Methods and Models for the Web Graph

What does the web graph look like?

Studies:

- *Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, Wiener.*
  Used a web crawl from 1999 with 200 million nodes, 1.5 billion links.

- *Kumar, Raghavan, Rajagopalan, Tomkins.*
  Used a web crawl from 1997 with 40 million nodes.

- *Barabasi and Albert.*
  Used the web pages in domain `nd.edu`.

Degree distributions.

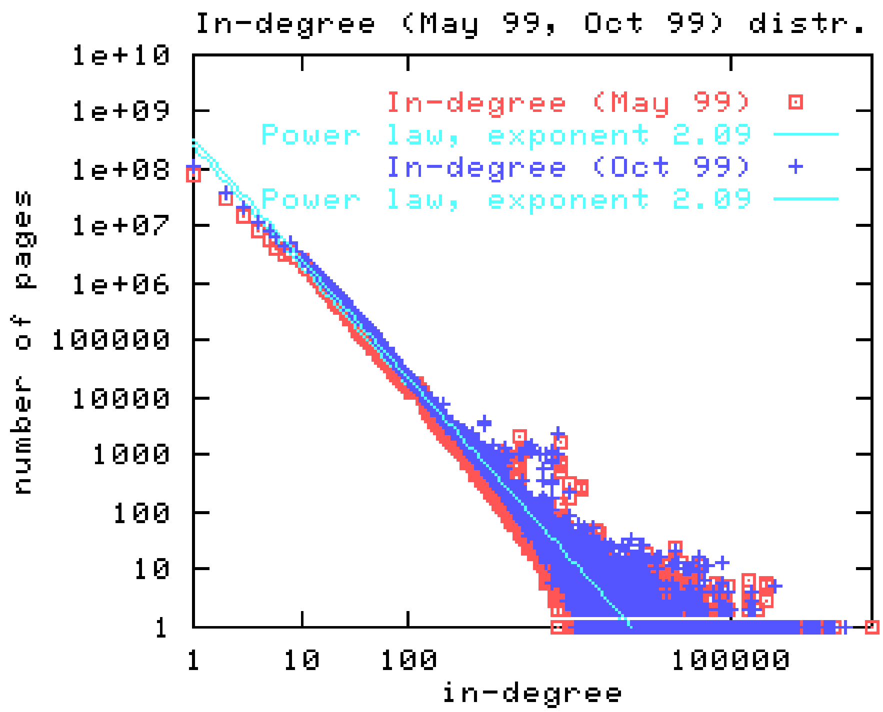
$P_i(k)$=proportion of nodes that have in-degree k.
$P_o(k)$=proportion of nodes that have out-degree k

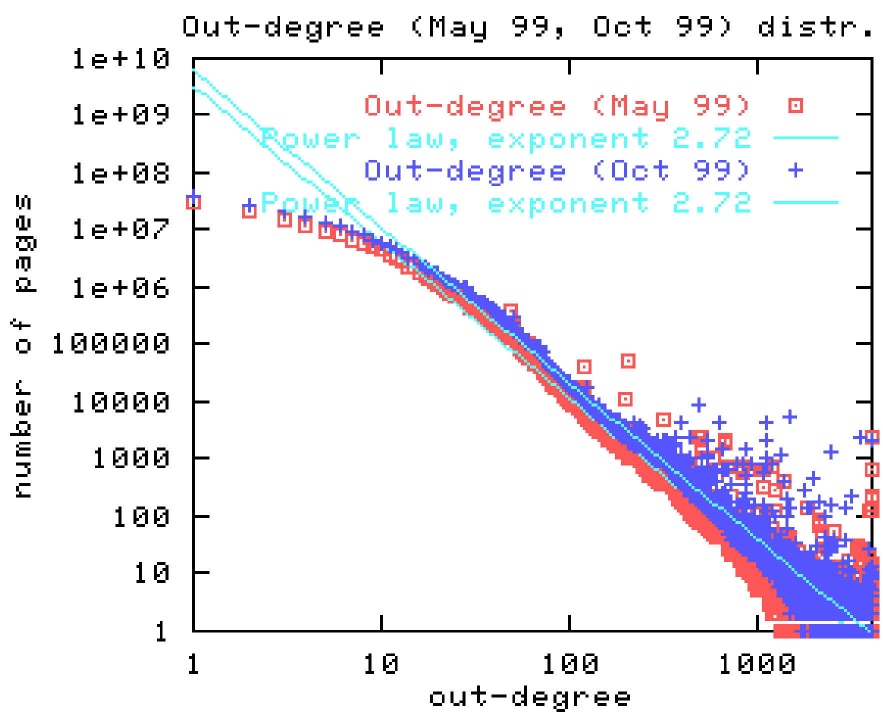
The in-degree distribution follows a power law:

$$P_i(k) \sim \frac{1}{k^t},$$

with exponent $t \approx 2.1$.

