Graph models for the Web and the Internet

Elias Koutsoupias
University of Athens and UCLA

Crete, July 2003

Outline of the lecture

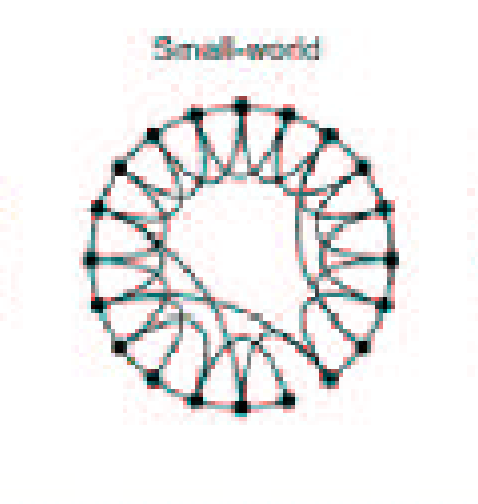
- Small world phenomenon
- The shape of the Web graph
- Searching and navigation
- Power law and similar phenomena
  - Experimental results
  - Erdös-Rényi classical model of random graphs
  - Graphs with prescribed degree-sequence
  - Models where power laws arise
- The Internet graph
  - Experimental results
  - Carson-Doyle's HOT
  - Fabrikant-K-Papadimitriou's model

Small-world phenomena

- In the late 60s Milgram noticed that social networks have small diameter: In an experiment, letters originating in Nebraska were sent to someone in Massachusetts via pairs of individuals that knew each other well (by first name). After 4-5 steps the letters reached their destination.
- The phenomenon gave rise to the claim "six-degrees of separation" (also a title of a popular play and film): Any two individuals of the planet are separated by a sequence of at most 6 "shake-hands".
- There are two "surprising" issues about the phenomenon:
  - The graph of acquaintances has **small diameter**.
  - There is a simple **local algorithm** that can route a message in a few steps.

The small-world model of Watts and Strogatz [1998]

A simple model to explain the small diameter of social, telephone, railroad, and other networks.



- Take a ring (circle) of n nodes in which every node is connected to the next k (say 2) nodes.
- Randomly re-wire each edge to a random destination with probability p.
- The resulting graph has logarithmic diameter with high probability:

$$diameter = O(\log n)$$

where $diameter = max_{u,v} distance(u, v)$

Clustering in networks

The **local clustering coefficient** at a vertex v is the fraction of the possible edges between neighbors of v that exist in the graph.
The **clustering coefficient of a graph** G:

$$C(G) = \frac{\text{number of triangles}}{\text{number of connected triples}}$$

Experiments suggest that social and other networks have

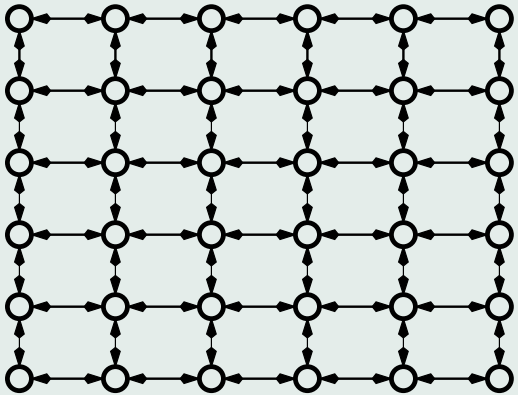
- large clustering coefficient
- small diameter

The model of Watts and Strogatz has both properties.
But do the real, engineered or self-governed, networks “look like” their simple model? Apparently not.

