Crawling on web graphs

Colin Cooper
Department of Mathematical and Computing Sciences,
Goldsmiths College,
University of London,
London SW14 6NW, UK
c.cooper@gold.ac.uk

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

February 19, 2002

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 [4] and [7] 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], [5], [6], [8], [9], [10], [12] and [13] 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 model may be replaced when a new search is made at a future time period or may be continously

*Supported in part by NSF grant CCR-9818411.

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 (spider, sniffer) 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).

We consider the following 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], [4], [5], [6], [7], [8], [9], [10], [12] and [13] 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 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 $v \in [t-1]$,

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

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,*

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

where for $s \leq \tau \leq t$,

$$0 \leq \zeta(s, \tau) \leq \frac{1}{m}.$$

Let

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

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,*

$$\eta_\ell = 2e^{9/(4\ell)}(\pi/\ell)^{1/2}\left(1 - \Psi\left(\frac{3}{\sqrt{2\ell}}\right)\right)$$

where $\Psi(x)$ *is the standardized Normal cumulate for the interval* $(-\infty, x]$*.*

In particular, $\eta_1 = 0.57\cdots$ *and* $\eta_\ell \sim 2(\pi/\ell)^{1/2}$ *as* $\ell \to \infty$*.*

(b) *For Model 2*

$$\eta_\ell = e^\ell \int_0^1 \exp\left(-\frac{\ell}{\sqrt{x}}\right)\, dx.$$

In particular, $\eta_1 = 0.59\cdots$*.*

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

2 Proof of Theorem 1

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. Next 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. This means that the τth 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 $\boldsymbol{\theta} = (\theta_\tau : s \le \tau \le t)$ be integers satisfying

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

Let

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

and for some event $\boldsymbol{C}$ let $\mathbf{Pr}_{\boldsymbol{\theta}}(\boldsymbol{C}) = \mathbf{Pr}(\boldsymbol{C} \mid \mathcal{D}(\boldsymbol{\theta}))$ be the corresponding conditional event. For $s \le \tau \le t$ let

$$\tau_0 = \tau - 10^6 \ln \tau. \tag{3}$$

We then prove that for $t/\ln t \le s < t$ and a $\boldsymbol{\theta}$ satisfying (2),

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\overline{\mathcal{A}}_s(t) \mid \mathcal{A}_s(t-1)) = \frac{\theta_{t_0}}{2mt_0}\left(1 - \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,t)\right)\right) + O(t^{-3})\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1} + \widetilde{O}(t^{-3/2}), \tag{4}$$

where if $N(s,\tau)$ denotes the set of neighbours of s in $G(\tau)$ then

$$\gamma(s,t) = \frac{1}{\theta_{t_0}} \sum_{y \in N(s,t_0)} \frac{1}{d(y,t)} \in \left[0, \frac{1}{m}\right].$$

Furthermore,

$$\mathbf{Pr}(\mathcal{A}_s(s)) = 1 - O(s^{-1}). \tag{5}$$

From this we deduce (1) of Theorem 1 as follows: If $\boldsymbol{\theta}$ satisfies (2) and $t/\ln t \leq s < t$ then

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

We see that if $\theta_t \leq \Delta_t^*$ then

$$\begin{aligned}
\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t)) &= \prod_{\tau=s}^{t} \left(1 - \frac{\theta_{\tau_0}}{2m\tau}(1 - \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,\tau)\right) + \widetilde{O}(t^{-1/2}))\right) && (6)\\
&= \prod_{\tau=s}^{t} \left(1 - \frac{\theta_{\tau}}{2m\tau}(1 - \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,\tau)\right) + \widetilde{O}(t^{-1/2}))\right) && (7)\\
&= \prod_{\tau=s}^{t} \left(1 - \frac{\theta_{\tau}}{2m\tau}(1 - \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,\tau)\right))\right) + \widetilde{O}(t^{-1/2})
\end{aligned}$$

Note that we can go from (6) to (7) because $\theta_\tau = \theta_{\tau-\tau_0}$ except for at most $\tau_0 \Delta_t^*$ instances.

Thus absorbing the cases where $\boldsymbol{\theta}$ does not satisfy (2) into the error term, summing out the conditional probabilities over walks, we get

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

where $\zeta(s,\tau) = \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,\tau)\right)$ and (1) follows.

We now consider the random walk made by the spider S. A *random walk* on an fixed undirected graph G is a Markov chain $\{X_t\} \subseteq V$ 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 $\pi(v)$ of the walk being at a vertex v is,

$$\pi(v) = \frac{d(v)}{d(G)}, \tag{8}$$

where $d(v)$ is the degree of v and $d(G)$ is the total degree of the graph G.

