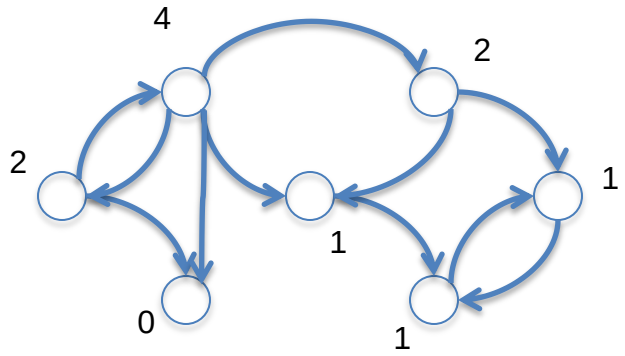
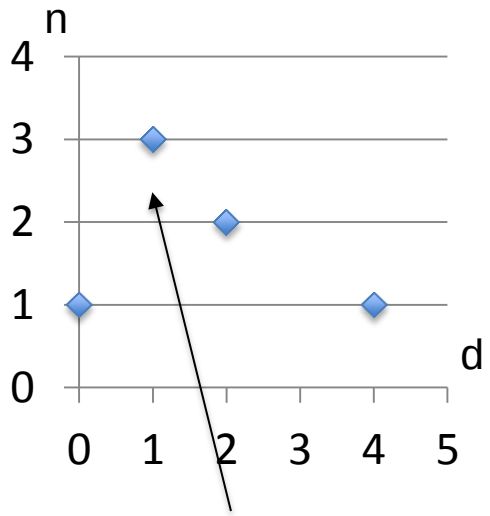
src: Dan Suciu

Structural Task: the Histogram of a Graph

- Outdegree of a vertex = number of outgoing edges
- For each integer d, let n(d) = number of vertices with outdegree d
- The outdegree histogram of a graph = the scatterplot (d, n(d))

