) \in \mathcal{W}_k(y)$. Let

$$\rho_W = \frac{\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(i) = w_i, i = 0, 1, \ldots, \lambda_0 \mid \mathcal{M})}{\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(i) = w_i, i = 0, 1, \ldots, \lambda_0)}. \tag{18}$$

Here $X_G(i), i = 0, 1, \ldots, \lambda_0$ is the sequence of vertices visited by S at times $t_0, t_0 + 1, \ldots, t$ and we will use $W_G = W_{w,G}$ to denote this walk. We let $X_H(i), i = 0, 1, \ldots, \lambda_0$ is the set of vertices of H visited by a random walk $W_H = W_{w,H}$ on H with start vertex w.

Then

$$1 \geq \rho_W \geq \left(\frac{m-5}{m} \right)^k.$$

This is because a vertex can have at most 5 edges joining it to s and then

$$\frac{\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(i) = w_i \mid X_G(i-1) = w_{i-1}, \mathcal{M})}{\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(i) = w_i \mid X_H(i-1) = w_{i-1})} \begin{cases} \geq \frac{d_G(w_{i-1})-5}{d_G(w_{i-1})} & w_{i-1} \in N^*. \\ = 1 & w_{i-1} \notin N^*. \end{cases}$$

Furthermore,

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B} \mid \mathcal{M}) &= \sum_{k \geq 0} \sum_{W \in \mathcal{W}_k} \mathbf{Pr}_{\boldsymbol{\theta}}(W_{w,G}(\lambda_0) = W \mid \mathcal{M}) \\ &= \sum_{k \geq 0} \sum_{W \in \mathcal{W}_k} \rho_W \mathbf{Pr}_{\boldsymbol{\theta}}(W_{w,H}(\lambda_0) = W) \\ &\geq \sum_{k \geq 0} p_k \left(\frac{m-5}{m} \right)^k \end{aligned}$$

where

$$p_k = \sum_{W \in \mathcal{W}_k} \mathbf{Pr}_{\boldsymbol{\theta}}(W_H(\lambda_0) = W) = \mathbf{Pr}_{\boldsymbol{\theta}}(W_H(\lambda_0) \in \mathcal{W}_k).$$

We will show later that

$$p_0 + p_1 + p_2 \geq 1 - O(m^{-1}) \tag{19}$$

which immediately implies that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B} \mid \mathcal{M}) \geq p_0 + p_1 \left(1 - \frac{5}{m} \right) + p_2 \left(1 - \frac{5}{m} \right)^2 \geq 1 - O(m^{-1}).$$

Now write

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y \mid \mathcal{B}, \mathcal{M}) &= \sum_{k \geq 0} \sum_{W \in \mathcal{W}_k(y)} \mathbf{Pr}_{\boldsymbol{\theta}}(W_G(\lambda_0) = W \mid \mathcal{M}) \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B} \mid \mathcal{M})^{-1} \\ &= \sum_{k \geq 0} \sum_{W \in \mathcal{W}_k(y)} \rho_W \mathbf{Pr}_{\boldsymbol{\theta}}(W_H(\lambda_0) = W) \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B} \mid \mathcal{M})^{-1}. \end{aligned}$$

Now if

$$\begin{aligned} p_{k,y} &= \frac{\mathbf{Pr}_{\boldsymbol{\theta}}(W_H \in \mathcal{W}_k(y))}{\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)} \\ &= \mathbf{Pr}_{\boldsymbol{\theta}}(W_H(\lambda_0) \text{ leaves } N^* \text{ exactly } k \text{ times} \mid X_H(\lambda_0) = y) \end{aligned}$$

then we have

$$\sum_{k \geq 0} p_{k,y} \left(\frac{m-5}{m}\right)^k \leq \frac{\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y \mid \mathcal{B}, \mathcal{M})}{\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)} \leq \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B} \mid \mathcal{M})^{-1}. \tag{20}$$

We need to be careful about probability spaces here. Let $N^{\#}$ denote the the set of neighbours of I in $G(t_0)$. In our conditional probability space, the $p_{k,y}$ are now to be thought of as random variables dependent on $N^{\#}$.

We will show later that

$$\mathbf{Pr}_{\boldsymbol{\theta},\#}(p_{0,y} + p_{1,y} + p_{2,y} \geq 1 - O(m^{-1})) = 1 - \widetilde{O}(t^{-1/2}). \tag{21}$$

where $\mathbf{Pr}_{\boldsymbol{\theta},\#}$ stresses that $N^{\#}$ is randomly chosen.