We will need a finite time approximation of the probability distribution π for the random walk on $H_s(t) = G(t) - s$ for $s \leq t$. We obtain this by considering the *mixing time* of the walk based on a conductance bound (9) of Jerrum and Sinclair [14].

Let s, t be fixed. Let P denote the transition matrix of the random walk on $H_s(t)$. Let $P^{i,\tau}$ denote the distribution of the τth step of a random walk on $H_s(t)$ which starts at vertex i. For $K \subset [t] \setminus \{s\}$ let $\overline{K} = [t] \setminus (K \cup \{s\})$ and

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

It follows from (8) 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

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

For the proof of Theorem 1 it is adequate to consider vertices s created after step $t/\ln t$. Let $\Delta_s(t)$ be the maximum degree of vertices in the interval $[s, t]$. We show that

Lemma 1. *Suppose $s \geq t/\ln t$ and let*

$$\mathcal{G}(s,t) = \left\{ G(t) : \Phi(s,t) > \frac{1}{\ln t}, \text{ and } \Delta_s(t) \leq \Delta_t^* \text{ and } d(s, \tau+1) - d(s,\tau) \leq 5, \; s < \tau < t \right\}.$$

Then, in both Model 1 and Model 2,

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

2.1 Proof of (4) and (5)

We can now apply the main result of [14].

$$|P^{i,\tau}(v) - \pi(v)| \leq \left(1 - \frac{\Phi^2}{2}\right)^{\tau} \frac{\pi(v)}{\pi_{\min}} \tag{9}$$

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

Using (9) and Lemma 1 we see that with $\mu_0 = 10^5 \ln t$, **whp**

$$|P^{i,\mu_0}(v) - \pi(v)| \leq t^{-3} \qquad \forall v \in [t] \setminus \{s\}. \tag{10}$$

We are glossing over one technical point here. Strictly speaking, (9) 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 $\boldsymbol{\theta}$ with $\theta_t \leq \Delta_t^*$ and assume that $t/\ln t \leq s \leq t$.

Now with $t_0 = t - 10\mu_0$ and

$$\begin{aligned}
I &= [t_0+1, t-1] \\
J_1 &= \{\sigma \in I :\ \exists \tau \in I \text{ such that } X_\tau = \sigma\} \\
\mathcal{E}_0 &= \{X_\tau \neq s,\ \tau \in I\} \\
\mathcal{E}_1 &= \{\exists j \in J :\ j \text{ has} \geq 2 \text{ neighbours in } \{X_\sigma :\ \sigma \in I\}\} \cup \\
&\qquad\qquad\qquad \{\exists j, j' \in I :\ j \in J \text{ and } j, j' \text{ are neighbours}\} \\
\mathcal{F}_k &= \{|J| = k\} \qquad k \geq 0 \\
\mathcal{F}_{\geq k} &= \{|J| \geq k\}
\end{aligned}$$

and write

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid \mathcal{A}_s(t-1)) = \\
&\quad \sum_{\substack{G \in \mathcal{G}(s,t_0) \\ x \in [t_0] \setminus \{s\}}} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{A}_s(t_0)) \mathbf{Pr}_{\boldsymbol{\theta}}(X_{t_0} = x, G(t_0) = G \mid \mathcal{A}_s(t-1)) + \\
&\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s, G(t_0) \notin \mathcal{G}(s,t_0) \mid \mathcal{A}_s(t-1)). \qquad (11)
\end{aligned}$$

It follows from Lemma 1 that

$$\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s, G(t_0) \notin \mathcal{G}(s,t_0) \mid \mathcal{A}_s(t-1)) = o(t^{-3}\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1}). \qquad (12)$$

To deal with the rest of (12) we write

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{A}_s(t_0)) = \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0) \\
&\quad = \sum_{k=0}^{1} \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_k) \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_k \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0) \\
&\qquad + \mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_{\geq 2}) \mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_{\geq 2} \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0). \qquad (13)
\end{aligned}$$

Given $x, G(t_0)$, conditioning on $\mathcal{F}_0$ is equivalent to S doing a random walk on $H = H_s(t_0)$ starting at x. Thus, we get

