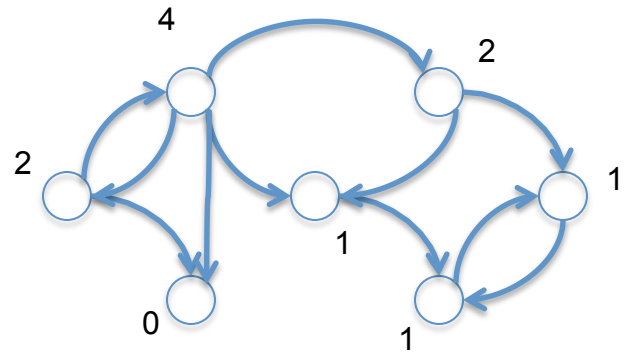
CSE 344, Lecture 25: Statistical Properties of the Web

Monday, May 22, 2011

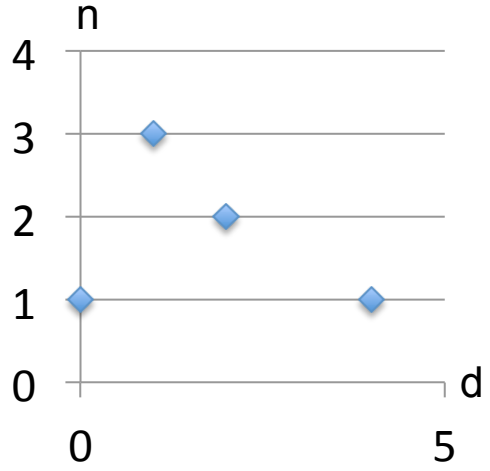
Based on: Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, Upfal: The Web as a Graph. PODS 2000

Histogram of a Graph

- Outdegree of a node = number of outgoing edges
- For each d, let n(d) = number of nodes with oudegree d
- The outdegree histogram of a graph = the scatterplot (d, n(d))

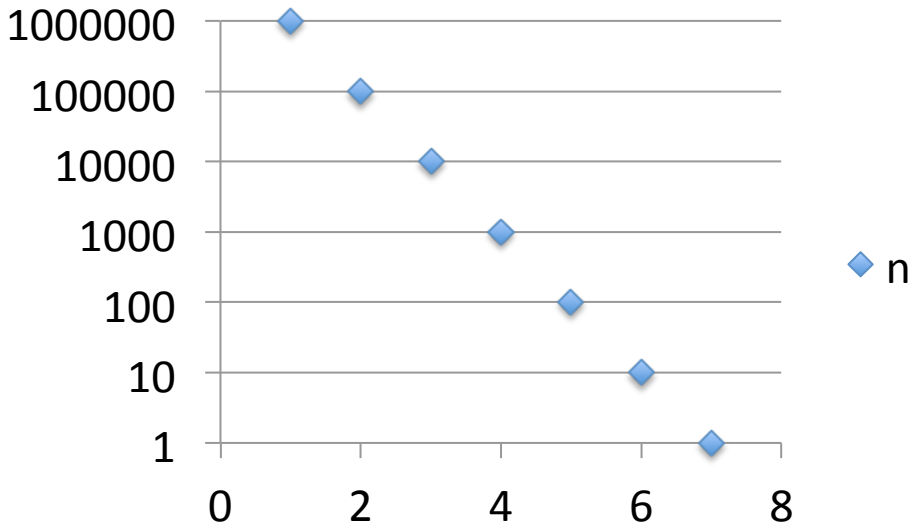


d	n(d)
0	1
1	3
2	2
3	0
4	1



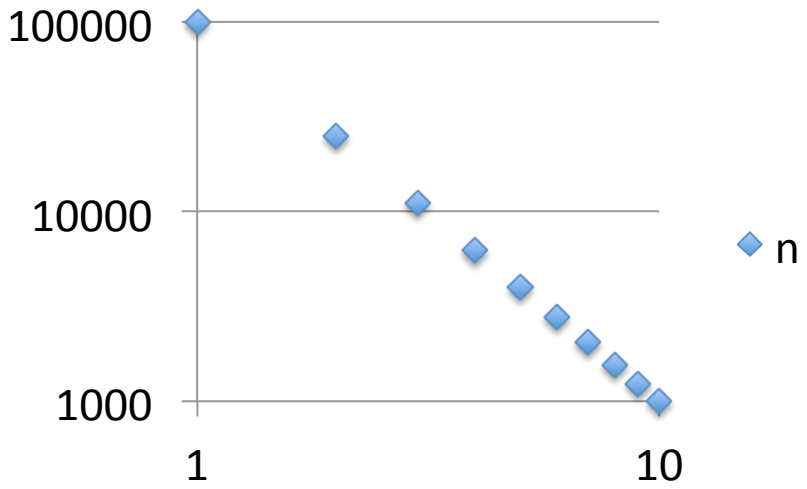
Exponential Distribution

- $n(d) \cong c/2^d$ (generally, cx^d, for some $x < 1$)
- A *random graph* has exponential distribution
- Best seen when n is on a log scale