So, from (20), we obtain

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y) \left(1 - \frac{5}{m}\right)^2 \mathbf{Pr}_{\boldsymbol{\theta},\#}(p_{0,y} + p_{1,y} + p_{2,y} \geq 1 - O(m^{-1})) \leq \\ \mathbf{Pr}_{\boldsymbol{\theta},\#}(X_G(\lambda_0) = y \mid \mathcal{B}, \mathcal{M}) \leq \\ \mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)(1 + O(m^{-1})) \end{aligned}$$

or using (17)

$$\left| \frac{\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y \mid \mathcal{B}, \mathcal{M})}{\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)} - 1 \right| = O\left(\frac{1}{m}\right).$$

Therefore

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y \mid \mathcal{B}, \mathcal{M}) &= (1 + O(m^{-1}))\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y) \\ \mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y, \mathcal{B}, \mathcal{M}) &= (1 + O(m^{-1}))\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B}, \mathcal{M}) \end{aligned}$$

or

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y, \mathcal{B}) + \tilde{O}(t^{-3}) = (1 + O(m^{-1}))\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y)(\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B}) + \tilde{O}(t^{-3})) \tag{22}$$

Now we will show later that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{B}) \geq \frac{1}{2} \tag{23}$$

and (11) implies

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_H(\lambda_0) = y) = \frac{d(y, t_0) - \delta_y}{2mt_0} + O(t^{-3}) = \Omega(t^{-1})$$

where $1 \leq \delta_y \leq 5$ is the number of (s, y) edges in $G(t_0)$. So from (22) we have

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_G(\lambda_0) = y \mid X_{t_0} = w, G(t_0) = G, \mathcal{B}) = (1 + O(m^{-1}))\frac{d(y, t_0)}{2mt_0}.$$

Thus,

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{B}) = (1 + O(m^{-1}))\mathbf{E}_{\boldsymbol{\theta}} \left(\sum_{y \in N(s,t_0)} \frac{d(y, t_0)}{2mt_0} \frac{1}{d(y, t)} \right. \\ \left. + O(m^{-1})\mathbf{Pr}_{\boldsymbol{\theta}}(X_{t-1} \in N(s, t-1) \setminus N(s, t_0)) \right). \end{aligned}$$

And then Lemma 10 follows from

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_{t-1} \in N(s, t-1) \setminus N(s, t_0)) = \widetilde{O}(t^{-2}), \tag{24}$$

which will be proved below.

3.6.1 Proof of (19,21)

Clearly (21) implies (19). We therefore start with (21). Observe first that $N^{\#}$ the set of neighbours of I in $G(t_0)$ satisfies

$$\mathbf{Pr}(W_H(\lambda_0) \cap N^{\#} \neq \emptyset) = \widetilde{O}(t^{-1/2}). \tag{25}$$

To see this we use the fact that the walk is defined on H and $N^{\#}$ is independent of H. We also need an upper bound of $\widetilde{O}(t^{1/2})$ for maximum degree in $G(t_0)$ and this follows from Lemmas 4(a) and 11(a).

Since $N^* = N^{\#} \cup N(s, t_0)$ we only have to show now that the probability that W_H leaves a vertex of $N(s, t_0)$ three times or more is $O(m^{-1})$. Furthermore, this event depends on $G(t_0)$, which is fixed and not on $G(t)$.

Given $G \in \mathcal{G}(t_0)$ and $W = W_H(\lambda_0)$, the total degree of the vertices of W is $\widetilde{O}(t^{1/2})$ in Model 2. In Model 1 we would have $\widetilde{O}(t^{-1})$ for the RHS of (25).

Now let $\mathcal{W}(a, b, \gamma)$ denote the set of walks in H from a to b of length γ and for $W \in \mathcal{W}(a, b, \gamma)$ let $\mathbf{Pr}(W) = \mathbf{Pr}(W_{a,H}(\gamma) = W)$. Then for $x \in V(H)$ we have

$$\begin{aligned}
\mathbf{Pr}(X_{w,H}(\lambda_0/2) = x \mid X_{w,H}(\lambda_0) = y) &= \sum_{\substack{W_1 \in \mathcal{W}(w,x,\lambda_0/2) \\ W_2 \in \mathcal{W}(x,y,\lambda_0/2)}} \frac{\mathbf{Pr}(W_1)\mathbf{Pr}(W_2)}{\mathbf{Pr}(\mathcal{W}(w, y, \lambda_0))} \\
&= \pi_{x,H}^{-1} \sum_{\substack{W_1 \in \mathcal{W}(w,x,\lambda_0/2) \\ W_2 \in \mathcal{W}(x,y,\lambda_0/2)}} \frac{\mathbf{Pr}(W_1)\pi_{x,H}\mathbf{Pr}(W_2)}{\mathbf{Pr}(\mathcal{W}(w, y, \lambda_0))}
\end{aligned}$$

and with W_3 equal to the reversal of W_2,

$$\begin{aligned}
&= \pi_{x,H}^{-1}\pi_{y,H} \sum_{\substack{W_1 \in \mathcal{W}(w,x,\lambda_0/2) \\ W_3 \in \mathcal{W}(y,x,\lambda_0/2)}} \frac{\mathbf{Pr}(W_1)\mathbf{Pr}(W_3)}{\mathbf{Pr}(\mathcal{W}(w, y, \lambda_0))} \\
&= \frac{\pi_{x,H}^{-1}\pi_{y,H}}{\mathbf{Pr}(\mathcal{W}(w, y, \lambda_0))} \mathbf{Pr}(\mathcal{W}(w, x, \lambda_0/2))\mathbf{Pr}(\mathcal{W}(y, x, \lambda_0/2)) \\
&= \frac{\pi_{x,H}^{-1}\pi_{y,H}}{\mathbf{Pr}(\mathcal{W}(w, y, \lambda_0))} (\pi_{x,H} - O(t^{-24}))^2 \\
&= \pi_{x,H} - O(t^{-23}).
\end{aligned}$$

It follows that the variation distance between the distribution of a random walk of length λ_0 from w to y and that of $W_1, W_3^{reversed}$ is $O(t^{-22})$ where W_1, W_3 are obtained by (i) choosing x from the steady state distribution and then (ii) choosing a random walk W_1 from w to x and a random walk W_3 from y to x. Furthermore, the variation distance between distribution of W_1 and a random walk of length $\lambda_0/2$ from w is $O(t^{-24})$. Similarly, the variation distance between distribution of W_3 and a random walk of length λ_0 from y is $O(t^{-24})$.