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_0) = \\
&\quad \mathbf{E}_{\boldsymbol{\theta}}\left(\sum_{y \in N(s,t_0)} \left(\left(\frac{d(y,t_0) - 1}{2mt_0 - O((\ln t)^5)} + O\left(\frac{1}{t^3}\right) \right) \cdot \frac{1}{d(y,t)} \right) \,\middle|\, x, G, \mathcal{E}_0, \mathcal{F}_0 \right) = \\
&\qquad\qquad \frac{\theta_{t_0}}{2mt_0} \left(1 - \mathbf{E}_{\boldsymbol{\theta}} \left(\gamma(s,t) + \frac{1}{\theta_{t_0}} \sum_{y \in N(s,t_0)} \frac{d(y,t_0) - d(y,t)}{d(y,t)} \right) \right) + \widetilde{O}(t^{-2}) \qquad (14)
\end{aligned}$$

We will next argue that

$$\begin{aligned}
\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{F}_{\geq k} \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0) &= \widetilde{O}(t^{-k}) \qquad k = 1,2. && (15) \\
\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_1) &= \widetilde{O}(t^{-1}). && (16) \\
\mathbf{E}_{\boldsymbol{\theta}}\left(d(y,t) - d(y,t_0)\right) &= \widetilde{O}(t^{-1/2}). && (17)
\end{aligned}$$

It follows from (11)–(17) that

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{\theta}}(X_t = s \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0) = \\
&\qquad\qquad \frac{\theta_t}{2mt}(1 - \mathbf{E}_{\boldsymbol{\theta}}\left(\gamma(s,t)\right) + O(t^{-3}\mathbf{Pr}_{\boldsymbol{\theta}}(\mathcal{A}_s(t-1))^{-1}) + \widetilde{O}(t^{-3/2})
\end{aligned}$$

and removing the conditioning on $X_{t_0} = x, G(t_0) = G$ yields (4).

2.1.1 Proof of (15)

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 $t_0+1, \dots, i$ 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 $t_0, t_0+1, \dots, i$ is that they are chosen from a set of size $t - o(t)$ either randomly (Model 1) or from the same set with the same probabilities as the steady state of the walk on $G(t)$ (Model 2). Of course for those i for which $\theta_i > \theta_{i-1}$, we have one neigbour s, which we don't include in this set of neighbours. Let $\mathcal{Y}_i = \{\text{YES at } i \text{ and } X_i \in \{t_0+1, \dots, i\}\}$. If the degrees of $X_j, j \in \{t_0+1, \dots, i\}$ is d_j then

$$\mathbf{Pr}(\mathcal{Y}_i) = O\left(\sum_{j=t_0+1}^{i} \frac{d_j}{t} \cdot \frac{1}{d_j}\right) = O\left(\frac{|I|}{t}\right). \tag{18}$$

Since $\mathcal{F}_{\geq 1} \subseteq \bigcup_{i \in I} \mathcal{Y}_i$ we have (15) for $k = 1$.

Now assume that i_1 is the first i for which $\mathcal{Y}_i$ occurs and that $X_{i_1-1} = j_1$ has neighbours $K_1 \subseteq [t_0+1, i_1]$.

There are 2 cases to consider.

Case 1: $|K_1| \geq 2$.
The probability of this is $\widetilde{O}(t^{-2}|I|^2)$.

Case 2: $|K_1| = 1$.
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 (15).

2.1.2 Proof of (16)

Let $J = \{j_1\}$. If j_1 is visited first at time $t_1 \leq t_0 + 5 \times 10^5(\ln t)$ 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 (10) for H' we can argue, as in the proof of (15) that the conditional probability $X_t = s$ is $\widetilde{O}(t^{-1})$ as required.

Suppose next that the first visit to j_1 after time $t_0 + 5 \times 10^5(\ln t)$. We now write

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

First observe that $\mathbf{Pr}(\mathcal{E}_1 \mid X_{t_0} = x, G(t_0) = G, \mathcal{E}_0, \mathcal{F}_1) = \widetilde{O}(t^{-3/2})$. Use (18) plus an extra $\widetilde{O}(t^{-1/2})$ factor for the extra neighbour(s).

So now assume that $\mathcal{E}_1$ does not occur. 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 (14). 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 $t - o(t)$.

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 $l_1 = v$ is at most twice this. It follows that in any subsequent step, the probability S is at v is at most $2\pi(v)$). 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 $\widetilde{O}(t^{-1})$ and (16) 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 l_1. Thus in any subsequent step, the probability S is at v is asymptotically equal to $\pi(v)$ and we can re-use the analysis for Model 1..

2.1.3 Proof of (17)

This follows from the fact that in Model 2, the maximum degree in $G(t)$ is $O(t^{1/2})$ **whp**, see e.g. [8]. For Model 1 the maximum degree is $o(\ln t)$ with sufficiently high probability.

