Due: 2018/04/13

Homework 6

Problem 1. (20 points) If the number of pages with in-degree i is proportional to $\frac{1}{i^{2.1}}$, what is the probability that a randomly chosen web page has in-degree 1?

Solution:
The sum $S_i = \sum \frac{1}{i^{2.1}}$ is a version of Riemann's Zeta Function
$Z(x) = \sum \frac{1}{i^x}$ with i going from 0 to infinity.
The function Z(x) has known properties:
Z(0) = -$\frac{1}{2}$
Z(1/2) = -1.4603...
Z(1) = infinity (a harmonic series known to diverge)
Z(3/2) = 2.612 (Bose-Einstein condensate)
Z(2) = 1.645
Z(5/2) = 1.341
Z(3) = 1.202 (Apery's constant)
Now, setting x = 2.1 we have our function $\frac{1}{i^{2.1}}$. The sum resides somewhere between 1.645 and 1.341.
Another proof is solution is let the number of pages with in-degrees $i = N(i) = \frac{k}{i^{2.1}}$, where k is a proportionality constant.
p as a random chosen page has in-degree 1 $= \frac{N(1)}{\sum_{i-1}^{\infty} N(i)} = \frac{k}{k.\xi(2.1)} = \frac{1}{1.5602} = 0.641$
where $\xi(x)$ is the Riemann's zeta function.

`http://en.wikipedia.org/wiki/Riemann_zeta_function`

Problem 2. (20 points) Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. 35% of A's pages are present in B's index, while 55% of B's pages are present in A's index. What is the number of pages in A's index relative to B's?

Solution: $x\ |E_A| = y\ |E_B| \implies \frac{|E_A|}{|E_B|}$
$\frac{x}{y} = \frac{5}{3}$
Number of pages in A's index is relative to B = $\frac{5}{3}$

Problem 3. (10 points) Why is it better to partition hosts (rather than individual URLs) between the nodes of a distributed crawl system?

Solution:
It is better to partition hosts rather than individual URLs between the nodes of a distributed

crawl system as the host address usually has direct correspondence with the physical location of a host while the URLs may have nothing to do with it.
This allows for "politeness" measures to be implemented, preventing excessive calls to a particular machine in short periods of time, by caching data from one node. Consider a web hosting company with multiple websites, i.e. a pool of ip addresses. Grouping hosts by ip address could also support "politeness". On the other hands, for URLs we cannot say in advance where the corresponding physical machine (cluster) is located for a number of reasons. First of all, there are international domains which are in use throughout the world, i.e. .com, .org and .net. Secondly, most country code top-level domain registries allow its accredited registrars to sell the domains under their delegation to the residents of third countries. For example, it is possible to buy a domain in Belgian zone .be while being an American resident and not planning to use it for mostly Belgian users and keep in Belgium (i.e. http://youtu.be). Also, even if buying a domain in your national zone to host a website for the local community, it is sometimes better to keep a server abroad for the sake of savings, security and/or other reasons.
Thus, if one will partition the nodes of the distributed crawler by certain URLs, all the nodes will end up crawling the servers all over the world that will lead to decreased performance as opposed to the hosts-based partitioning.

Problem 4. (10 points) Why should the host splitter precede the Duplicate URL Eliminator?

Solution:
Host splitters can be cached, allowing the host splitter to discard duplicate URLs, thereby reducing network traffic. The same goes for DUE but the DUE requires significant RAM storage on each machine. Using the host splitter first, slows the growth rate of the cache and reduces redundant RAM usage.

Problem 5. (40 points) Web search engines A and B each crawl a random subset of the same size of the Web. Some of the pages crawled are duplicates − exact textual copies of each other at different URLs. Assume that duplicates are distributed uniformly amongst the pages crawled by A and B. Further, assume that a duplicate is a page that has exactly two copies − no pages have more than two copies. A indexes pages without duplicate elimination whereas B indexes only one copy of each duplicate page. The two random subsets have the same size before duplicate elimination. If, 45% of A's indexed URLs are present in B's index, while 50% of B's indexed URLs are present in A's index, what fraction of the Web consists of pages that do not have a duplicate?

Solution:
Let number of duplicates pages be D. Let $S(A)$ and $S(B)$ denote the sizes of the random subset to be indexed by A and B respectively. Given that
$S(A) = S(B)$
Both S(A) and S(B) contain $(\frac{D}{2})$ pages which have exactly one duplicate.
Number of pages indexed by A = $S(A)$

Number of Pages indexed by B = $S(B) - (\frac{D}{2})$
Hence,
$0.45 * S(A) = 0.5 * (S(B) - D(\frac{D}{2})$
Since S(A)=S(B), we get $D = \frac{S(A)}{5}$
Assuming that S(A) is the same as size of the web.
Fraction of web with duplicates = $\frac{1}{5}$
Fraction of web without duplicates = $\frac{4}{5}$