On the Stability of Web Crawling and Web Search

Reid Anderson[1], Christian Borgs[1], Jennifer Chayes[1],
John Hopcroft[2], Vahab Mirrokni[3], and Shang-Hua Teng[4]

[1] Microsoft Research
[2] Cornell University
[3] Google Research
[4] Boston University

Abstract. In this paper, we analyze a graph-theoretic property motivated by web crawling. We introduce a notion of stable cores, which is the set of web pages that are usually contained in the crawling buffer when the buffer size is smaller than the total number of web pages. We analyze the size of core in a random graph model based on the bounded Pareto power law distribution. We prove that a core of significant size exists for a large range of parameters $2 < \alpha < 3$ for the power law.[1]

1 Introduction

Since the World Wide Web is continually changing, search engines [1] must repeatedly crawl the web, and update the web graph. In an ideal world one would search, discover, and store all web pages on each crawl. However, in practice constraints allow indexing and storing only a fraction of the web graph [3]. This raises the question as to what fraction of the web one needs to crawl in order to maintain a relatively stable set of pages that contains all sufficiently important web pages.

When a link to a web page is encountered, the page is said to be *discovered.* When the page is retrieved and explored for its links, it is said to be *explored.* Thus, we can partition the web into three types of pages. (1) Pages the crawl has explored; (2) Pages the crawl has discovered but not explored; and (3) All other pages.

Web crawling can be viewed as a dynamic process over the entire web graph. As time goes by, these three sets change dynamically, depending on the crawling algorithm as well as the space/time constraints. Let S_t be the set of pages that have been discovered and explored at time t. The search engine typically ranks pages in S_t. When users make queries during phase t, only pages from S_t are returned. Presumably these are the pages that were deemed sufficiently important to index and are used to answer queries. Let C_t be the set of pages that have been discovered but not explored at time t. At this point the edges from pages in C_t are not known since the search has not crawled these pages.

[1] Most of this work was done while the authors were at Microsoft Research.

S.-H. Hong, H. Nagamochi, and T. Fukunaga (Eds.): ISAAC 2008, LNCS 5369, pp. 680–691, 2008.

Consider the subgraph of the web that consists of all web pages in $S_t \cup C_t$ and all directed edges from pages in S_t ending in either S_t or C_t. At the next stage the search engine would calculate the page rank of all pages in this graph and select the set of size b of pages of highest page rank to be S_{t+1}. The pages in S_{t+1} would then be explored to produce a new set C_{t+1} of pages reachable from pages in S_{t+1} but which are not in S_{t+1}. Here b is determined by the available amount of storage.

The space constraints immediately raise several basic questions in web crawling and web search. An important question is how large b needs to be in order for the search engine to maintain a core that contains all sufficiently important pages from the web. Assuming the web is on the order of 100 billion pages, is a buffer of size of 5 billion sufficient to ensure that the most important 100 million pages are always in the buffer and hence available to respond to queries? In general, what percent of pages are stable in the sense that they are always in the buffer? What percent of pages are regularly moving in and out of the buffer? What percent of the buffer is just random pages? What is the relationship between the importance of a page (such as high page rank or high degree) with the frequency that page is in the buffer? These questions are particular interesting when the graph is changing with the time? How frequently must we do a crawl where frequency is measured by the percentage of the graph that changed between the crawls? How should we design high a quality crawl given the space and time constraints? For example, should we completely explore the highest page ranked pages or should we explore some fraction of links from a larger number of pages, where the fraction is possibly determined by the page rank or degree? How accurately do we need to calculate the page rank in order to maintain the quality? Could we substitute in-degree for page rank?

Clearly the theoretical answers to above questions depend on how the underlying graph is modeled. We investigate the behaviors of the web crawling process and web crawling dynamics. The motivation behind our investigation is to design better and robust algorithms for web crawling and web search.

We will focus on the stability of web crawling and its potential impact to web search. In particular, we analyze a graph-theoretic property that we discovered based on our initial experiments involving web crawling dynamics.

Suppose b pages are stored at each stage of the crawling. We have observed that, for various choices of $\mathsf{EXPLORE}$ and STORE functions and large enough b, the sets S_t do not converge. However, there exists t_0 such that $\left(\cap_{t\geq t_0}^{\infty} S_t\right)$ converges to a non-empty set. We call $K = \left(\cap_{t\geq t_0}^{\infty} S_t\right)$ the *core* of the crawling dynamics with ($\mathsf{EXPLORE}$, STORE) transitions. Naturally, the size of the core depends on b as well as on ($\mathsf{EXPLORE}$, STORE). When b is small, the core might be empty. Naturally, when $b = |V|$, the whole graph is the core. When $b = 1$, we do not expect a non-empty core.