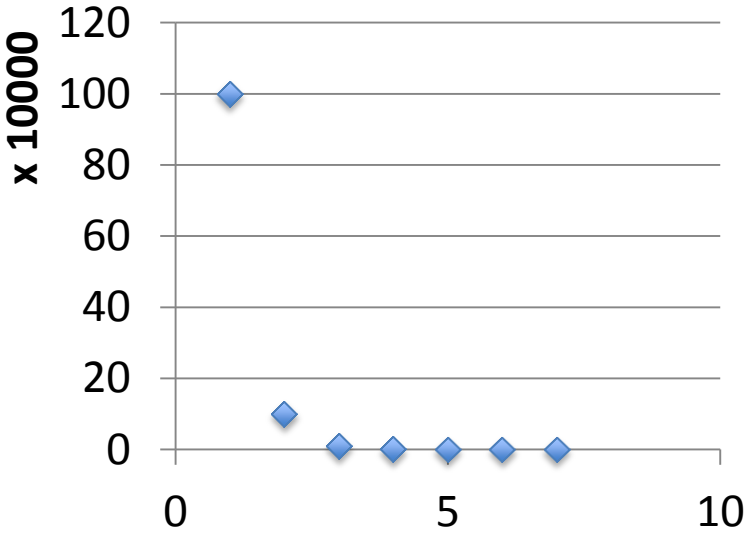


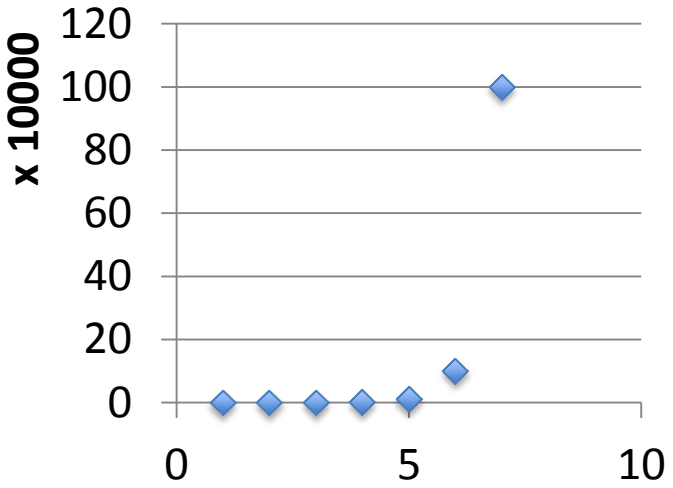
d	n(d)
0	1
1	3
2	2
3	0
4	1



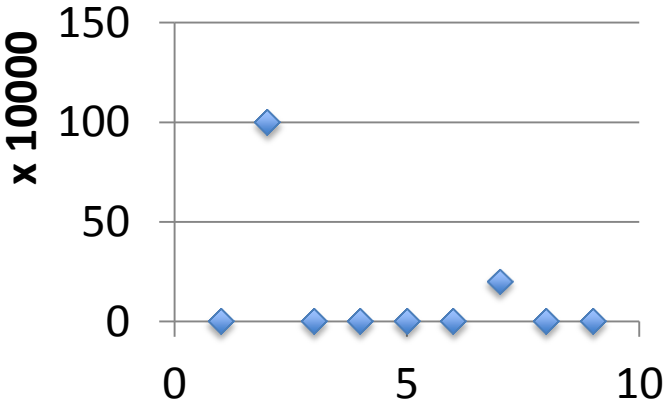
src: Dan Suciu

Histograms Tell Us Something About the Graph





What can you say about these graphs?