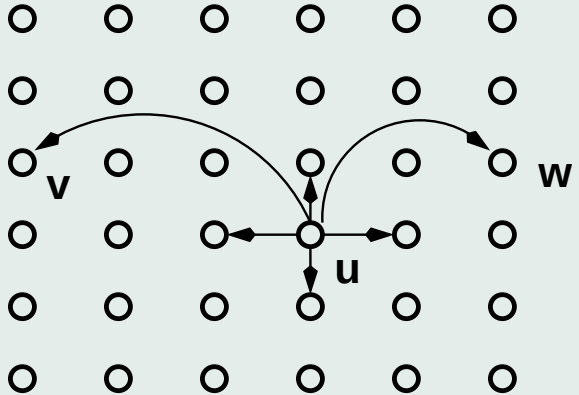
Kleinberg's model [2000]

Jon Kleinberg attempted to "explain" the small world phenomenon as follows:

A)



B)



Consider the 2-dimensional grid. For each node u, we add a "long" edge (u, v) to some node v selected with **probability proportional to** $[d(u, v)]^{-r}$, where $d(u, v)$ is the distance between u and v, and r is a parameter.

- $r = 0$: v is selected uniformly among all nodes
- $r = 2$: For every x, the probability that v is in distance between x and $2x$ is constant, independent of x.

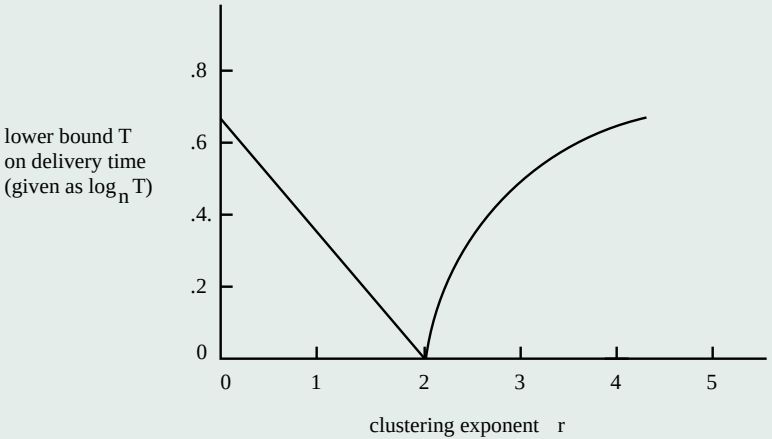
Kleinberg's model

Define as **local** a routing algorithm that knows only

- Its position in the grid
- Its neighbors
- The destination of the message

Theorem 7.1 *For this model*

- *When* $r = 2$, **there is a local algorithm** *with expected delivery time* $O(\log^2 n)$.
- *When* $r \neq 2$, *the expected delivery time of* **every local algorithm** *is* $\Omega(n^\epsilon)$, *for some* ϵ *that depends on* r.