In this paper, we consider a simplified crawling algorithm with limited space. Let

$$C_t = \mathsf{EXPLORE}(S_{t-1}) = \{v\ |(u \rightarrow v) \in E, \text{for some } u \in S_{t-1}\} - S_{t-1},$$

be the set of direct neighbors of S_{t-1}. For each page $v \in S_{t-1} \cup C_t$, let $\Delta_t(v)$ be the number of links from pages in S_{t-1} to v. Then, $\mathsf{STORE}(S_{t-1}, C_t)$ is the set of b pages with the largest Δ_t value, where ties are broken according to some predefined rule.

We analyze the size of the core in a random graph model based on the bounded Pareto power law distribution [2,4]. We prove that a core of significant size exists when the power law parameter α lies in the range $[2 : 3)$.

2 The Core

Web crawling defines a sequence B_0, ..., B_t,...,B_∞, where B_t is the content of the buffer at time t. If a page enters the buffer at some stage and stays in the buffer after that, then we say the page is in the *core* of of the sequence.

For example, suppose the web graph is fixed and the crawl process is deterministic, then since the number of the subsets of web pages of size b is finite, the above sequence eventually become periodic. In other words, there exist a t_0 and p such that $B_{t_0} = B_{t_0+p}$. In this case, the *core* of the sequence is equal to $\cap_{t=t_0}^{t_0+p} B_p$. When the graph is fixed, but the web crawling is stochastic, we define the core as those pages that stay in the buffer with high probability.

In the rest of the paper, we assume B_0 is a set of size b uniformly chosen from the vertices of the input graph. The core of this graph is then defined according to the sequence produced by the crawling process.

3 Bounded Pareto Degree Distributions and Its Graphs

One of the objectives of this paper is to estimate the core size as a function of b, for a directed graph G. Naturally, this quantity depend on G and the initial set B_0. It is well known that the web graph has power law degree distribution [5]. To present our concepts and results as clearly as possible we use the following "first-order approximation" of the power-law graphs with bounded Pareto degree distributions. We first define a degree vector, which is the expected degree of a bounded Pareto degree distribution [2]. Then, we consider random graphs with expected degrees specified by the degree vector. We will show the core size depends on both the degree distribution and on the size of the buffer.

The expected number of vertices of degree k in the bounded Pareto degree distribution with parameters $\alpha > 1$ and positive integer n is $Cnk^{-\alpha}$ where $C = 1/\left(\sum_{x=1}^{\infty} x^{-\alpha}\right)$. We can construct a typical degree sequence as follows. Let h_i be the largest integer such that $\sum_{k=h_i}^{\infty} Cnk^{-\alpha} \geq i$. The sequence starts with the highest degrees $(h_1, ..., h_i, ...)$. Note that h_i is approximately

$$\left(\frac{C}{\alpha-1}\right)^{1/(\alpha-1)} \left(\frac{n}{i}\right)^{1/(\alpha-1)}$$

To construct the degree vector $\mathbf{d}_{\alpha,n}$, we start at the right with the degree one vertices and work to the left. Let k_0 be the smallest integer such that $Cnk^{-\alpha} < 1$.

- For each $1 \leq k < k_0$, from right to left, assign the k to the next $Cnk^{-\alpha}$ entries in $\mathbf{d}_{\alpha,n}$. To be more precise, when $Cnk^{-\alpha}$ is not an integer, we first assign k to $\lfloor Cnk^{-\alpha} \rfloor$ entries, and then, with probability $Cnk^{-\alpha} - \lfloor Cnk^{-\alpha} \rfloor$, add one more entry with value k. Suppose this step assigns n' entries.
- For $j = 1 : n - n'$, assign the value h_j to $\mathbf{d}_{\alpha,n}[n - n' - j]$.

In other words, $\mathbf{d}_{\alpha,n}$ is a sorted vector of expected degrees, from the largest to the smallest. In this vector, the smallest degree k that appear s times approximately solves $Cnk^{-\alpha} = s$, implying $k \approx s^{-1/\alpha} (Cn)^{1/\alpha}$.