2.1.4 Proof of (5)

We consider the process from time $s-\mu_0$ to s. At time s, the chance that the spider is at a neighbour y of s is either $O(m(s-1)^{-1})$ (Model 1) or $O(d(y,s-1)/s)$ (Model 2, because of (10),) and then in both models, the probability of moving from y to s is $1/d(s-1,t)+1$ and we get (5).

2.2 $\ell \geq 1$

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).

3 Proof of Theorem 2

Theorem 3. Model 1

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

$$\eta_\ell = e^{9/(4\ell)} \sqrt{\pi/\ell} \left(1 - \Psi\left(\frac{3}{\sqrt{2\ell}}\right)\right),$$

where $\Psi(x)$ is the standardized Normal cumulate for the interval $(-\infty, x]$.

Model 2

$$\eta_\ell = e^\ell \int_0^1 \exp\left(-\frac{\ell}{\sqrt{x}}\right) \, dx.$$

Proof Model 1
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

$$\begin{aligned}\prod_{\tau=s}^{t}\left(1-\frac{d(s,\tau)}{2m\tau}\gamma(s,\tau)\right)^{\ell} &= \exp\left(-\left(1+O\left(\frac{1}{m}\right)\right)\ell\sum_{\tau=s}^{t}\frac{d(s,\tau)}{2m\tau}\right) \\ &= \exp\left(-\left(1+O\left(\frac{1}{m}\right)\right)\frac{\ell}{2m}\sum_{r=s}^{t}X_r\ln t/r\right) \\ &= \left(1+O\left(\frac{1}{m}\right)\right)\prod_{r=s}^{t}\left(\frac{r}{t}\right)^{\frac{\ell X_r}{2m}}.\end{aligned}\tag{19}$$

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

Thus

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

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

Using the transformation $x=e^{-y}$ we obtain

$$\begin{aligned}\nu_{\ell,m}(t) &= \left(1+O\left(\frac{1}{m}\right)\right)t\int_0^{\infty}\exp\left(-\frac{3}{2}y-\frac{\ell}{4}y^2\right)\,dy \\ &= \left(1+O\left(\frac{1}{m}\right)\right)t\;2e^{9/(4\ell)}\sqrt{\pi/\ell}\left(1-\Psi\left(\frac{3}{\sqrt{2\ell}}\right)\right).\end{aligned}$$

Model 2

Lemma 2.

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

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

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

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

$$p(s,t,r:\boldsymbol{\tau}) = \left(1+O\left(\tfrac{(m+r)^3}{s}\right)\right)\frac{m(m+1)\cdots(m+r-1)}{2^r}\left(\frac{s}{t}\right)^{m/2}\frac{1}{t^{r/2}}\frac{1}{\sqrt{\tau_1}}\cdots\frac{1}{\sqrt{\tau_r}}. \qquad (20)$$

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. □

Next let

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

As in the proof of Lemma 2 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

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

Thus, combining $\rho(s,t,r:\boldsymbol{\tau})$ with $p(s,t,r:\boldsymbol{\tau})$ from (20) and summing over $\boldsymbol{\tau}$ we have

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

Thus summing over r we get

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

Thus

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

as required. When $\ell=1$, $\eta_1 = 0.59634...$.

4 Proof of Lemma 1

4.1 Calculations for Model 1

4.1.1 Degree sequence of $G(t)$

We begin with the degree sequence of $G(t)$. The degree $d(s,t)$ of vertex s in $G(t)$ is distributed as

$$m + B(m,(s+1)^{-1}) + \cdots + B(m,t^{-1}) \tag{21}$$

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

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

Lemma 3.

(a) $\bar{d}(s,t) = m(1+H_t-H_s) \le m(2+\ln t/s)$
where $H_k = 1+\frac{1}{2}+\cdots+\frac{1}{k}$.

(b) $\mathbf{Pr}(\Delta(G(t)) \ge 2m\ln t) = o(t^{-3})$
where $\Delta(G(t))$ *is the maximum degree in* $G(t)$.

(c) $\mathbf{Pr}(\exists K\subseteq[t]:\ |K|\ge 3t/4$ *and* $d(K,t)\le(1.1)mt) = o(e^{-cmt})$ *for some absolute constant* $c>0$.

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

Proof (a) follows from (21) and (b) follows from Theorem 1 of Hoeffding [11]. (d) is easy, since $d(s,\tau+1) - d(s,\tau) = B(m,\tau^{-1})$.