Now consider W_1 and let Z_i be the distance of $X_H(i)$ from s. We observe that since $s \notin B_{t_0} \subseteq B_t$, while the walk is within σ_0 of s the distance to s must go up or down in one step and that

$$\mathbf{Pr}\left(Z_{i+1} = Z_i + 1 \mid Z_i < \sigma_0\right) \geq 1 - \frac{1}{m-5}.$$

We will deduce from this that, where $N_H(s)$ is the set of $G(t_0)$ neighbours of s,

$$\mathbf{Pr}(W_1 \text{ or } W_3 \text{ make a } \textit{return} \text{ to } N_H(s)) = O(1/m) \tag{26}$$

and this together with (25) implies (21).

To verify (26) we first see that

$$\begin{aligned} \mathbf{Pr}_{\boldsymbol{\theta}}(\exists 1 \le i \le \lambda_0/2 : Z_i = 1 \mid Z_0 = \sigma_0) &\le \lambda_0 \sum_{k=0}^{\lambda_0/2} \binom{\sigma_0+2k}{\sigma_0+k} \left(\frac{1}{m-5}\right)^{\sigma_0+k} \\ &\le \lambda_0 \sum_{k=0}^{\lambda_0/2} \left(\frac{(\sigma_0+2k)e}{(m-5)(\sigma_0+k)}\right)^{\sigma_0+k} \\ &\le \lambda_0^2 (2e/(m-5))^{\sigma_0} \end{aligned} \tag{27}$$

Then we have

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists i > 0 : \ Z_{2i} = 1, 1 < Z_1, Z_2, \dots, Z_{2i-1} \le \sigma_0 \mid Z_0 = 1) \le \sum_{i>0} \binom{2i}{i} \left(\frac{1}{m-5}\right)^i < \sum_{i>0} \left(\frac{2}{m-5}\right)^i = \frac{2}{m-7}. \tag{28}$$

Equation (26) follows from (27) and (28).

3.6.2 Proof of (23)

This follows from $\mathbf{Pr}(\mathcal{E}_0 \mid X_{t_0} = w, G(t_0) = G, \mathcal{F}_0) = 1 - O(m^{-1})$, and this follows much as in the proof of (21). In particular, we see that if a walk in G starts at $w \neq s \notin B_t$ then the probability it visits s in λ_0 steps is $O(m^{-1})$. Then we will see that $\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_0 \mid X_{t_0} = w, G(t_0) = G) = 1 - o(1)$ (see (29) below). In the proof below we condition on $\mathcal{E}_0$ but the proof is valid without this conditioning.

This completes the proof of Lemma 10. $\square$

We will next argue that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_{\ge k} \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) = \widetilde{O}(t^{-k}) \qquad k = 1, 2. \tag{29}$$

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_1) = \widetilde{O}(t^{-1}). \tag{30}$$

$$\mathbf{E}_{\boldsymbol{\theta}}\left(d(y,t) - d(y,t_0)\right) = \widetilde{O}(t^{-1/2}). \tag{31}$$

It follows from (13)–(16) and (29)–(31) that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) = \frac{\theta_{t_0}}{2mt}\left(1 + O\left(\frac{1}{m}\right)\right) + O(t^{-3}\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1}). \tag{32}$$

and removing the conditioning on $X_{t_0} = w, G(t_0) = G$ yields Lemma 3a.

For part (b) we see that $X_s = s$ if and only if [i] s chooses X_{s-1} as one of its m neighbours and then [ii] S moves to s. If we condition on $X_{s-1} = x$ and $d(x, s-1) = d$ then $\mathbf{Pr}_{\boldsymbol{\theta}}([i]) \le \frac{md}{2m(s-1)}$ and $\mathbf{Pr}_{\boldsymbol{\theta}}([ii] \mid [i]) \le \frac{2}{d}$ (we write $\le \frac{2}{d}$ in place of the more natural $\frac{1}{d+1}$ to account for x being chosen more than once). This proves part (b).

3.6.3 Proof of (29)

Let us generate $X_i, i \in I$ using as little information about the edges incident with I as possible. Thus, at step i we first establish whether any of $t_0+1,\ldots,i$ are neighbours of X_{i-1}. If the answer is NO, we do not determine these neighbours. Thus up to the first time we get the answer YES, the conditional distribution of the neighbours of $t_0, t_0+1,\ldots,i$ is that they are chosen from a set of size $t-o(t)$ either randomly (Model 1) or from the same set with probabilities proportional to degree (Model 2). Let $\mathcal{Y}_i = \{\text{YES at } i \text{ and } X_i \in \{t_0+1,\ldots,i\}\}$. If $d(X_{i-1},i)=d$ then

$$\mathbf{Pr}(\mathcal{Y}_i \mid d) = O\left(|I| \cdot \frac{d}{t} \cdot \frac{1}{d}\right) = O\left(\frac{|I|}{t}\right). \tag{33}$$

Since $\mathcal{F}_{\geq 1} \subseteq \bigcup_{i \in I} \mathcal{Y}_i$ we have (29) for $k=1$.

