317 Ch 10

Note Title 2016-04-17

A common solution is to rank the pages in order of the number of links to that page (often called *backlinks* of the page), starting with the page that has the highest number of pointers into it. We refer to this strategy as **citation counting.**

Question: Suppose that we could determine the number of backlinks of each page (number of links pointing to the page). Why would that *not* necessarily be a good measure of the importance of the page?

Google's Solution: Google's solution is to define page rank recursively: "A page has high rank if the sum of the ranks of its backlinks is high." Observe that this covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks.

Question: It is easy to say that "a page has high rank if the sum of the ranks of its backlinks is high," but how does that help us figure out the rank of a page?

Answer: The "aha" that the Google founders made was to realize that the recursive definition is actually saying

$$\pi_j = \sum_{i=1}^{n} \pi_i P_{ij}.$$

Google's PageRank Algorithm:

1. Create a DTMC transition diagram where there is one state for each web page and there is an arrow from state i to state j if page i has a link to page j.
2. If page i has $k > 0$ outgoing links, then set the probability on each outgoing arrow from state i to be $1/k$.
3. Solve the DTMC to determine limiting probabilities. Pages are then ranked based on their limiting probabilities (higher probability first).

Example

Suppose the entire web consists of the three pages shown in Figure 10.1. Then the corresponding DTMC transition diagram is shown in Figure 10.2.

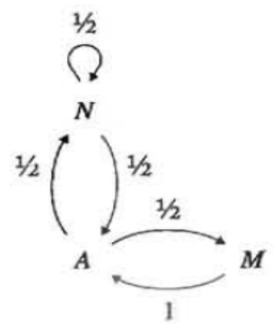


Figure 10.2. Corresponding DTMC transition diagram.

$$\pi_A = \frac{1}{2}\pi_N + 1\cdot\pi_M$$

$$\pi_N = \frac{1}{2}\pi_N + \frac{1}{2}\pi_A \Rightarrow \frac{1}{2}\pi_N = \frac{1}{2}\pi_A \Rightarrow \pi_N = \pi_A$$

$$\pi_M = \frac{1}{2}\pi_A$$

$$\pi_A + \pi_N + \pi_M = 1$$

$$\Rightarrow \pi_M = \frac{1}{5}, \ \pi_N = \pi_A = \frac{2}{5}$$

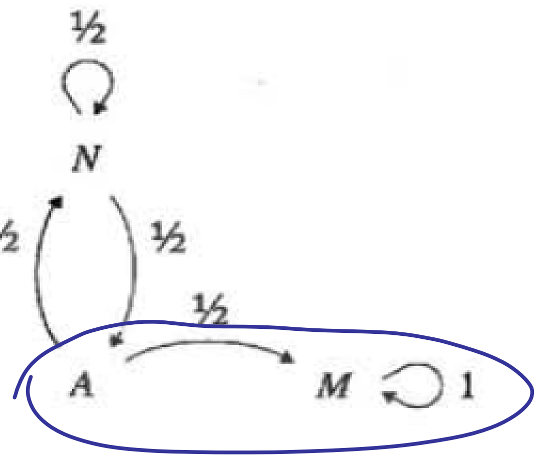
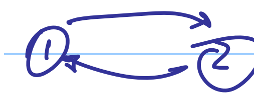


Figure 10.3. DTMC for a web graph with a dead end or spider trap at M.



$1 \cdot \pi_{12} = 2 \cdot \pi_{21}$

$$\begin{cases} \pi_N = \frac{1}{2}\pi_N + \frac{1}{2}\pi_A \\ \pi_A = \frac{1}{2}\pi_N \\ 0 \cdot \pi_M = \frac{1}{2}\pi_A \\ \pi_M + \pi_A + \pi_N = 1 \end{cases}$$

$\Rightarrow \pi_A = 0 \Rightarrow \pi_N = 0$

$\Rightarrow \pi_M = 1$

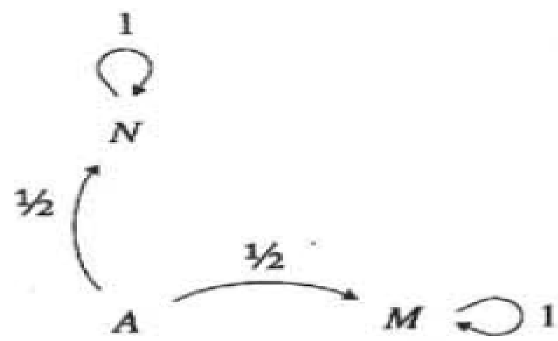


Figure 10.4. DTMC for a web graph with two spider traps.