For (c) 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 (c) and the lemma. □

4.1.2 Conductance for 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 4.

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

Proof 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 $L = \overline{K} = [t] \setminus K$ and $L = K$. It follows from Lemma 3(c) that with probability $1 - o(t^{-3})$

$$\Phi(s,t) \geq \min_{\pi(K) \leq 1/2} \frac{e(K:\overline{K}) - |K|}{m|K| + e(K:\overline{K})} \geq \min_{|K| \leq 3t/4} \frac{e(K:\overline{K}) - |K|}{m|K| + e(K:\overline{K})}. \tag{22}$$

($e(K:\overline{K}) - |K|$ 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 (22) this proves the lemma.

$$\mathbf{Pr}_{\boldsymbol{\theta}}(\exists K:\ e(K:\overline{K}) \leq m|K|/150) = o(t^{-3}). \tag{23}$$

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

$$\mathbf{E}_{\boldsymbol{\theta}}\left(Y_K\right) \geq \sum_{\tau=\kappa}^{t-4k/7-1} \frac{3(m-5)k/7}{\tau + 4k/7} \geq \frac{3(m-5)k}{7} \ln\left(\frac{t-1}{\kappa + 4k/7}\right) \geq \frac{mk}{16}.$$

Explanation: Consider the $\geq t - \kappa - 4k/7 - 1$ vertices of $[t] - [\kappa] \cup \{s\}$. 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_-.

Applying Theorem 1 of [11] 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 (23) 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 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 (23). □

4.2 Calculations for Model 2

We need to prove bounds corresponding to those given in Lemma 3 for Model 1.

4.2.1 Maximum Degree

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

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

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) $s, se^{1/4}, ...se^{r/4}, ..., se^{k/4}$. Here $\lceil se^{k/4} \rceil = t$, so that $k \leq 4 \ln\ln t$, as $s \geq t/\ln t$.

Suppose that, at the start of $I_r = (\lceil se^{r/4} \rceil, \lceil se^{(r+1)/4} \rceil]$ 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 $\leq 2d(r)$, t he number X_τ of edges acquired at step $\tau \in I_r$ is dominated by $B(m, d(r)/(m\tau))$, so that the number of edges gained during this time has expected value

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

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

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

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 $B(10(\ln t)^5, (2m\tau)^{-1})$. □

4.2.2 Conductance

Lemma 6. *There is an absolute constant $\xi > 0$ such that*

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

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

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

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

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

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

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

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 $m > 2/(c\zeta)$ then the probability there is a set of size $\lceil\zeta t\rceil$ which has degree exceeding $6m\zeta^{1/4}$ is exponentially small. In which case, every set K of size at least $t - \lceil\zeta t\rceil$ has total degree $d(K,t) \geq 2mt - 6m\zeta^{1/4}$ and the lemma follows by traking ζ sufficiently small. □

Lemma 7. *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 $c = e^{-8}$.

4.2.3 Case of K small

Let $Q = \{K \cap [\sqrt{kt}]\}$ and let $R = K \setminus Q$.

Case of $|Q| \geq k/2$.
Let $X_t = X_t(Q)$ be the number of those edges directed into Q 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. Thus

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

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

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

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

Case of $|R| \geq k/2$. We consider the evolution of the set $R = \{u_1, u_2, \ldots, u_r\}$ 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 $Z = Z(R) = \sum_{j=1}^{r} Y_j$ 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 $\alpha > 0$,

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

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 R not incident with K on the condition that $\delta \leq m/10$, completing the analysis of this case.

4.2.4 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 6. Let $M = [T]$ and $N = [T+1, t]$. Let $Q = K \cap M$ and let $R = K \cap N$. We calculate the expected number of edges $\mu(Q,R)$ of $L = (R \times (N \setminus Q)) \cup ((N \setminus R) \times Q)$ 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

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

Thus

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

Let $\mu(k) = \min_{Q,R}\mu(Q,R)$. Then 'somewhat crudely'

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

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 δ be a very small positive constant, and 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

$$\begin{array}{rcl} \mathbf{Pr}\left(\exists Q, R \subseteq N : \; Z \leq \delta\mu(k)\right) & \leq & 2^t \exp\left(-\mu(k)\left(1 - \delta \ln e/\delta\right)\right) \\ & \leq & 2^t e^{-\eta m t/2} \\ & = & e^{-\eta m t/3}, \end{array}$$

provided 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 conductivity Φ_K is always $\Omega(1/k)$.

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. [7], [8]. 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] 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`

[5] B. Bollobás, O. Riordan and J. Spencer, *The degree sequence of a scale free random graph process*, to appear.

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

[7] 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`

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

[9] E. Drinea, M. Enachescu and M. Mitzenmacher, *Variations on random graph models for the web.*

[10] 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.

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

[12] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. *The web as a graph.* `www.almaden.ibm.com`

[13] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. *Stochastic models for the web graph.* `www.almaden.ibm.com`

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

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.