Zipf Distribution

- $n(d) \cong 1/d^x$, for some value $x>0$
- Human-generated data has Zipf distribution: letters in alphabet, words in vocabulary, etc.
- Best seen in a log-log scale (why ?)



The Histogram of the Web

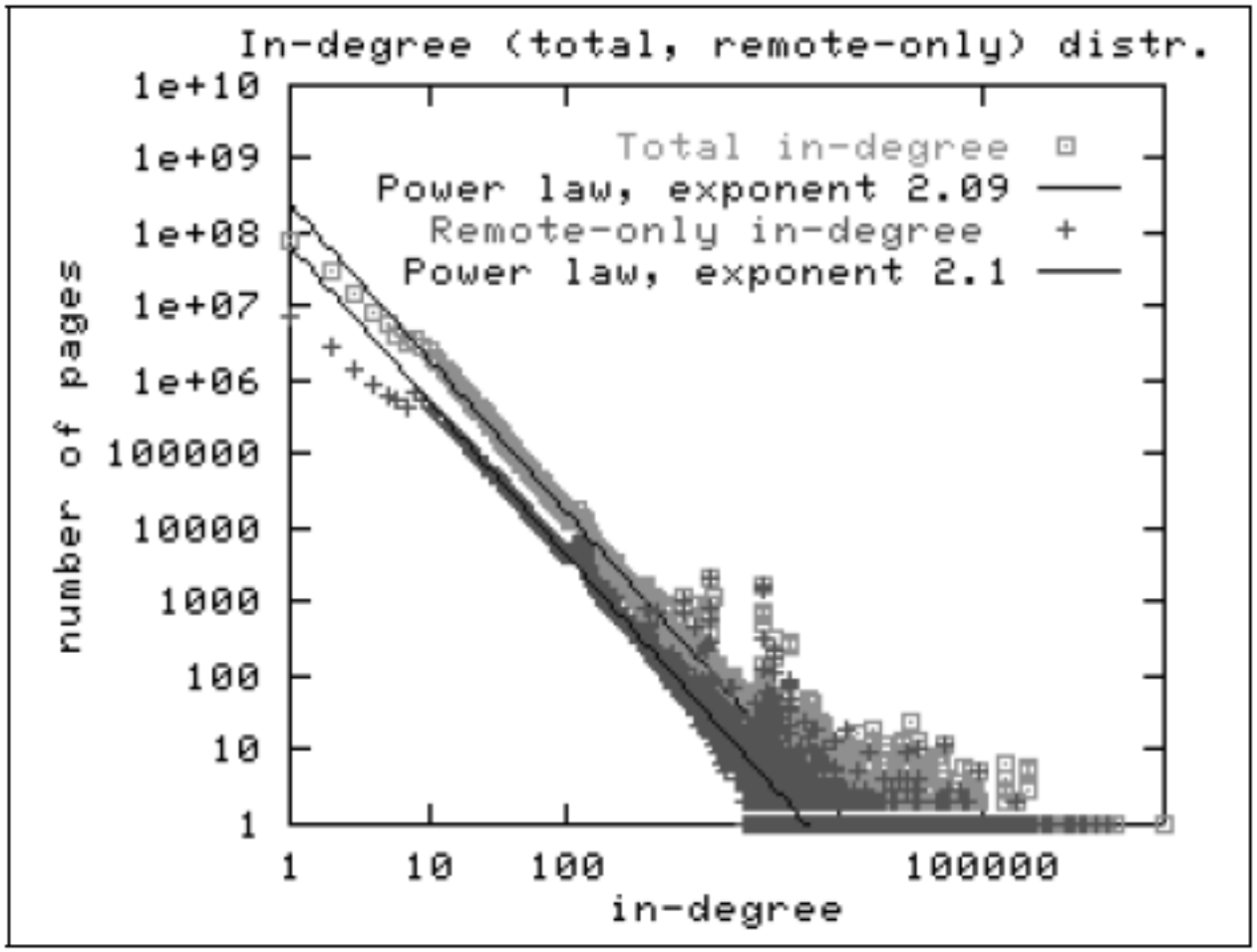


Figure 2: In-degree distribution.