Note that for $\alpha > 2$, the expected number of edges is proportional to n and for $\alpha = 2$ the expected number of edges is proportional to $n \log n$. That is,

$$E\left[||\mathbf{d}_{\alpha,n}||_1\right] = \begin{cases} \Theta(n) & \text{if } \alpha > 2, \text{ and} \\ \Theta(n \log n) & \text{if } \alpha = 2. \end{cases}$$

The graph we analyze has n vertices, labeled 1 to n, and is generated by the following random process: Let $m = ||\mathbf{d}_{\alpha,n}||_1$. Independently choose m directed edges, by first selecting a vertex i randomly according to $\mathbf{d}_{\alpha,n}$, and then choosing another vertex j randomly, also according to $\mathbf{d}_{\alpha,n}$. Note that this graph model allows multiple edges and self-loops.

Call a random graph from this distribution a random (α, n)-*BBPL* graph. This class of graphs has several statistical properties. The expected number of vertices with in-degree 0 is highly concentrated around $\sum_{k=1} e^{-k} Cnk^{-\alpha}$, which is a constant fraction of n.

Lemma 1. *For $h \geq 3$, the expected number of vertices with in-degree h or larger in a random (α, n)-BPPL graph is highly concentrated around*

$$C \sum_{k=1}^{\Theta(n^{1/(1-\alpha)})} \binom{n}{h} \left(\frac{k}{n}\right)^h \frac{n}{k^\alpha} \leq \Theta\left(\frac{n}{h^{(\alpha-1)(1+1/(2h))}}\right).$$

4 Estimating the Core Size for Power-Law Graphs

Consider a buffer B of b vertices. These vertices induce a graph which we will refer to as the *buffer-induced graph* of B. The buffer-induced degree of a vertex is its degree in the buffer-induced graph.

4.1 A Simple Thought Process

When the buffer size is a fraction of n, a vertex with constant degree may have buffer-induced degree of 0 and thus may drop out of the buffer. This implies that the core size might be $o(n)$, depending on the tie breaking rule. However, in this section, we show that for any ϵ, the core size is $\Omega(n^{1-\epsilon})$.

Suppose we do a crawl with a buffer large enough to hold every vertex of in-degree at least one in the web. This does not implies that the vertices with indegree at least in the original graph may stay or enter the buffer, since its buffer-induced may be 0. We show with high probability, the two highest degree vertices, to be called 1 and 2, are mutually connected and will enter the buffer in step 1. The fact that they are mutually connected means that they will always remain in the buffer. In subsequent steps, all vertices reachable from 1 and 2 will be added to the buffer and will remain in the buffer. Thus, the core contains all these vertices. We now give a lower bound on the expected number of vertices reachable from 1 and 2, which provides a lower bound on the core size.

Lemma 2. *The probability that vertices 1 and 2 are mutually connected to each other in a random (α, n)-BPPL graph is*

$$1 - e^{\Theta\left(-n^{\frac{3-\alpha}{\alpha-1}}\right)}.$$

Proof. Let m be the number of edges in a random (α, n)-BPPL graph. Thus, $m = ||\mathbf{d}_{\alpha,n}||_1$ and is linear in n. The probability that 1 and 2 are not mutually connected to each other is at most

$$\left(1 - \Theta\left(\frac{n^{1/(\alpha-1)}}{m}\frac{n^{1/(\alpha-1)}}{m}\right)\right)^m = e^{-\Theta\left(n^{\frac{3-\alpha}{\alpha-1}}\right)}.$$

□

With a relatively small buffer of size $\Theta(n^{1-1/(\alpha-1)} \log n)$ containing randomly chosen vertices, the probability that vertices 1 and 2 will be in the buffer in the next step is high. Note that for $\alpha = 3$, this buffer size is only $\Theta(\sqrt{n})$.

Lemma 3. *Suppose $G = (V, E)$ is a random (α, n)-BPPL graph and S is a set of b randomly chosen vertices of V. There exists a constant c such that if $b \geq cn^{1-1/(\alpha-1)} \log n$, then with high probability, there are edges from vertices in S to both vertices* 1 *and* 2*.*

Proof. Because in our model, the expected degree of each vertex is at least 1, $\sum_{u \in S} \mathbf{d}_{\alpha,n}[u] \geq b$. The expected indegrees of vertices 1 and 2 are $\Theta(n^{1/(\alpha-1)})$. Their expected indegrees counting only edges from S are

$$\Theta\left(\frac{n^{1/(\alpha-1)}}{n}\right) \sum_{u \in S_0} \mathbf{d}_{\alpha,n}[u] = \Theta\left(n^{1/(\alpha-1)}\frac{b}{n}\right) \geq \Theta(c \log n).$$

As this bound is highly concentrated, when c is large enough, with high probability (e.g., $1 - n^{-\Theta(c)}$), the buffer-induced in-degrees of vertices 1 and 2 are larger than 1. □