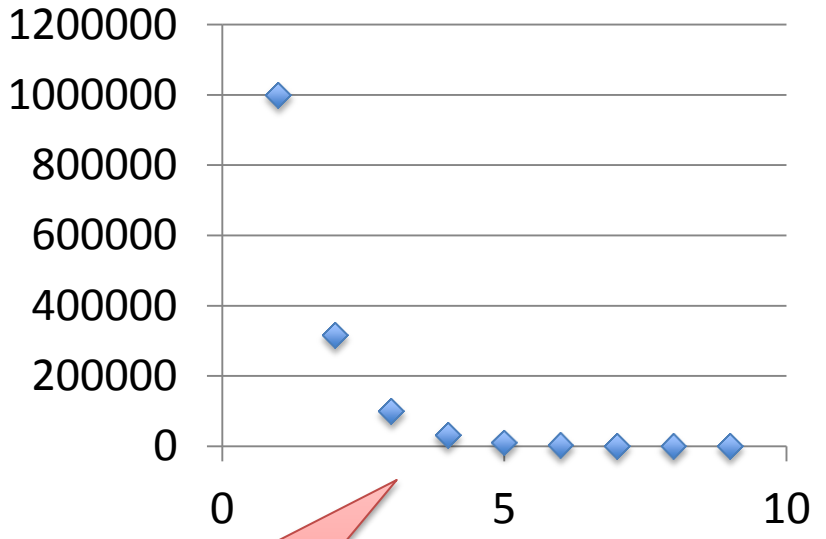


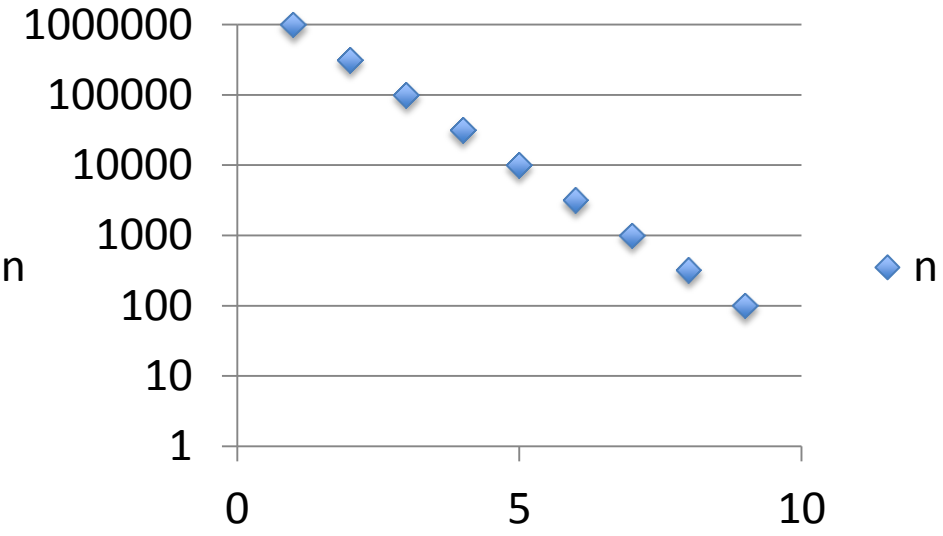
src: Dan Suciu

Exponential Distribution

- $n(d) \simeq c\left(\frac{1}{2}\right)^d$ (generally, cx^d, for some $x < 1$)

- A *random graph* has exponential distribution
- Best seen when n is on a log scale



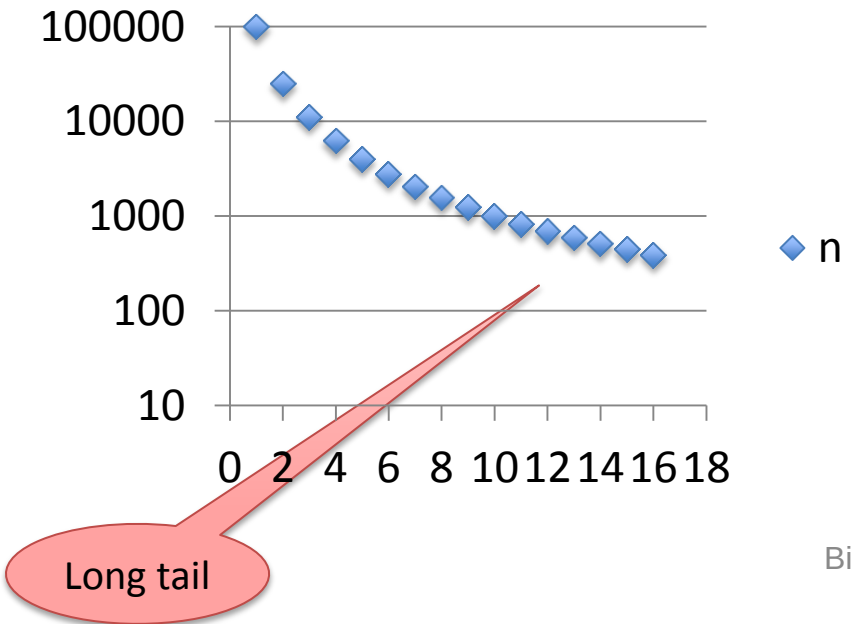


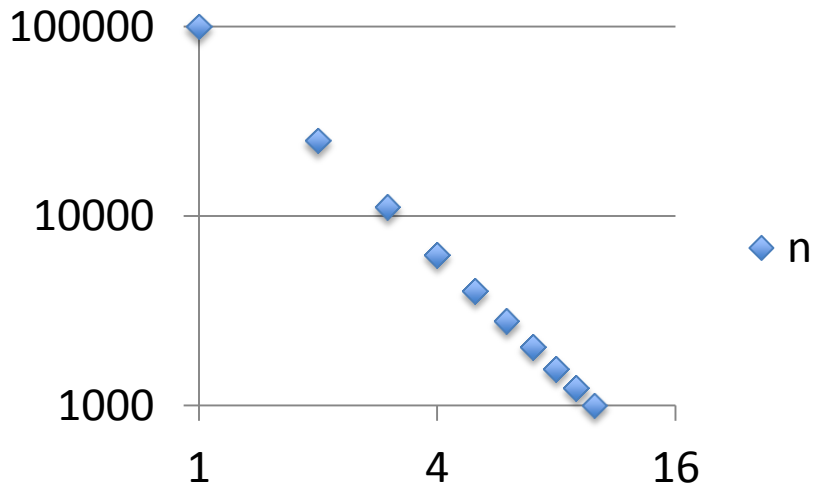
Quickly vanishing

src: Dan Suciu

Zipf Distribution

- $n(d) \simeq \frac{1}{d^x}$ 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