The tail of the out-degree distribution (for $k \geq 10$) also follows a power law, with exponent $t \approx 2.7$.


In-degree (May 99, Oct 99) distr.
In-degree (May 99)
Power law, exponent 2.09
In-degree (Oct 99)
Power law, exponent 2.09
number of pages
1e+10
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
100000
in-degree


Out-degree (May 99, Oct 99) distr.
Out-degree (May 99)
Power law, exponent 2.72
Out-degree (Oct 99)
Power law, exponent 2.72
number of pages
1e+10
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
1000
out-degree

Connectivity.

If the direction of the edges is ignored:

- 91% of the nodes are in one giant weakly connected component (=maximal connected subgraph)

- The sizes of the weakly connected components follow a power law distribution with exponent ≈ 2.5.

Robustness.

Size of largest weakly connected component when nodes with degree greater than k are removed:

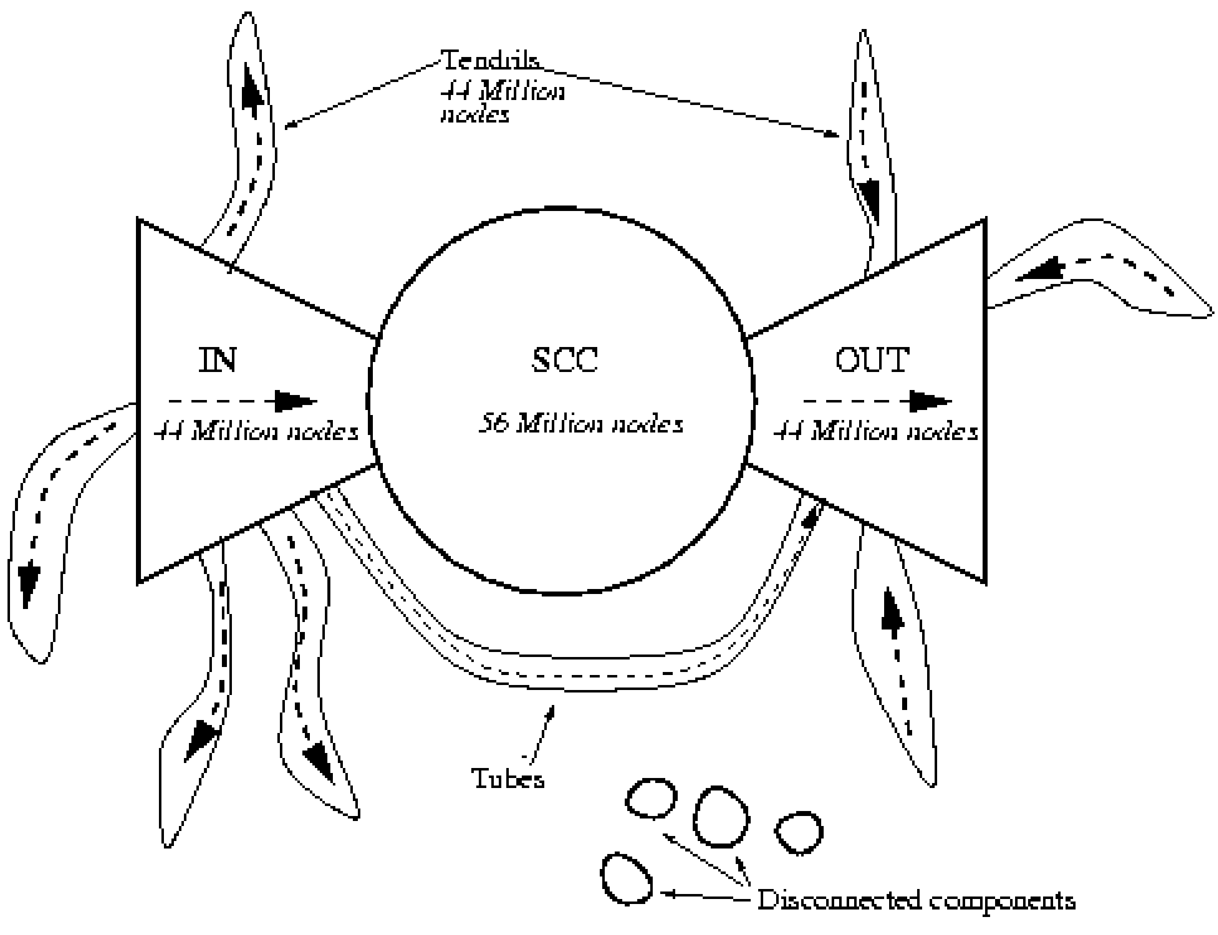
k	(none)	1000	100	10	5	4	3
Size (in millions)	182	177	167	105	59	41	15

Distance

- Only an estimated 24% of all node pairs are connected by a directed path.
- The average length of a directed path, if it exists, is estimated to be around 16.
- The maximum finite shortest directed path length could be as large as 905, and is at least 503.
- The average length of an undirected path is around 7.