Lemma 4. *For $2 < \alpha < 3$, with high probability, the number of vertices reachable from $\{1, 2\}$ in a random (α, n)-BPPL graph G is $\Omega(n^{1-\epsilon})$.*

Here we will sketch an outline of the proof but skip some technical details since we will give a stronger result later with all the details. To better illustrate the analysis, instead of writing a proof for all $\alpha : 2 < \alpha < 3$, we choose a typical value in this range and provide an explicit derivation. The proof is easily adapted to handle all $\alpha : 2 < \alpha < 3$. Our choice of "typical" value is $\alpha = 11/4$. In this case, note that the expected degree of vertex 1 is $\Theta(n^{4/7})$. The expected number of vertices directly reachable from $\{1, 2\}$ is

$$\Theta\left(\sum_k \left[1 - \left(1 - \frac{k}{n}\right)^{n^{4/7}}\right] \frac{n}{k^{11/4}}\right) = \Theta(n^{4/7}).$$

The expected total degree of the nodes directly reachable from $\{1, 2\}$ is

$$\Theta\left(\sum_k k\left[1 - \left(1 - \frac{k}{n}\right)^{n^{4/7}}\right] \frac{n}{k^{11/4}}\right) = \Theta\left(\int_1^{n^{3/7}} \frac{n^{4/7}}{k^{3/4}}\right) = \Theta(n^{19/28})$$

Let $S_0 = \{1, 2\}$. Let S_t be the set defined by the set of vertices t hops away from $\{1, 2\}$. Let $\Delta(S_t)$ be their expected degree. We thus have

$$E[|S_1|] = \Theta(n^{4/7}), \text{ and } E[|\Delta(S_1)|] = \Theta(n^{19/28}).$$

The key to the analysis is that $E[|\Delta(S_1)|]$ is magnitudely larger than $E[|S_1|]$, which means that the frontiers of the Breadth-First Search starting from $\{1, 2\}$ have good expansions. There are two types of out-links from S_t: the edges to $S_0 \cup ... \cup S_t$ and the edges to S_{t+1}. We now bound the expected size of S_2 and the expected total degree $\Delta(S_2)$ of S_2. A similar analysis can be extended to any t.

Let $F_1 = \{v \mid (u \to v) \in E, \text{for some } u \in S_1\}$ and let $B_1 = F_1 \cap (S_0 \cup S_1)$. We have $S_2 = F_1 - B_1$. Note that $E[|B_1|] \leq E[|S_1|] + 2 = \Theta(n^{4/7})$. Thus,

$$\begin{aligned}
E[|S_2|] &= E[|F_1|] - E[|B_1|] \\
&= \left(\sum \left[1 - \left(1 - \frac{k}{n}\right)^{n^{19/28}}\right] \frac{n}{k^{11/4}}\right) - E[|B_1|] \\
&= \Theta\left(\int \left[1 - \left(1 - \frac{k}{n}\right)^{n^{19/28}}\right] \frac{n}{k^{11/4}}\right) - E[|B_1|] \\
&= \Theta\left(\int_1^{n^{9/28}} \left[n^{19/28}\left(\frac{k}{n}\right)\right] \frac{n}{k^{11/4}}\right) - E[|B_1|] \\
&= \Theta(n^{19/28}) - \Theta(n^{4/7}) = \Theta(n^{19/28}).
\end{aligned}$$

We now bound $E[\Delta(S_2)]$, which is $E[\Delta(F_1)] - E[\Delta(B_1)]$. Because $B_1 = F_1 \cap (S_0 \cup S_1)$, $E[\Delta(B_1)] \leq E[\Delta(S_0) + \Delta(S_1)] = \Theta(n^{19/28})$. Thus,

$$\begin{aligned}
E[|\varDelta(S_2)|] &= E[|\varDelta(C_1)|] - E[|\varDelta(B_1)|] \\
&= \left(\sum k \left[1 - \left(1 - \frac{k}{n}\right)^{n^{19/28}}\right] \frac{n}{k^{11/4}}\right) - E[|\varDelta(B_1)|] \\
&= \varTheta\left(\int \left[1 - \left(1 - \frac{k}{n}\right)^{n^{19/28}}\right] \frac{n}{k^{7/4}}\right) - E[|B_1|] \\
&= \varTheta\left(\int_1^{n^{9/28}} \left[n^{19/28}\left(\frac{k}{n}\right)\right] \frac{n}{k^{7/4}}\right) - E[|B_1|] \\
&= \varTheta(n^{85/112}) - \varTheta(n^{19/28}) = \varTheta(n^{85/112}).
\end{aligned}$$

Note that S_2 still has a polynomial expansion. So S_3 will continue to grow. As these bounds a highly concentrated, by repeating this argument a constant number of times, to be formalized in the next subsection, we can show that the expected number of vertices reachable from $\{1, 2\}$ is $\varOmega(n^{1-\epsilon})$ for any $\epsilon > 0$.