Late 1990's
200M Webpages

Exponential ?

Zipf ?

The Bowtie Structure of the Web

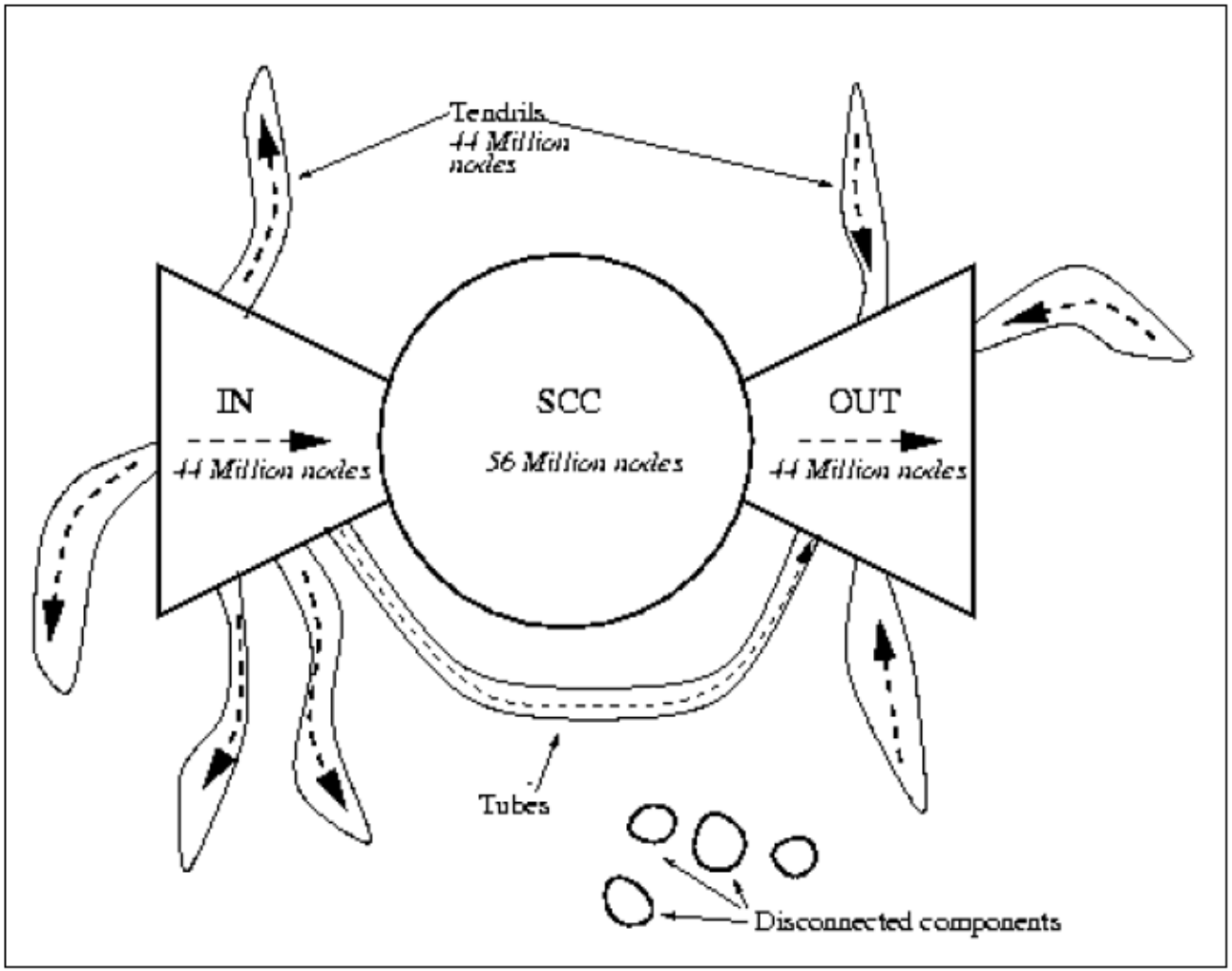


Figure 4: The web as a bowtie. SCC is a giant strongly connected component. IN consists of pages with paths to SCC, but no path from SCC. OUT consists of pages with paths from SCC, but no path to SCC. TENDRILS consists of pages that cannot surf to SCC, and which cannot be reached by surfing from SCC.