Broder et al. 2000

A Histogram of the Web

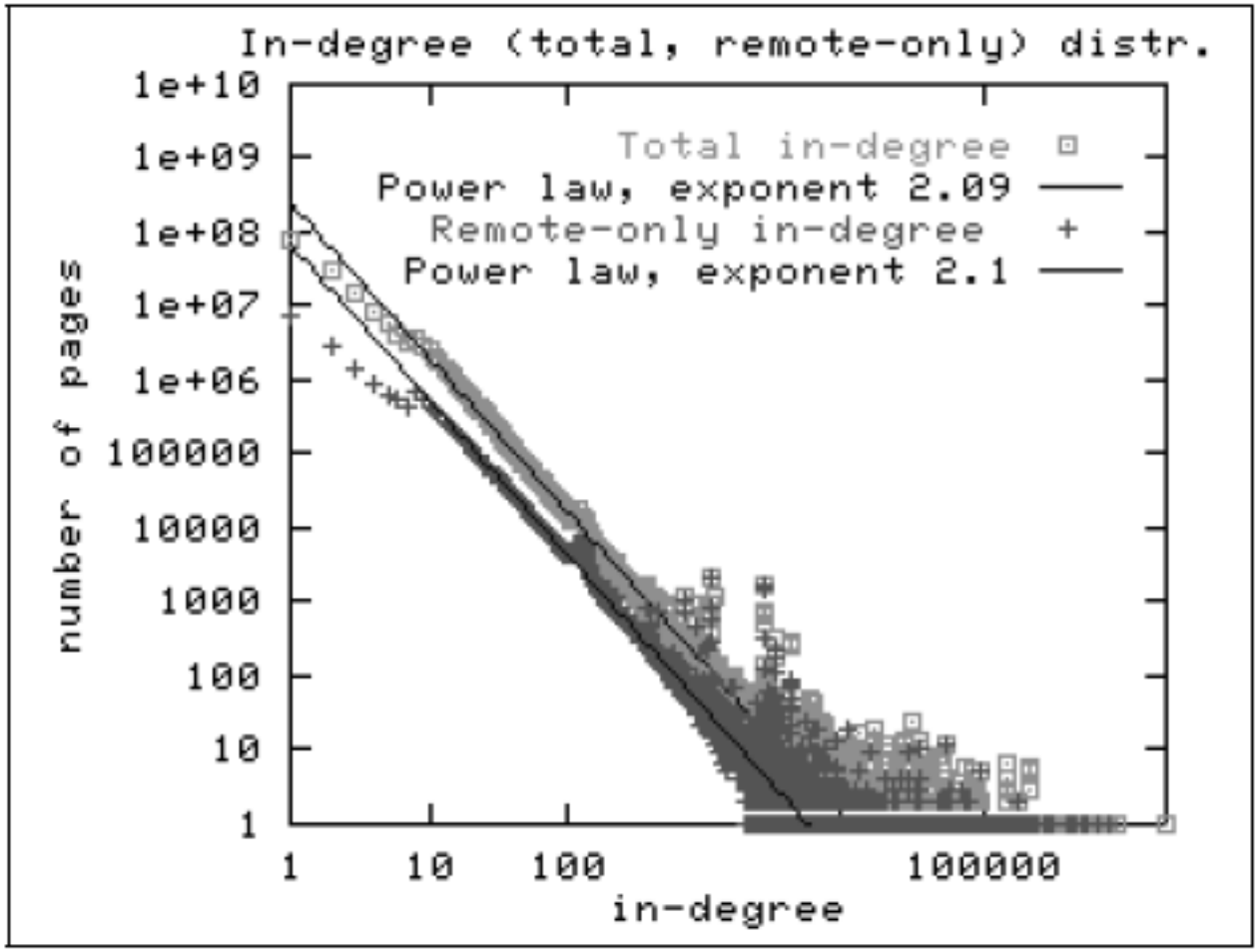


Figure 2: In-degree distribution.

Late 1990's
200M Webpages

Exponential ?

Zipf ?