4.2 Crawling with Buffer of Size Constant Fraction of n

We now consider the case when the buffer is too small to contain all vertices of in-degree 1. Let h be an integer such that the buffer is large enough to contain all vertices of in-degree at least $h - 1$. We will use the following structure to establish a lower bound on the core size: Let $S_0 = [1 : h]$. Let the h-PYRAMID of S_0, denoted by PYRAMID(S_0), be the following subgraph. For each i, let

$$S_i = \text{NEIGHBORS}(S_{i-1}) - \cup_{j=1}^{i-1} S_j.$$

Then, PYRAMID(S_0) is the subgraph induced by $\cup_i S_i$.

We will use the following lemma whose proof is straightforward.

Lemma 5. *Suppose $G = (V, E)$ is a directed graph and S_0 is a subset of V of size b. If there is a t_0 such that S_{t_0} contains a subset C_0 satisfying that the indegree of every vertex in C_0 in the induced subgraph $G(C_0)$ over C_0 is at least h, then* PYRAMID(S_0) *is in the core if b is larger that the number of vertices in G whose indegrees are h or more.*

Below, we will show the h highest degree vertices form a clique. Furthermore, if we start with a random set of b vertices, then with high probability, these h vertices get in the buffer in the first step and will remain there. Again, we will focus on $\alpha = 11/4$. Let CLIQUE(h) be the event that the subgraph induced by $[1 : h]$ is a complete direct graph. We use $[A]$ to denote that an event A is true.

Lemma 6. *In a random (α, n)-BBPL graph G with $\alpha = 11/4$,*

$$Pr[[\text{CLIQUE}(h)]] \geq 1 - e^{-\frac{n^{1/7}}{h^{8/7}}}.$$

Proof. The expected degree of vertex i is $\Theta\left(\left(\frac{n}{i}\right)^{4/7}\right)$. By a union bound,

$$Pr[[\text{not CLIQUE}(h)]] \leq \sum_{i,j \leq h} \left(1 - \frac{(\frac{n}{i})^{4/7}}{n} \frac{\left(\frac{n}{j}\right)^{4/7}}{n}\right)^n \leq \sum_{i,j \leq h} e^{-\frac{n^{1/7}}{(ij)^{4/7}}} \approx e^{-\frac{n^{1/7}}{h^{8/7}}}$$

□

Note that if $h < n^{1/8}$, then with high probability, $[1 : h]$ induces a complete directed clique.

Lemma 7. *Let $G = (V, E)$ be a random (α, n)-BPPL graph, for $2 < \alpha < 3$. With high probability, there exists a constant c, such that for any h (not necessarily a constant), for a set of $b \geq cn^{1-1/(\alpha-1)} h^{1+1/(\alpha-1)} \log n$ randomly chosen vertices S, the buffer-induced in-degrees of vertices 1,..., h are larger than h.*

Proof. Note that at least 1, we have $\sum_{u \in S} \mathbf{d}_{\alpha,n}[u] \geq b$. The expected in-degrees of vertices 1,..., h are bounded by

$$\Theta\left(\frac{n^{1/(\alpha-1)}}{n}\right) \sum_{u \in S} \mathbf{d}_{\alpha,n}[u] = \left(\frac{n}{h}\right)^{1/(\alpha-1)} \frac{b}{n} \geq ch \log n.$$

As this bound is highly concentrated, thus if c is large enough, with high probability (e.g., $1 - n^{-\Theta(c)}$), the buffer-induced in-degrees of vertices 1,...,h are at least h. □

Lemma 8. *In a random BPPL graph with parameters n and $\alpha = 11/4$, the total expected degrees of vertices $[1 : h]$ is $\Theta\left(n^{4/7} h^{3/7}\right)$.*

Proof. $\sum_{x=1}^{h} \left(\frac{n}{x}\right)^{4/7} = \Theta(n^{4/7} h^{3/7})$. □

We now analyze the size of pyramid.