Now assume that i_1 is the first i for which $\mathcal{Y}_i$ occurs and that $X_{i_1-1}=j_1$. Arguing as in the first paragraph of this subsection, we see that the conditional probability that $\mathcal{Y}_i$ occurs for $i_2 > i_1$, with $X_{i_2-1}=j_2 \neq j_1$ is also $\widetilde{O}(t^{-1}|I|)$ and this completes the proof of (29).

3.6.4 Proof of (30)

Let $J=\{j_1\}$ and let j_1 be visited for the first at time t_1. If $t_1 \leq t_0 + 5\times 10^5(\ln t)$ then we can view the walk from time t_1 onwards as a walk of length $\geq 5 \times 10^5(\ln t)$ on the graph $H' = H + j_1$. Using (12) for H' we can argue, as in the proof of (16) that the conditional probability $X_t = s$ is $\widetilde{O}(t^{-1})$ as required.

Suppose next that $t_1 > t_0 + 5\times 10^5(\ln t)$.

We write

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \mathcal{F}_1) = \\
&\quad \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \overline{\mathcal{E}}_1, \mathcal{F}_1)\mathbf{Pr}(\overline{\mathcal{E}}_1 \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \mathcal{F}_1)+ \\
&\quad \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \mathcal{E}_1, \mathcal{F}_1)\mathbf{Pr}(\mathcal{E}_1 \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \mathcal{F}_1).
\end{aligned}$$

Observe that $\mathbf{Pr}(\mathcal{E}_1 \mid X_{t_0}=w, G(t_0)=G, \mathcal{E}_0, \mathcal{F}_1) = \widetilde{O}(t^{-3/2})$. Use (33) plus an extra $\widetilde{O}(t^{-1/2})$ factor for the extra neighbour(s).

So now assume that $\mathcal{E}_1$ does not occur and let $k_1 = X_{t_1+1}$. Suppose first that $k_1 \neq= X_{t_1-1}$ so that k_1 is chosen from a set of size $t-o(t)$. Model 1: If $v \in [t_0] \setminus \{s\}$ then its steady state random walk probability $\pi(v)$ is at least $1/(2t_0)$ and the probability that $k_1 = v$ is at most twice this. It follows that in any subsequent step of a simple random walk, the probability S is at v is at most $2\pi(v)$. At this point we are conditioned on $\mathcal{E}_0$ and $\mathcal{F}_1$ has no further effect. Thus we are essentially conditioned on $\mathcal{B}$ which has probability at least $1/2$. Thus the probability S ever returns to j_1 is $\widetilde{O}(t^{-1}|I|)$. Failing this, by considering the vertices visited after the last visit to l_1 we deduce that the probability we arrive at a neighbour of s, at time $t-1$ is also $\widetilde{O}(t^{-1})$ and (30) follows.

Model 2: Now if $v \in [t_0] \setminus \{s\}$ its steady state random walk probability $\pi(v)$ is asymptotically equal to the probability it is chosen as k_1 and we can use the analysis for Model 1.

Now let t_2 be the last time that j_1 is visited before time t We first deal with the possibility that $t_2 = t-1$ and $j_1 \in N(s,t-1)$. We first argue that

$$\begin{aligned}
&\mathbf{Pr}(I \cap N(s,t-1) \neq \emptyset \mid |J| = 1, X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) \leq \\
&\quad \mathbf{Pr}(I \cap N(s,t-1) \neq \emptyset \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0) \times \mathbf{Pr}(|J| = 1 \mid X_{t_0} = w, G(t_0) = G, \mathcal{E}_0)^{-1} \\
&\qquad = \widetilde{O}(t^{-1}). \quad (34)
\end{aligned}$$

This is because

Now let t_2 be the last time that j_1 is visited before time t Let k_1 be the unique neighbour of j_1 on our walk. If k_1 is never visited, or if each visit to k_1 is the middle of the sequence of visits j_1, k_1, j_1 then S's walk is essentially a random walk on H and we can argue as in (16). If there is a visit to k_1 at time t_1 say, and $X_{t_1+1} = l_1 \neq k_1$ then l_1 will have been chosen from a set of size $t - o(t)$.

3.6.5 Proof of (31)

This follows from the fact that in Model 2, the maximum degree in $G(t)$ is $O(t^{1/2})$ **whp**, see e.g. [16]. For Model 1 the maximum degree is $O(\ln t)$ with sufficiently high probability.

3.6.6 Proof of (24)

We first observe that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(|N(s,t-1) \setminus N(s,t_0)| \geq 2) = \widetilde{O}(t^{-2})$$

and so we only need to consider the case $N(s,t-1) \setminus N(s,t_0) = \{y_1\}$ where $y_1 \in I$.