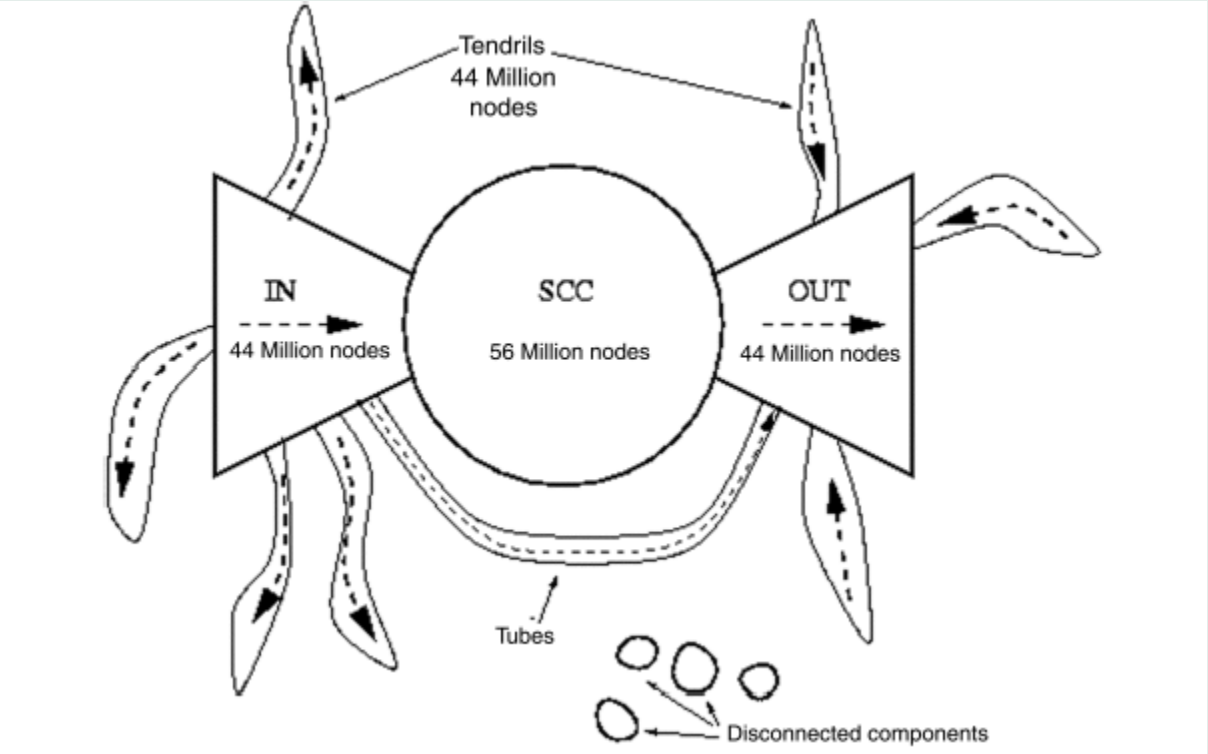


The outline of the web graph

The web graph: Nodes=Pages, Edges=Hyperlinks

- It is huge, ever-expanding graph.
- We don’t know it.
- We don’t know what percentage we know (but we have some estimate)

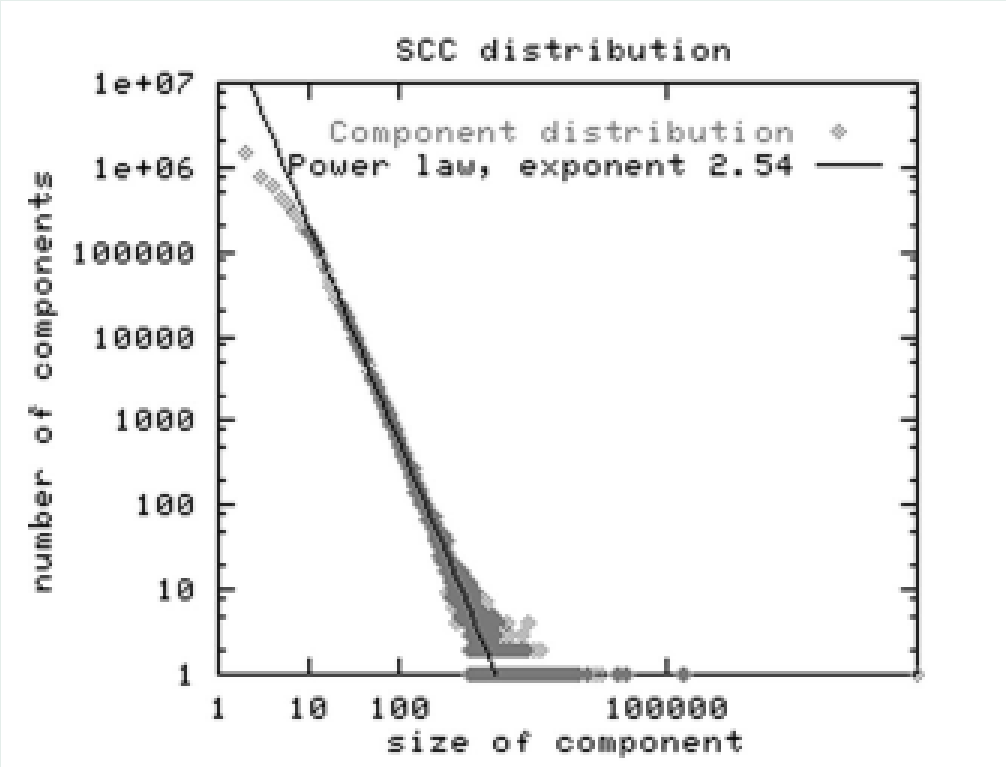
This is a picture from an article in Computer Networks by Broder, Kumar, et al



It gives a broad outline of some part of the web graph.