Lemma 9. *Let T_0 be a subset of vertices of $[1 : n]$ whose expected total degree is $\Theta(n^\gamma)$ and $|T_0| = o(n^\gamma)$, for a constant γ. Suppose T_0 has n^γ random outgoing edges chosen according to the BPPL distribution with parameters n and $\alpha = 11/4$. Let $T_1 = \text{NEIGHBORS}(T_0) - T_0$. Then, for any constant $h \geq 2$,*

$$E[|T_1|] \geq \Theta\left(n^{(7/4)\gamma - 3/4}\right).$$

Proof. Let $T_1' = \text{NEIGHBORS}(T_0)$. So, $T_1 = T_1' - T_0$. For $k \leq n^{1-\gamma}$ and $h \ll n^\gamma$, the probability that a weight k vertex receives at least h edges from T_0 is

$$\binom{n^r}{h} \left(\frac{k}{n}\right)^h = \Theta\left(\min\left((n^\gamma)^h \left(\frac{k}{n}\right)^h, 1\right)\right) \quad (1)$$

Thus, as $h \geq 2$, the expected size of T_1 is

$$\begin{aligned} E[|T_1|] &= E[|T_1'|] - E[|T' \cap T_0|] \\ &= \Theta\left(\sum_1^{n^{1-\gamma}} \frac{n}{k^{11/4}} \cdot n^{\gamma h} \left(\frac{k}{n}\right)^h\right) - E[|T' \cap T_0|] \\ &= \Theta\left(n^{h\gamma - h + 1} \int_1^{n^{1-\gamma}} k^{h-11/4}\right) - o(n^\gamma) \qquad (2) \\ &= \Theta\left(n^{h\gamma-h+1} n^{(1-\gamma)(h-7/4)}\right) = \Theta\left(n^{(7/4)\gamma - 3/4}\right). \qquad \square \end{aligned}$$

This bound is independent of h. If $\gamma = 4/7$, then $(7/4)\gamma - 3/4 = 1/4$. We would like to remark that if $h = 1$, then the calculation follows from what we did in the previous subsection. The integral there was a constant but is not here. The next lemma bounds $\Delta(T_1)$, the expected total degrees of vertices in T_1.

Lemma 10. *Let T_0 be a subset of vertices of $[1:n]$ whose total expected degrees is $\Theta(n^\gamma)$ and $|T_0| = o(n^\gamma)$, for a constant γ. Suppose T_0 has n^γ random outgoing edges chosen according to the BPPL distribution with parameters n and $\alpha = 11/4$. Let $T_1 = \text{NEIGHBORS}(T_0) - T_0$. Then, for any constant $h \geq 2$,*

$$E[\Delta(T_1)] \geq \Theta\left(n^{(3/4)\gamma + 1/4}\right).$$

Proof. Let $T_1' = \text{NEIGHBORS}(T_0)$. We have $T_1 = T_1' - T_0$.

$$\begin{aligned} E[\Delta(T_1)] &= E[\Delta(T_1')] - E[\Delta(T_1' \cap T_0)] \\ &\geq \Theta\left(\sum_1^{n^{1-\gamma}} k\frac{n}{k^{11/4}} \cdot n^{\gamma h} \left(\frac{k}{n}\right)^h\right) - |\Delta(T_0)| \\ &= \Theta\left(n^{h\gamma-h+1} \int_1^{n^{1-\gamma}} k^{h-7/4}\right) - \Theta(n^\gamma) \qquad (3) \\ &= \Theta\left(n^{h\gamma-h+1} n^{(1-\gamma)(h-3/4)}\right) = \left(n^{(3/4)\gamma+1/4}\right). \qquad \square \end{aligned}$$

We now apply Lemmas 9, 10 and 6, to prove the following theorem. Because we need to apply these lemmas iteratively, we need to know the concentration of the bounds in Lemmas 9 and 10. Recall the the original Hoeffding bounds states that if X_i are independent random variables in $[0, 1]$ (not necessarily binary) and $S = \sum X_i$, then

$$\Pr[S > (1+\lambda)E] \leq e^{-\lambda^2 \mathrm{E}[S]/2} \qquad (4)$$

$$\Pr[S < (1-\lambda)E] \leq e^{-\lambda^2 \mathrm{E}[S]/3}. \qquad (5)$$

In Lemma 9, the bound of $\mathrm{E}[|T_1|]$ is the sum of random 0 and 1 variables. We use the standard Chernoff bound to show that the sum is exponentially concentrated, i.e., with probability $1-e^{-n^{\Theta(1)}}$. The bound of $\mathrm{E}\left[\Delta(T_1)\right]$ in Lemma 10 is no longer the sum of random 0/1 variables or random variables whose value is in the range of $[0,1]$. We need the following restatement of Heoffding bound: If X_i are independent random variables in $[0,A]$ and $S=\sum X_i$, then