If $y_1 - t_0 \leq 5\mu_0$ then we can prove $\mathbf{Pr}_{\boldsymbol{\theta}}(X_{t-1} = y_1 \mid X_{y_1} = w', \ldots) = \widetilde{O}(t^{-1})$ essentially by replacing t_0 by y_1. If $y_1 - t_0 > 5\mu_0$ then either (i) there exists $u \in I$ such that $X_u = Y_1$ or (ii) we can coindition on $X_u \neq y_1, u \in I$ and proceed as for (16). Finally note that the probability of (i) is $\widetilde{O}(t^{-1})$ by the argument for (16).

3.7 $\ell \geq 1$

We follow the above analysis and note that the degrees do not change during the spider's walk and that error estimates do not increase (no new vertices are added).

4 Proof of Theorem 2

4.1 Model 1

Theorem 3.

$$\mathbf{E}\,\eta_{\ell,m} = (1 + O(m^{-1})) \int_0^1 \exp\left((m + \tfrac{1}{2}) \ln x + \frac{2m^2}{\ell}\left(1 - x^{\frac{\ell}{2m}}\right)\right)\, dx.$$

$$\eta_\ell = \sqrt{\frac{2}{\ell}} e^{(\ell+2)^2/(4\ell)} \int_{(\ell+2)/\sqrt{2\ell}}^{\infty} e^{-z^2/2}\, dz.$$

Thus, when $\ell = 1$, $\eta_1 = 2\sqrt{\pi}e^{9/4}(1 - \Phi(3/\sqrt{2}))$ *where* $\Phi(x)$ *is the standard normal cumulate. Thus* $\eta_1 \sim 0.5717....$ *Furthermore, as* $\ell \to \infty$, $\eta_\ell \sim 2/\ell$.

Proof We write $d(s,t)$ as

$$d(s,t) = X_s + X_{s+1} + \cdots + X_\tau + \cdots + X_t,$$

where $X_s = m$ and for $\tau > s$, the $X_\tau = B(m, \frac{1}{\tau-1})$ are independent.

Now

$$\begin{aligned} \sum_{\tau=s}^{t} \frac{d(s,\tau)}{\tau} &= \sum_{\tau=s}^{t} \frac{1}{\tau} \sum_{r=s}^{\tau} X_r \\ &= \sum_{r=s}^{t} X_r \left(\sum_{\tau=r}^{t} \frac{1}{\tau} \right). \end{aligned}$$

$$\sum_{\tau=r}^{t} \frac{1}{\tau} = \ln \frac{t}{r} + O\left(\frac{1}{r}\right).$$

Thus

$$\begin{aligned} \prod_{\tau=s}^{t} \left(1 - \frac{d(s,\tau)}{2m\tau}\left(1 + O\left(\frac{1}{m}\right)\right)\right)^{\ell} &= \exp\left(-\left(1 + O\left(\frac{1}{m}\right)\right) \ell \sum_{\tau=s}^{t} \frac{d(s,\tau)}{2m\tau}\right) \qquad (35) \\ &= \exp\left(-\left(1 + O\left(\frac{1}{m}\right)\right) \frac{\ell}{2m} \sum_{r=s}^{t} X_r \ln t/r\right) \\ &= \prod_{r=s}^{t} \left(\frac{r}{t}\right)^{\frac{\ell X_r}{2m}(1+O(\frac{1}{m}))}. \end{aligned}$$

Then we can write

$$\prod_{r=s}^{t} \left(\frac{r}{t}\right)^{\frac{\lambda_1 X_r}{2m}} \le \prod_{\tau=s}^{t} \left(1 - \frac{d(s,\tau)}{2m\tau}\left(1 + O\left(\frac{1}{m}\right)\right)\right)^{\ell} \le \prod_{r=s}^{t} \left(\frac{r}{t}\right)^{\frac{\lambda_2 X_r}{2m}}$$

where $\ell - \frac{A}{m} \le \lambda_1 \le \ell \le \lambda_2 \le \ell + \frac{A}{m}$ for some constant $A > 0$.

Now

$$\begin{aligned} \mathbf{E} \prod_{r=s}^{t} \left(\frac{r}{t}\right)^{\frac{\lambda_1 X_r}{2m}} &= \prod_{r=s}^{t} \mathbf{E} \left(\frac{r}{t}\right)^{\frac{\lambda_1 X_r}{2m}} \\ &= \left(\frac{s}{t}\right)^{\frac{\lambda_1}{2}} \prod_{r=s+1}^{t} \left(1 - \frac{1}{r-1} + \frac{1}{r-1}\left(\frac{r}{t}\right)^{\frac{\ell}{2m}}\right)^{m} \\ &= (1 + o(1)) \left(\frac{s}{t}\right)^{\frac{\lambda_1}{2}} \exp\left\{ m \ln \frac{s}{t} + \frac{2m^2}{\lambda_1}\left(1 - \left(\frac{s}{t}\right)^{\frac{\lambda_1}{2m}}\right)\right\}. \end{aligned}$$

Thus

$$\begin{aligned} \nu_{\ell,m}(t) &\ge (1 + o(1))t \int_0^1 \exp\left(\left(m + \frac{\lambda_1}{2}\right)\ln x + \frac{2m^2}{\lambda_1}\left(1 - x^{\frac{\lambda_1}{2m}}\right)\right)\, dx \\ &= \left(1 + O\left(\frac{1}{m}\right)\right) t \int_0^1 \exp\left(\left(m + \frac{\ell}{2}\right)\ln x + \frac{2m^2}{\ell}\left(1 - x^{\frac{\ell}{2m}}\right)\right)\, dx. \end{aligned}$$

Replacing λ_1 by λ_2 to get an upper bound, we deduce that

$$\nu_{\ell,m}(t) = \left(1 + O\left(\frac{1}{m}\right)\right) t \int_0^1 \exp\left(\left(m + \frac{\ell}{2}\right) \ln x + \frac{2m^2}{\ell}\left(1 - x^{\frac{\ell}{2m}}\right)\right) dx.$$

The values of this integral are easily tabulated. For $\ell = 1$ they quickly reach a value of about 0.57 as m grows. The approximation is accurate to the second decimal place for $m \geq 4$.