More findings about the web graph

- Average directed distance between two nodes: 16
- Average undirected distance between two nodes: 6
- Distribution of strongly connected components

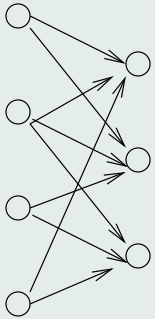


The big picture vs the local structure

- We can distinguish between global properties of the web graph and its local structure. Although it has been created in an anarchic decentralized way, it shows a great amount of self-organization.
- Its local properties are useful for searching it, for finding communities and topics
- The local structure provides important cues to search engines.
- There seems to be possible to exploit the local structure to improve algorithms. How?

Hubs and authorities

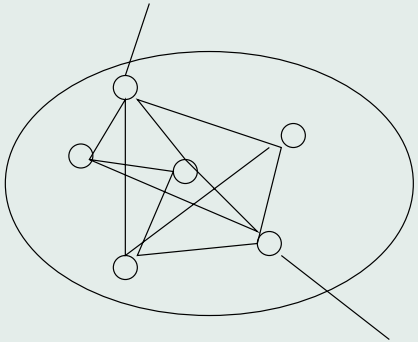
- A **hub** is a page that points to many authorities



- An **authority** is a page that is pointed to by many hubs
- This definition makes sense only if there are large dense bipartite subgraphs of the web; this has been verified empirically. Such subgraphs can be exploited by algorithms that search for particular topic.
- One great advantage of such cues, which content-based cues lack, is that they cannot be exploited by individuals to mislead the search engines.

Communities on the web

The local structure of the web often reveals the social structure that created it. Communities on the web are subgraphs that are highly interconnected but have few connections to the rest of the graph.



Usually community pages have related topics.

The big picture

- One of the striking properties of the web graph is that the degrees of its nodes follow a general law.
- In particular, the degree distribution does not match the degree distribution of random graphs
- The degrees of a random graph are highly concentrated around a particular value but the degrees of the web graph are more diverse and follow a **power law**.

The classical theory of random graphs

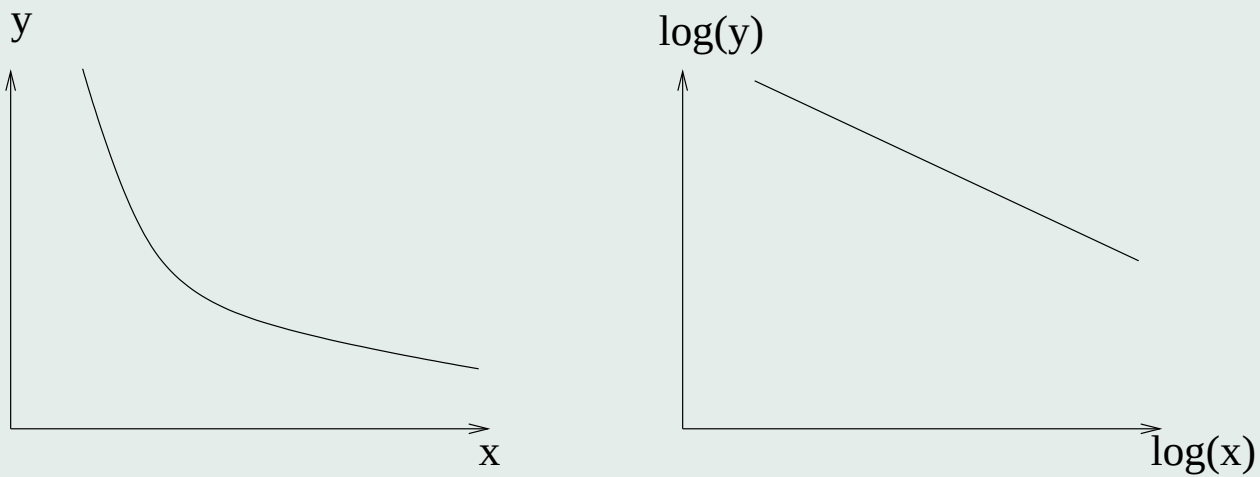
- Random graphs studied originally by Erdös and Rényi has evolved into a very influential area in the last 40 years.
- What is a random graph? Take n nodes; each possible edge between them is selected with a fixed probability p.
- Here are some facts about random graphs where $c = np$:
- **Degree distribution:** The expected degree of a node is approximately c. The probability that a node has degree away from this value drops exponentially fast.
- **Giant component:** If $c < 1$ then with high probability all components have size at most $O(\log n)$. If $c > 1$ then with high probability the largest component has size $\Theta(n)$ and all other components have size at most $O(\log n)$.
- **Diameter:** If $c = \omega(\log n)$ then the graph is connected and has diameter at most $O(\log n / \log \log n)$.