$$\Pr\left[S>(1+\lambda)E\right] \le e^{-\lambda^2 \mathrm{E}[S/A]/2} \tag{6}$$

$$\Pr\left[S<(1-\lambda)E\right] \le e^{-\lambda^2 \mathrm{E}[S/A]/3}. \tag{7}$$

To obtain a concentration bound, we observe that k in Equation (3) is in the range of $[1:n^{1-\gamma}]$. Thus, the bound in Equation (3) is the sum of random variables in range $[1:n^{1-\gamma}]$. So, as long as $n^{1-\gamma} \ll n^{(3/4)\gamma+1/4}$, we can use this restatement of Hoeffding bound to an $1-e^{-n^{\Theta(1)}}$ concentration. In our argument below that uses Lemma 10, we will have $\gamma \ge 4/7$. Thus, all our bounds are exponentially concentrated.

Theorem 1 (Size of Pyramid: h is a constant). *Let $G=(V,E)$ be a random (α,n)-BPPL graph with $\alpha=11/4$. For any constant h, let S_0 be a random set of size b, where $b=\Theta\left(\frac{n}{h^{\alpha-1}(1+1/(2h))}\right)$. Then, for any constant $\epsilon>0$, the expected size of* PYRAMID(S_0) *is $\Theta(n^{1-\epsilon})$.*

Proof. Because S_0 is a random set of b elements, by Lemma 7, with high probability, S_1 contains $[1:h+1]$. Let $\gamma=4/7$ and $\beta=3/7$, i.e., $\gamma=1-\beta$. By Lemmas 9 and 10, we have that the expected value of $|S_2|$ and $\Delta(S_2)$ are

$$\left[\Theta\left(n^{1-\frac{7}{4}\beta}\right),\Theta\left(n^{1-\frac{3}{4}\beta}\right)\right] \tag{8}$$

By iteratively applying this analysis, for any constant t, the expected values of $|S_t|$ and $\Delta(S_t)$ are

$$\left[\Theta\left(n^{1-\frac{7}{4}\left(\frac{3}{4}\right)^{t-1}\beta}\right),\Theta\left(n^{1-\left(\frac{3}{4}\right)^{t}\beta}\right)\right]$$

Moreover, these random variables are highly concentrated. Thus, for $t=\lceil\log_{4/3}\epsilon\rceil$, we have $\mathrm{E}\left[|\mathrm{PYRAMID}(S_0)|\right] \ge \mathrm{E}\left[|S_t|\right]=\Theta(n^{1-\epsilon})$. □

By Lemma 1, if we set buffer size $b=\Theta\left(\frac{n}{h^{\alpha-1}(1+1/(2h))}\right)$, with a sufficiently large constant, every vertex that receives at least h votes will be in the buffer.

Theorem 2. *For any constants $2<\alpha<3$, $0<c<1$, and $\epsilon>0$, with high probability, our crawling process with buffer size $b=cn$ starting on a randomly chosen set S_0 of vertices of a random (α,n)-BPPL graph G has a core of expect size $\Theta(n^{1-\epsilon})$.*

4.3 As Buffer Becomes Even More Smaller

When h is a function of n, e.g., $h=n^{\delta}$, we need to be a little more careful. But, our analysis can still be extended to establish the following theorem similar to Theorem 1.

Theorem 3 (*h*: $h = n^{\Theta(1)}$)**.** *Let $G = (V, E)$ be a random (α, n)-BPPL graph with $\alpha = 11/4$. For any $\delta \leq 1/8$ and $\epsilon > 0$, letting $h = n^{\delta}$, if our crawling process starting with a random set S_0 of size $b = \Theta\left(\frac{n}{h^{\alpha-1}(1+1/(2h))}\right)$ has a core with expect size $\Theta(n^{1-\frac{7}{4}\delta-\epsilon})$.*

Proof. The main difference is that if $h = n^{\delta}$, then the estimation of the probability that a weight k vertex receives at least h edges from S_0 is

$$\Theta\left(\left(\frac{en^{\gamma}}{h}\right)^{h}\left(\frac{k}{n}\right)^{h}\right) = \Theta\left(\left(en^{\gamma-\delta}\right)^{h}\left(\frac{k}{n}\right)^{h}\right) \tag{9}$$

instead of Equation (1) used in the proof of Lemma 9. With the help of this bound, the bound of Equation of 2 becomes

$$E[|S_1|] = E[|S_1'|] - E[|S' \cap S_0|] = \Theta\left(n^{(7/4)\gamma-\frac{3}{4}\delta-3/4}\right), \tag{10}$$

and the bound of Equation of 3 becomes

$$E[\Delta(S_1)] = E[\Delta(S_1')] - E[\Delta(S_1' \cap S_0)] = \Theta\left(n^{(3/4)\gamma-\frac{7}{4}\delta+1/4}\right). \tag{11}$$