As $m \to \infty$, by using the transformations $x = e^{-y}$ and $z = \sqrt{\ell/2}y + (l+2)/(\sqrt{2\ell})$ we obtain

$$\begin{aligned}
\eta_\ell &= \int_0^\infty \exp\left(-\frac{\ell+2}{2}y - \frac{\ell}{4}y^2\right) dy \\
&= \sqrt{\frac{2}{\ell}} e^{(\ell+2)^2/(4\ell^2)} \int_{(\ell+2)/\sqrt{2\ell}}^\infty e^{-z^2/2} \, dz \\
&\sim \frac{2}{\ell}
\end{aligned}$$

since $e^{x^2/2}\int_x^\infty e^{-y^2/2}dy \sim 1/x$ as $x \to \infty$.

4.2 Model 2

Theorem 4.

$$\eta_\ell = e^\ell 2\ell^2 \int_\ell^\infty y^{-3} e^{-y} \, dy.$$

When $\ell = 1$, $\eta_1 = 0.59634....$ Furthermore, as $\ell \to \infty$, $\eta_\ell \sim 2/\ell$.

Lemma 11.

(a)

$$\mathbf{Pr}(\exists s,t: \ d(s,t) \geq 200m\sqrt{t/s}(\ln t)^2) = \widetilde{O}(t^{-10}).$$

(b) *If $t/\ln t \leq s \leq t$ and $r \leq 2m\sqrt{t/s}(\ln t)^2)$ then*

$$\mathbf{Pr}(d(s,t) = m + r) = \binom{m+r-1}{r}\left(\frac{s}{t}\right)^{m/2}\left(1 - \left(\frac{s}{t}\right)^{\frac{1}{2}}\right)^r \left(1 + O\left(\tfrac{(m+r)^3}{s}\right) + O\left(\tfrac{r}{\sqrt{s}}\right)\right).$$

Proof (a) Fix $s \leq t$ and let $X_\tau = d(s,\tau)$ for $\tau = s, s+1, \ldots, t$ and let $\lambda = \frac{(s/t)^{1/2}}{10\ln t}$. Now conditional on $X_\tau = x$, we have

$$X_{\tau+1} = X_\tau + B\left(m, \frac{x}{2m\tau}\right)$$

and so

$$\begin{aligned}
\mathbf{E}\left(e^{\lambda X_{\tau+1}} \mid X_\tau = x\right) &= e^{\lambda x}\left(1 - \frac{x}{2m\tau} + \frac{x}{2m\tau}e^\lambda\right)^m \\
&\leq \exp\left\{\lambda x - \frac{x}{2\tau} + \frac{x}{2\tau}(1 + \lambda + \lambda^2)\right\} \\
&= \exp\left\{\lambda x\left(1 + \frac{1+\lambda}{2\tau}\right)\right\}.
\end{aligned}$$

Thus

$$\mathbf{E}\left(e^{\lambda X_{\tau+1}}\right) \leq \mathbf{E}\left(\exp\left\{X_\tau \lambda\left(1 + \frac{1+\lambda}{2\tau}\right)\right\}\right).$$

If we put $\lambda_t = \lambda$ and $\lambda_{\tau-1} = \lambda_\tau\left(1 + \frac{1+\lambda_\tau}{2\tau}\right)$ then provided $\lambda_s \leq 1$ we will have

$$\mathbf{E}\left(e^{\lambda X_t}\right) \leq e^{m\lambda_s}.$$

Now provided $\lambda_\tau \leq \Lambda = \frac{1}{\ln t}$ we can write

$$\lambda_{\tau-1} \leq \lambda_\tau\left(1 + \frac{1+\Lambda}{2\tau}\right)$$

and then

$$\begin{aligned}\lambda_s &\leq \lambda\prod_{\tau=s}^{t}\left(1 + \frac{1+\Lambda}{2\tau}\right)\\ &\leq 10\lambda(t/s)^{1/2}\end{aligned}$$

which is $\leq \Lambda$ by the definition of λ.

Putting $u = 200m(t/s)^{1/2}(\ln t)^2$ we get

$$\begin{aligned}\mathbf{Pr}(X_t \geq u) &\leq e^{m\lambda_s - \lambda u}\\ &\leq \exp\{\lambda(10m(t/s)^{1/2} - u\}\\ &\leq t^{-19)}\end{aligned}$$

and part (a) follows.