Power laws

What is a power law?
Two quantities y and x are related by a power law if

$$y \approx x^c$$

for some constant c.



Brief history of power laws

Pareto studied the distribution of income. He asked
"How many people have income greater than $100K? Than $20K?"

He observed that the probability that the income I is greater than a value v is

$$\text{Prob}[I > v] \approx v^{-k}$$

for some constant k.

Brief history of power laws (cont.)

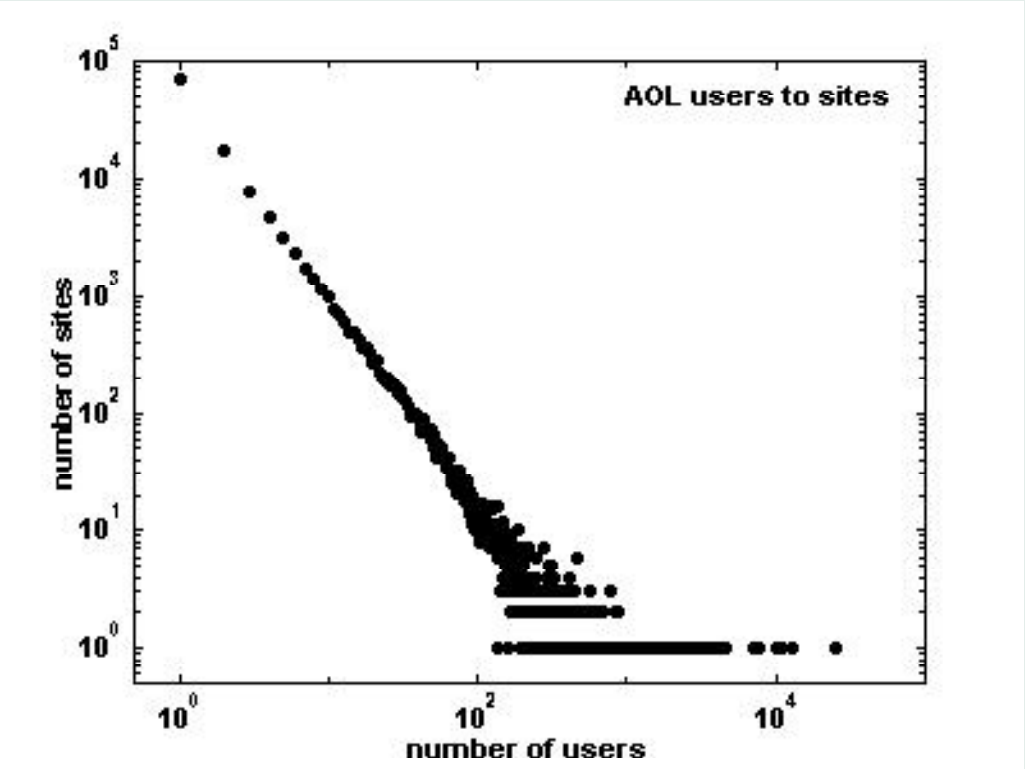
Zipf studied the distribution of frequencies of words. He asked
"What is the frequency of the most used word? Of the 100th most used?"