Applying these bounds in the analysis of Theorem 1, setting $\gamma = 1 - \beta = 4/7$, the expected value of $|S_2|$ and $\Delta(S_2)$ are

$$\left[\Theta\left(n^{1-\frac{3}{4}\delta-\frac{7}{4}\beta}\right), \Theta\left(n^{1-\frac{7}{4}\delta-\frac{3}{4}\beta}\right)\right]. \tag{12}$$

By iteratively applying this analysis, if $\delta \leq 1/8$ (which ensures that Proposition 6 holds), then for any constant t, the expected values of $|S_t|$ and $\Delta(S_t)$ are

$$\left[\Theta\left(n^{1-\frac{7}{4}\delta-\frac{7}{4}\left(\frac{3}{4}\right)^{t-1}\beta}\right), \Theta\left(n^{1-\frac{7}{4}\delta-\left(\frac{3}{4}\right)^{t}\beta}\right)\right].$$

Again, these random variables are highly concentrated. Thus, the core is at least $n^{1-(7/4)\delta-\epsilon}$ for all $\epsilon > 0$, i.e., for large enough t, the expected values of $|T_t|$ and $\Delta(T_t)$ are

$$\left[\Theta\left(n^{1-\frac{7}{4}\delta-\epsilon}\right), \Theta\left(n^{1-\frac{7}{4}\delta-\epsilon}\right)\right].$$

□

4.4 Discussion

First of all, in our proof, we in fact consider the graph generated by the BBPL process and remove the multiple edges and self-loops. If we use the self-loops and multiple edges, we can further simplify the proof by starting with vertex 1 only, because its self-loop contribution is sufficient to keep it in the buffer. In other words, we do not need to start with an h-clique.

Our analysis can be easily modified to apply to the following family of random graphs: For vertex i, we add $\mathbf{d}_{\alpha,n}$ outward edges whose endpoints are chosen according to $\mathbf{d}_{\alpha,n}$. Again, in this model, we can remove self-loops and multiple edges. All our lemmas and theorems can be extended to this model. The analysis can also be extended to the following model with exact in and out degree. Let A and B be the array of length $||\mathbf{d}_{\alpha,n}||_1$, in which there are $\mathbf{d}_{\alpha,n}(i)$ entries with value i. Now randomly permute A and B, and add a directed edge from $A(i)$ to $B(i)$. Again, this graph may have multiple edges and self loops.

5 Final Remarks on Experiments and Future Directions

This paper is a step towards modeling web processing with limited space and time. Its objective is to provide some theoretical intuition indicating why table cores of non-trivial size exist. However, the models we consider here, both in terms of the crawling process and in terms of the graphical models, are in some respects unlike these usually encountered in practice. We have conducted limited experiments with some other models of power law graphs, for example, as discussed in [4] as well as some segments of web graphs. These experiments have shown the existence of non-trivial stable cores.

As the next step of this research, we would like to extend our result to other more realistic power-law models. The following are a few examples. (1) This is a growth model. Start with one node and at time t do the following based on a uniform three-way coins: (i) add a new node and connect it from a link from the existing nodes according the out degree distribution (plus some constant); (ii) add a new node and connect it to a link from the existing nodes according the in degree distribution (plus some constant); and (iii) choose a vertex according to the out degree and a vertex according to in degree, and insert this edge. (2)Given two vectors, IN and OUT and an integer m. Repeat m times, at each time, choose a vertex according to the out degree and a vertex according to the in degree, and insert this edge. In this model, we would like to study the graph based on the properties of IN and OUT, such as, IN and OUT follows some kind of power law. For example, in this paper, we analyze a particular (IN, OUT) pair. We would like to analyze the process for a larger family of (IN, OUT) distributions. (3) Start with one node and at time t, insert one new vertex and three edges. One out of the new vertex and one into the new vertex, and of course, according the in or out degree. (4) Other models in Chung and Lu's book.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. 30, 107–117 (1998)
2. Bollobas, B., Riordan, O., Spencer, J., Tusnady, G.: The degree sequence of a scale-free random process. Random Structures and Algorithms 18, 279–290 (2001)
3. Castillo, C.: Effective Web Crawling, Ph.D. Thesis, University of Chile (2004)
4. Chung, F., Lu, L.: Complex Graphs and Networks. AMS (2007)
5. Faloutsos, C., Faloutsos, M., Faloutsos, P.: On power-law relationships of the internee topology. In: Proc. SIGCOMM (1999)