(b) Let $\boldsymbol{\tau} = (\tau_1, ..., \tau_r)$ where τ_j is the step at which the transition from degree $m+j$ to degree $m+j+1$ occurs. Let $\tau_0 = s$ and let $\tau_{r+1} = t$. Let $p(s,t,r:\boldsymbol{\tau}) = \mathbf{Pr}(d(s,t) = m+r$ and $\boldsymbol{\tau})$. Then

$$p(s,t,r:\boldsymbol{\tau}) = \prod_{j=0}^{r}\left(\Phi_j(\tau_j)\prod_{\tau_j<T<\tau_{j+1}}\left(1 - \frac{m+j}{2mT}\right)^m\right),$$

where $\Phi_0 = 1$ and

$$\Phi_j = \left(1 + O\left(\tfrac{m+j}{\tau_j}\right)\right)\frac{m(m+j-1)}{2m\tau_j}\left(1 - \frac{m+j-1}{2m\tau_j}\right)^{m-1}.$$

In the above and in the following we use the fact that $\tau_j \geq s \gg (m+r)^3$ and $r = o(s^{1/2})$.

Now

$$\begin{aligned}\prod_{\tau_j<T<\tau_{j+1}}\left(1 - \frac{m+j}{2mT}\right)^m &= \exp\left(-\frac{m+j}{2}\sum_{\tau_j<T<\tau_{j+1}}\left(\frac{1}{T} + O\left(\frac{m+j}{T^2}\right)\right)\right)\\ &= \exp\left(-\frac{m+j}{2}\left(\log\frac{t_{j+1}}{\tau_j} + O\left(\frac{m+j}{\tau_j}\right)\right)\right)\\ &= \left(\frac{\tau_j}{\tau_{j+1}}\right)^{\frac{m+j}{2}}\left(1 + O\left(\tfrac{(m+j)^2}{\tau_j}\right)\right).\end{aligned}$$

Thus

$$\mathbf{Pr}(d(s,t) = m+r) = \sum_{\boldsymbol{\tau}} p(s,t,r:\boldsymbol{\tau})$$

where

$$p(s,t,r:\boldsymbol{\tau}) = \left(1 + O\left(\tfrac{(m+r)^3}{s}\right)\right)\frac{m(m+1)\cdots(m+r-1)}{2^r}\left(\frac{s}{t}\right)^{m/2}\frac{1}{t^{r/2}}\frac{1}{\sqrt{\tau_1}}\cdots\frac{1}{\sqrt{\tau_r}}. \tag{36}$$

Now

$$\begin{aligned}\sum_{\boldsymbol{\tau}} \frac{1}{\sqrt{\tau_1}} \cdots \frac{1}{\sqrt{\tau_r}} &= \frac{1}{r!}\left(\int_s^t \frac{1}{\sqrt{\tau}}\, d\tau + O\left(\frac{1}{\sqrt{s}}\right)\right)^r \\ &= \left(1 + O\left(\tfrac{r}{\sqrt{s}}\right)\right) \frac{2^r}{r!}\left(\sqrt{t} - \sqrt{s}\right)^r.\end{aligned}$$

The result follows. □

Assuming the same conditions on r, s as in Lemma 11b, define

$$\rho(s,t,r) = \prod_{\tau=s}^{t} \exp\left(-\ell \frac{d(s,\tau)}{2m\tau}\right).$$

As in the proof of Lemma 11 let $d(s,t) = m + r$ and let $\boldsymbol{\tau} = (\tau_1, ..., \tau_r)$ denote the transition steps of $d(s,t)$ from m to $m + r$. As before, let $\tau_0 = s$ and $\tau_{r+1} = t$. Let $\rho(s,t,r : \boldsymbol{\tau})$ be the value of ρ given $\boldsymbol{\tau}$.

Then

$$\begin{aligned}\rho(s,t,r:\boldsymbol{\tau}) &= \exp\left(-\frac{l}{2m}\sum_{j=0}^{r} \sum_{\tau_j \le T < \tau_{j+1}} \frac{m+j}{T}\right) \\ &= \exp\left(-\frac{\ell}{2m}\sum_{j=0}^{r}(m+j)\left(\log\frac{\tau_{j+1}}{\tau_j} + O\left(\frac{1}{\tau_j}\right)\right)\right) \\ &= \left(1 + O\left(\tfrac{(m+r)^2}{s}\right)\right)\left(\frac{s}{t}\right)^{\frac{\ell}{2}} t^{-r\ell/2m}{\tau_1}^{\ell/2m} \ldots {\tau_r}^{\ell/2m}.\end{aligned}$$

Thus, combining $\rho(s,t,r:\boldsymbol{\tau})$ with $p(s,t,r:\boldsymbol{\tau})$ from (36) and summing over $\boldsymbol{\tau}$ we have

$$\begin{aligned}\mathbf{E}\,\rho(s,t,r) &= \sum_{\boldsymbol{\tau}} \rho(s,t,r:\boldsymbol{\tau})p(s,t,r:\boldsymbol{\tau}) \\ &= \left(1 + O\left(\tfrac{(m+r)^2}{s}\right)\right)\left(\tfrac{s}{t}\right)^{(m+\ell)/2}\binom{m+r-1}{r}\frac{r!}{2^r}\frac{1}{t^{r(1+\ell/m)/2}}\sum_{\boldsymbol{\tau}}\prod_{j=1}^{r}\frac{1}{\tau_j^{(1-\ell/m)/2}} \\ &= \left(1 + O\left(\tfrac{(m+r)^2}{s}\right) + O\left(\tfrac{r}{s^{(1-\ell/m)/2}}\right)\right)\binom{m+r-1}{r}\left(\frac{1-\left(\frac{s}{t}\right)^{(1+\ell/m)/2}}{1+\ell/m}\right)^r.\end{aligned}$$