He observed that the frequency f of the r-th most common word is

$$f \approx r^{-1}$$

Power laws and the Web

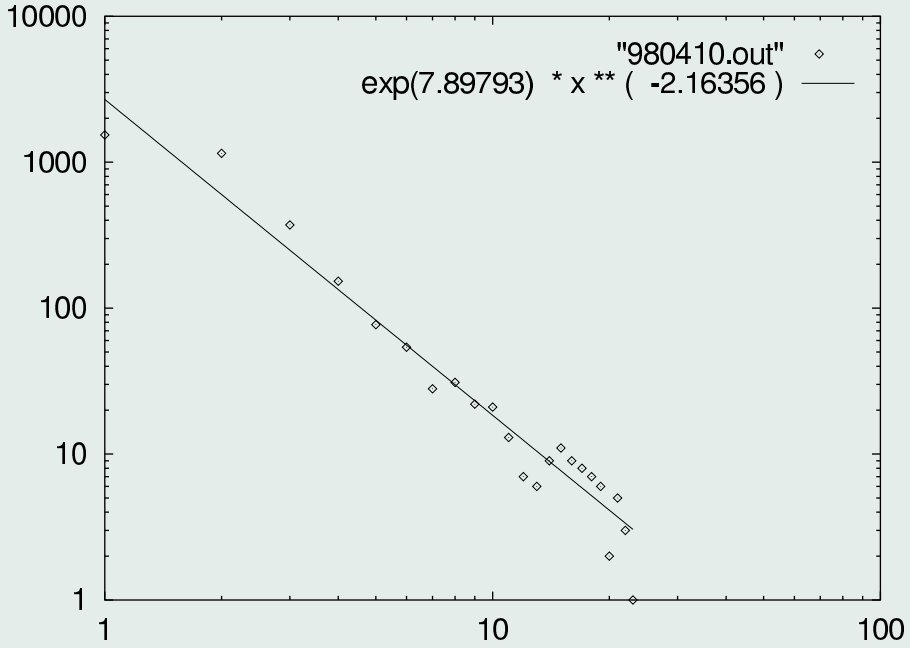
It has been observed that power laws appear in many aspects of the Internet and the Web. For example, the number of users per site obeys a power law.



Faloutsi's picture [1999]

Mihalis, Petros, and Christos Faloutsos studied the degree distribution of the Internet graph. Here is the log-log plot of the frequency vs the outdegree.

$$y(x) = \text{frequency of nodes with degree x}$$



Why such an abundance of power laws?

Power laws have been observed almost in every process involving human activity — they have been termed “the signature of human activity”.
But why?
Many times the explanation is easy: If you get enough logarithms most functions become linear!

Work in power laws

- Experimental results that measure various properties of the Web, the Internet, and other networks (railway, power lines, etc)
- Graphs with prescribed degree-sequence
- Web models where power laws arise
  - Albert-Barabasi
  - Kumar-Raghavan-Rajagopalan-Sivakumar-Tomkins-Upfal
  - Cooper-Frieze
- Internet models where power laws arise
  - Carson-Doyle's HOT model
  - Fabrikant-K-Papadimitriou's model

The Barabasi-Albert model [1999]

Albert and Barabasi proposed the following simple model for random graphs:

- Start with one node (or a small fixed graph)
- Add one-by-one nodes
- Each new node is connected to m nodes selected randomly with **probability proportional to their degree**.

The idea behind the model is simple: New pages tend to have links to "popular" pages.

The model is called the Barabasi-Albert model, or the preferential attachment model, or the rich-get-richer model, or the LCD model.