Broder et al. 2000

The Bowtie Structure of the Web

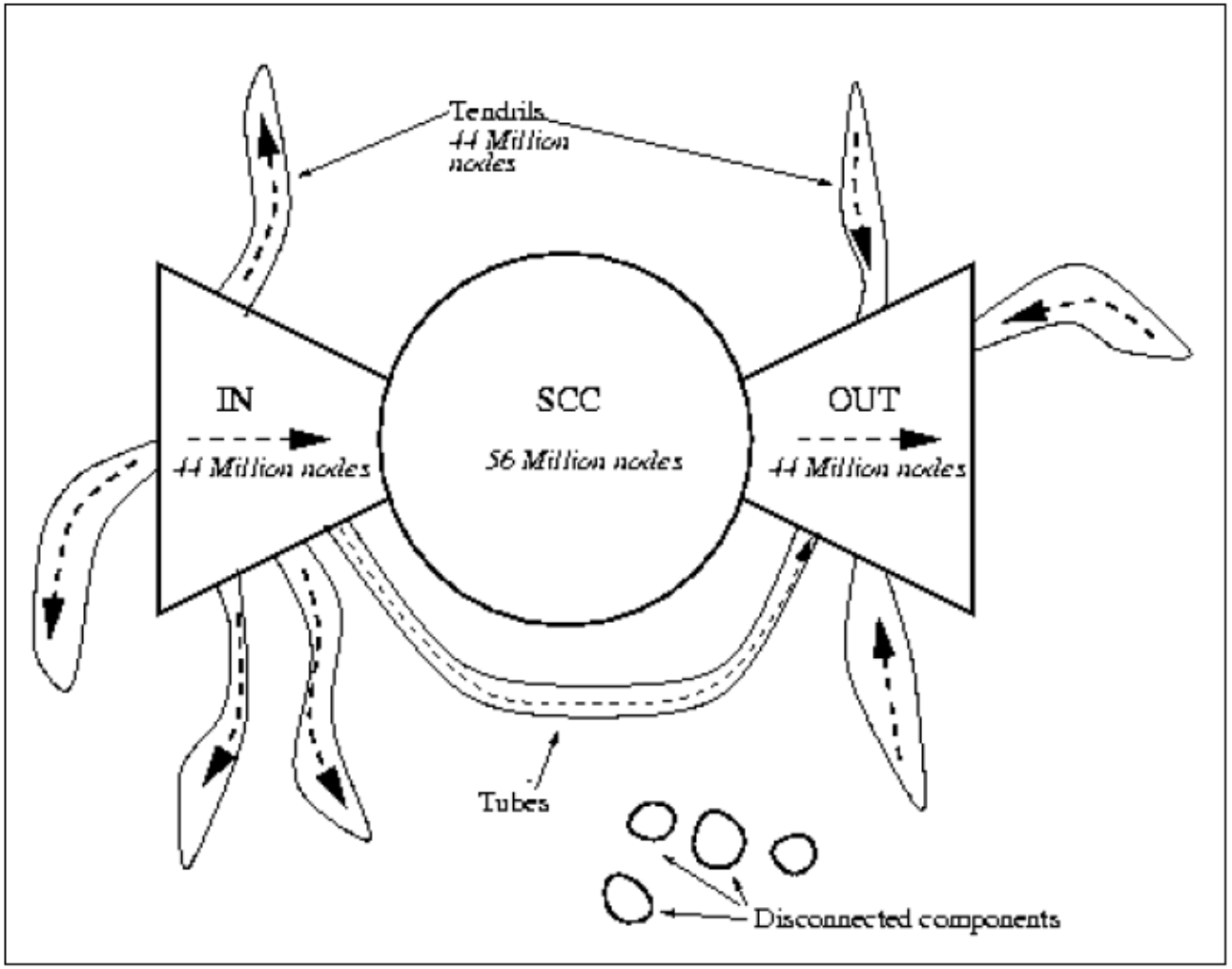


Figure 4: The web as a bowtie. SCC is a giant strongly connected component. IN consists of pages with paths to SCC, but no path from SCC. OUT consists of pages with paths from SCC, but no path to SCC. TENDRILS consists of pages that cannot surf to SCC, and which cannot be reached by surfing from SCC.