Thus summing over r we get

$$\mathbf{E}\,\rho(s,t) = (1 + o(1))\left(\frac{1+\frac{\ell}{m}}{1+\frac{\ell}{m}\left(\frac{t}{s}\right)^{(1+\ell/m)/2}}\right)^m.$$

Thus, using the transformations, $x = s/t$ and $y = \ell/\sqrt{x}$, we find

$$\begin{aligned}\lim_{m,t\to\infty}\frac{\mathbf{E}\,\nu_{\ell,m}(t)}{t} &= \lim_{m,t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\left(\frac{1+\frac{\ell}{m}}{1+\frac{\ell}{m}\left(\frac{t}{s}\right)^{(1+\ell/m)/2}}\right)^m \\ &= e^{\ell}\int_0^1 e^{-\ell/\sqrt{x}}\,dx \\ &= e^{\ell}2\ell^2\int_{\ell}^{\infty} y^{-3}e^{-y}\,dy.\end{aligned}$$

as required.

5 Extensions and further research

There are some natural questions to be explored in the context of the above models.

- It should be possible to extend the analysis to other models of web-graphs e.g. [12], [16]. In principal, one should only have to establish that random walks on these graphs are rapidly mixing.
- One can consider non-uniform random walks. Suppose for example that each $v \in [t]$ is given a weight $\lambda(v)$ and when at a vertex v the spider chooses its next vertex with probability proportional to $\lambda(v)$. If $\Lambda(v) = \sum_{N(v)} \lambda(v)$ ($N(v)$ denotes the neighbours of v) then the steady state probability $\pi(v)$ of being at v in such a walk is proportional to $\Theta(v) = \lambda(v)\Lambda(v)$. Again, once one shows rapid mixing it should be possible to obtain an expression like (1) for the number of unvisited vertices.
- We have only estimated the expectation of the number of unvisited vertices. It would be interesting to establish a concentration result.

References

[1] D. Achlioptas, A. Fiat, A.R. Karlin and F. McSherry, Web search via hub synthesis, *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science* (2001) 500-509.

[2] M. Adler and M. Mitzenmacher, *Toward Compressing Web Graphs*, To appear in the 2001 Data Compression Conference.

[3] W. Aiello, F. Chung and L. Lu, Random evolution in massive graphs, *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science* (2001) 510-519.

[4] W. Aiello, F. Chung and L. Lu, A random graph model for power law graphs, *Experiment. Math.* 10 (2001) 53-66.

[5] R. Albert, A. Barabasi and H. Jeong, *Diameter of the world wide web.* Nature 401:103-131 (1999) see also `http://xxx.lanl.gov/abs/cond-mat/9907038`

[6] A. Barabasi, R. Albert and H. Jeong, Scale-free characteristics of random networks: The topology of the world wide web, *Physics A* 272 (1999) 173-187.

[7] A. Barabasi and R. Albert, Emergence of scaling in random networks, *Science* 286 (1999) 509-512).

[8] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale free random graph process, Random Structures and Algorithms 18 (2001) 279-290.

[9] B. Bollobás and O. Riordan, *The diameter of a scale free random graph*, to appear.

[10] B. Bollobás and O. Riordan, *Mathematical results on scale-free random graphs*, to appear.

[11] B. Bollobás and O. Riordan, *Robustness and vulnerability of scale-free random graph*, to appear.

[12] A. Broder, R. Kumar, F.Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. *Graph structure in the web.* `http://gatekeeper.dec.com/pub/DEC/SRC /publications/stata/www9.htm`

[13] Buckley, P.G. and Osthus, D., Popularity based random graph models leading to a scale-free degree sequence, preprint available from http://www.informatik.hu-berlin.de/~osthus/.

[14] F. Chung and L. Lu, Connected components in a random graph with given degree sequences, *Annals of Combinatorics*, to appear.

[15] F. Chung and L. Lu, The average distances in random graphs with given expected degrees, *PNAS* 99 (2002) 15879-15882.

[16] C. Cooper and A.M. Frieze, A general model of web graphs, *Proceedings of ESA 2001*, 500-511.

[17] E. Drinea, M. Enachescu and M. Mitzenmacher, Variations on random graph models for the web, Harvard Computer Science Technical Report TR-06-01.

[18] M.R. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, Measuring Index Quality Using Random Walks on the Web, *WWW8 / Computer Networks* 31 (1999) 1291-1303.

[19] W. Hoeffding, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58 (1963) 13-30.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. The web as a graph, *Proceedings of the 19th Annual ACM Symposium on Principles of Database Systems* (2000) 1-10.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. Stochastic models for the web graph, Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (2000) 57-65.

[22] L. Lu, The diameter of random massive graphs, Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms (2001) 912-921.

[23] A. Sinclair and M. Jerrum, Approximate counting, uniform generation, and rapidly mixing Markov chains, *Information and Computation* 82 (1989) 93-133.