Seen as an undirected graph, and restricted to the giant WCC, the web graph is a *small world network*.

The web is a bowtie:



Mathematical methods applied to the web graph

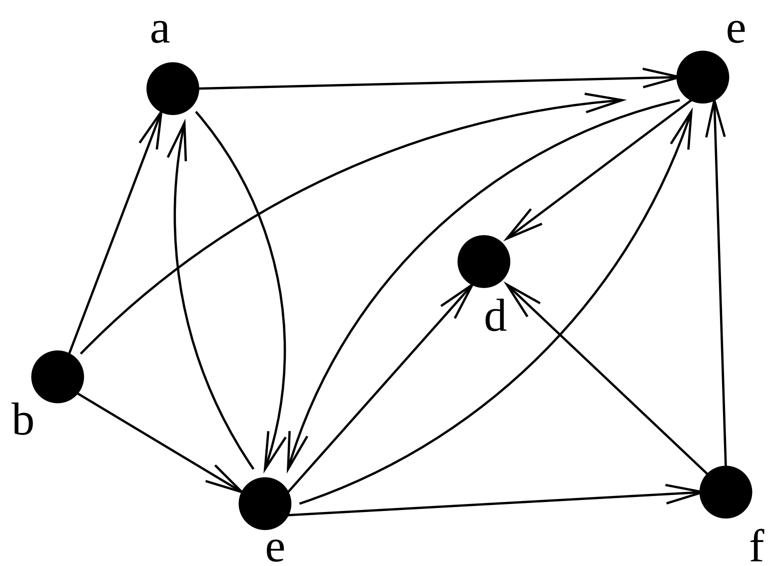
Ranking of search results

- PageRank (Brin and Page, '98)
  (used by Google)

- HITS (Kleinberg,'98)

Both based on eigenvectors

Example:



Adjacency matrix:

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Transition matrix:

$$A = \begin{bmatrix} 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 1/4 & 0 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Perron-Frobenius theorem

Let A be an irreducible $n \times n$ matrix with nonnegative entries. Then A has a positive real eigenvalue λ_1, and

- λ_1 has an associated eigenvector x_1 with strictly positive entries,
- λ_1 has algebraic multiplicity 1,
- If all main-diagonal entries of A are strictly positive, then all eigenvalues λ of A satisfy $|\lambda| \leq |\lambda_1|$.

Corollary:

If A is an irreducible $n \times n$ matrix with nonnegative entries and strictly positive main-diagonal entries, then for any vector x, $A^i x$ converges to αx_1.

PageRank p

$$p_i(v) = (1-\epsilon)\sum_{u \to v} p_{i-1}(u)\frac{1}{d^+(u)} + \epsilon \sum_u p_{i-1}(u),$$

where $d^+(u)$ is the out-degree of u.

Vector notation:

$$p_i = ((1-\epsilon)M^T + \epsilon U)p_{i-1},$$

Where M transition matrix, i.e. the adjacency matrix with each row normalized to 1, and U is a matrix with all entries equal to $1/n$ (n is the number of nodes).

HITS Algorithm

Hub vector h
Authority vector a

$$a_i(v) = \sum_{u \to v} h_{i-1}(u)$$
$$h_i(v) = \sum_{v \to u} a_i(u)$$

Vector notation:

$$a_i = A^T h_{i-1}$$
$$h_i = A a_i$$

Finding web communities

- Kleinberg (98): analysis of non-principal eigenvectors.
- Kumar, Raghavan, Rajagopalan, Tomkins (99): bipartite cores
- Flake, Lawrence, Giles (00): minimum cut
- An, J, Milios: communities in citation graph

What is a community?

Flake, Lawrence, Giles:
A set of nodes, so that the number of edges among these nodes exceeds the number of edges from these nodes to the rest of the graph.

A good web graph model should be:

- simple,
- plausible,
- dynamic,
- have a Power law or Zipfian in-degree distribution,
- have a locally dense, globally sparse structure.

(from Kumar et al.)

Web graph models

- Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, Upfal (00)
  Evolving graph model: copy with error

- Barabasi and Albert (00)
  Preferential attachment

Preferential attachment

- Start with a small graph.

- At each time step, create a new node, u, and draw its outdegree according to some predetermined distribution.

- For each edge originating from u, draw its destination according to a distribution proportional to the in-degree of the nodes, so the probability that the edge will have destination v is proportional to $1/d^{-}(v)$.

Graph evolution

- Start with a small graph.

- At each time step, create a new node, u.

- Choose a prototype node, v from the graph, uniformly.

- For each edge vw originating from v, do the following:

    With probability $1-\alpha$, create an edge with origin u and destination w,

    With probability α, draw a node w uniformly at random, and create an edge with origin u and destination w.