Balance equs:

$$\begin{cases} 0 \cdot \pi_N = \frac{1}{2}\pi_A \\ 0 \cdot \pi_M = \frac{1}{2}\pi_A \\ \pi_A = 0 \end{cases} \Rightarrow \pi_N + \pi_M = 1$$

(infinitely many solutions)

$$\pi_N + \pi_M + \pi_A = 1$$

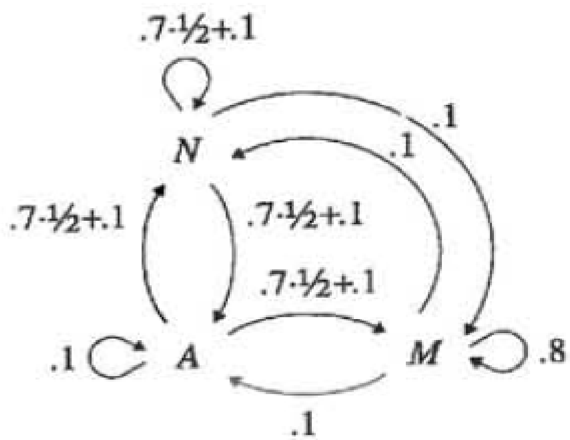


Figure 10.5. DTMC transition diagram for Figure 10.3 after 30% tax.

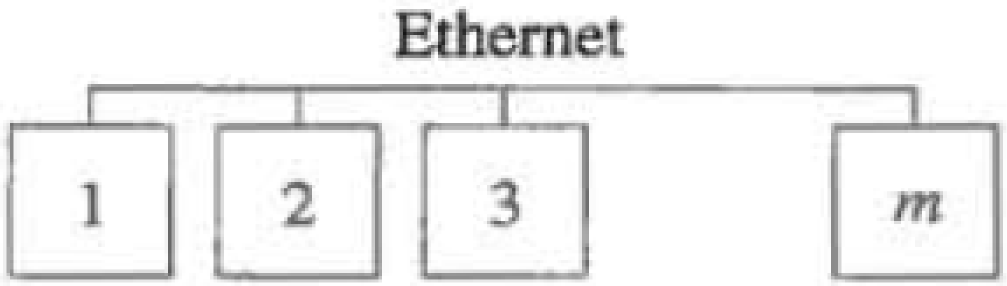


Figure 10.6. Ethernet with m hosts.

10.2.1 *The Slotted Aloha Protocol*

The Slotted Aloha protocol is defined as follows: Time is divided into discrete time steps or "slots." There are m transmitting hosts. At each time step, each of the m hosts independently transmits a new message (frame) with probability p (assume $p < \frac{1}{m}$).

If exactly 1 message (frame) is transmitted in a slot, the transmission is deemed "successful" and the message leaves the system. However, if more than 1 message is transmitted during a slot, the transmission is deemed "unsuccessful." In this case *none* of those messages leave the system. Every message involved in an "unsuccessful transmission" is then *retransmitted* at every step with probability q, until it successfully leaves the system. To keep things stable, we may need to make q very small; for the time being, assume that q is a very small constant. Note that, regardless of the backlog of messages, each of the m hosts continues to transmit *new* messages with probability p at each step.

$$p_k = \binom{m}{k} p^k (1-p)^{m-k}, \forall k = 0, 1, \ldots, m.$$

$$q_k^n = \binom{n}{k} q^k (1-q)^{n-k}, \forall k = 0, 1, \ldots, n.$$

$$P_{0,0} = (1-p)^m + mp(1-p)^{m-1}.$$
$$P_{0,1} = 0.$$
$$P_{0,j} = \binom{m}{j} p^j (1-p)^{m-j} \quad , \forall j = 2, \ldots, m.$$
$$P_{0,j} = 0 \quad , \forall j > m.$$

- $k \to k-1$: No new transmissions and 1 retransmission. The retransmitted message gets out of the system. Probability: $(1-p)^m kq(1-q)^{k-1}$.
- $k \to k$: There are several ways this can happen:
  1. One new transmission and no retransmissions:
     Probability: $mp(1-p)^{m-1}(1-q)^k$.
  2. No new transmission and no retransmissions:
     Probability: $(1-p)^m(1-q)^k$.
  3. No new transmission and at least 2 retransmissions:
     Probability: $(1-p)^m \left(1-(1-q)^k - kq(1-q)^{k-1}\right)$.
- $k \to k+1$: One new transmission and *at least* 1 retransmission. As a result, collision occurs, and the number of messages increases by 1.
  Probability: $mp(1-p)^{m-1}\left(1-(1-q)^k\right)$.

- $k \to k+j$, $j = 2, \ldots, m$: There are j new transmissions. The number of retransmissions does not matter, because collision occurs anyway.
  Probability: $\binom{m}{j} p^j (1-p)^{n-j}$.

$$
\begin{aligned}
P_{k,j} &= 0, \quad \forall j \leq k-2. \\
P_{k,k-1} &= (1-p)^m kq(1-q)^{k-1}. \\
P_{k,k} &= m(1-q)^k p(1-p)^{m-1} + \left(1 - kq(1-q)^{k-1}\right)(1-p)^m. \\
P_{k,k+1} &= mp(1-p)^{m-1}\left(1-(1-q)^k\right). \\
P_{k,k+j} &= \binom{m}{j} p^j (1-p)^{m-j}, \quad \forall j = 2, \ldots, m. \\
P_{k,j} &= 0, \quad \forall j > k+m.
\end{aligned}
$$