Results about the preferential attachment model

- Albert and Barabasi contacted experiments with their model and they found that the degrees obey a power law. They also, together with Jeong, gave a heuristic argument.
- Bollobás, Riordan, Spencer, and Tusnády [2001] proved the following theorem:

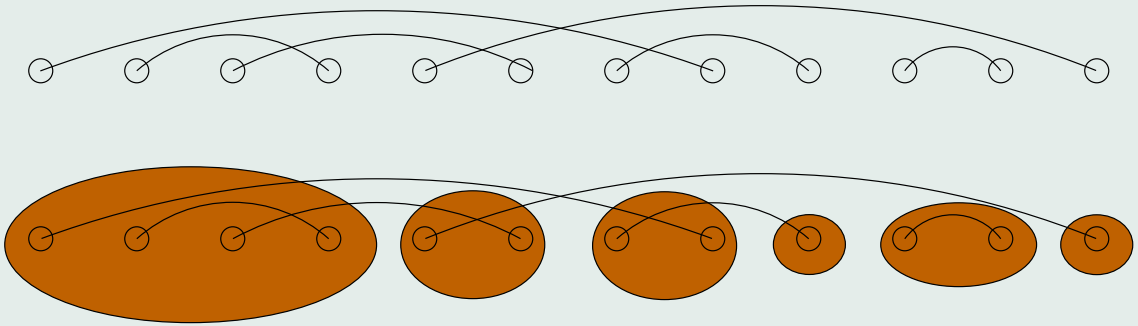
  Theorem 24.1 *The fraction of nodes with degree d is proportional to d^{-3}. More precisely, the expected number of nodes with indegree d is*

$$n \times \frac{2m(m+1)}{(d+m)(d+m+1)(d+m+2)}$$

The Bollobás-Riordan static model

Growth models vs. Static probability distributions:
Bollobás and Riordan gave a simple static description of the random graphs produced by preferential attachment:

- Consider $2n$ nodes with labels $1, 2, \ldots, 2n$ and a random pairing (matching) of them. Starting from the left identify all endpoints up to and including the first right endpoint. This is node 1. Then identify all further endpoints up to the next right endpoint. This is node 2. And so on.



Preferential attachment graphs: diameter and clustering

Theorem 26.1 *The expected diameter of a graph with n nodes is given by*

- *If $m = 1$ then the diameter is $\Theta(\log n)$*
- *If $m > 1$ then the diameter is $\Theta(\log n / \log\log n)$*

Theorem 26.2 *The expected value of the clustering coefficient of a graph with n nodes is*

$$\frac{m-1}{8}\,\frac{\log^2 n}{n}$$

The copying model

The preferential attachment model cannot explain the observation that the Web has many dense bipartite subgraphs. Another model, the copying model, by Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, and Upfal [2000] is more accurate.

To produce a random graph we start with one node and we add one-by-one the nodes. Each new node has d edges to previous nodes selected as follows: First select a (uniformly) random existing node v. For each one of the d nodes we select either a neighbor of v (with probability $1 - a$) or a random existing node (with probability a).

The intuition behind the model is the following. A new page has a topic and it is likely that most of its links will be similar to the links of some other page on the same topic. The parameter a controls what fraction of links will be new independent links.

Power laws for the copying model

Theorem 28.1 *The degree distribution of graphs produced by the copying model satisfies a power law: The expected fraction of nodes with degree d is*

$$\Theta\left(d^{-(2-a)/(1-a)}\right)$$

The graphs are rich in bipartite cliques:

Theorem 28.2 *The expected number of bipartite cliques of size $i \times d$ is*

$$ne^{-i}$$

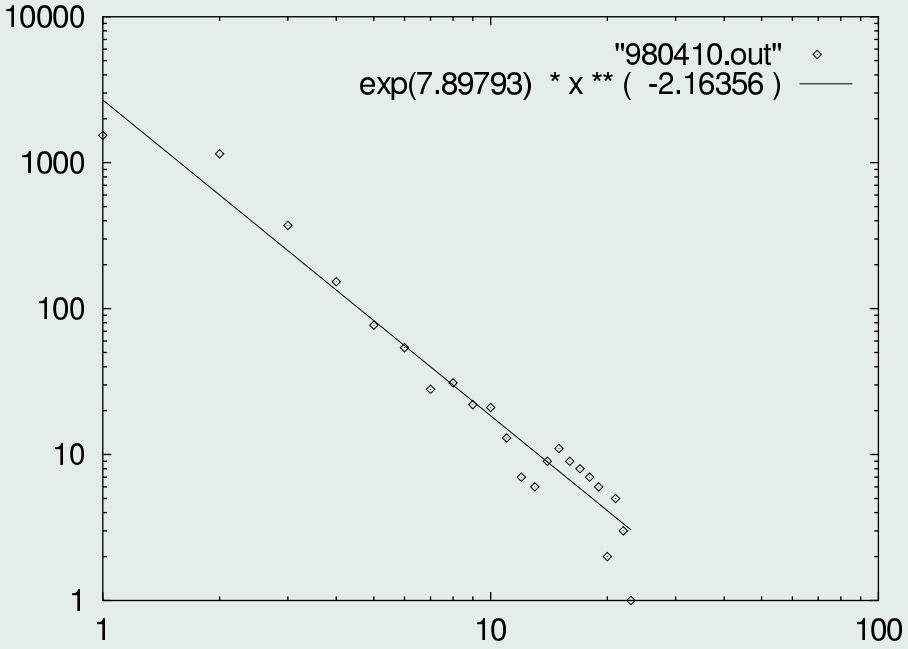
The Cooper-Frieze model

Cooper and Frieze proposed and analyzed a general model that has many parameters that can be tuned. It is a generalization of the preferential attachment model and the copying model.
The difference with the other two models is that at each step we either add a new node or add edges to an existing node. The parameters of the model make it less attractive, but it is important that power laws appear in a wide range of models.

Power laws for the Internet

The Internet graph: Nodes=Computers and Routers, Edges=links The degree distribution of the Internet seems to satisfy a power law. Here is the log-log plot of the frequency vs the outdegree from the Faloutsi paper.

$y(x) =$ frequency of nodes with degree x



Internet is not Web

The graphs of the Internet and the Web share some characteristics but differ completely on others.

Apparently the models for the Web (a virtual network) are not good for the Internet (a physical network). The most important difference is geography: A web page can link to all pages with the same cost. On the other hand, the cost of connecting two computers depends on their distance.

The HOT model of Carlson and Doyle

Carlson and Doyle, proposed HOT, a different model that predicts (and produces) power laws: Power laws are sometimes the result of optimal but reliable designs. For example, the distribution of forest fires is attributed to the firebreaks.

The Fabrikant-K-Papadimitriou model

We proposed a simple model of Internet growth, and proved that it results in power-law-distributed degrees.
A tree is built as nodes arrive uniformly at random in the unit square. When the i-th node arrives, it attaches itself on j, one of the previous nodes. But which one?

The Fabrikant-K-Papadimitriou model (cont.)

It wants to optimize two conflicting objectives:

- the distance d_{ij} to the other node (the last-mile cost)
- the centrality h_j of the other node (the operation cost)

h_j measures the centrality of node j and it can be

1. the average number of hops from other nodes
2. the maximum number of hops from another node
3. the number of hops from a fixed center of the tree

The Fabrikant-K-Papadimitriou model (cont.)

In our model, node i attaches itself to the node j that minimizes the weighted sum of the two objectives:

$$\min_{j<i} \quad \alpha \cdot d_{ij} + h_j,$$

where α is a parameter that may depend on the final number n of points.
The model attempts to capture in a simple way the *trade-offs* that are inherent in networking, but also in all complex human activity.

Results

The behavior of the model depends crucially on the value of α.

(1) If $\alpha < 1/\sqrt{2}$, then the tree is a star.

(2) If $\alpha = \Omega(\sqrt{n})$, then the degree distribution is exponential.

The expected number of nodes that have degree at least D is at most $n^2 \exp(-cD)$ for some constant c:

$$E\left[\,|\{i : \text{ degree of } i \ \geq D\}|\,\right] < n^2 \exp(-cD).$$

Main result

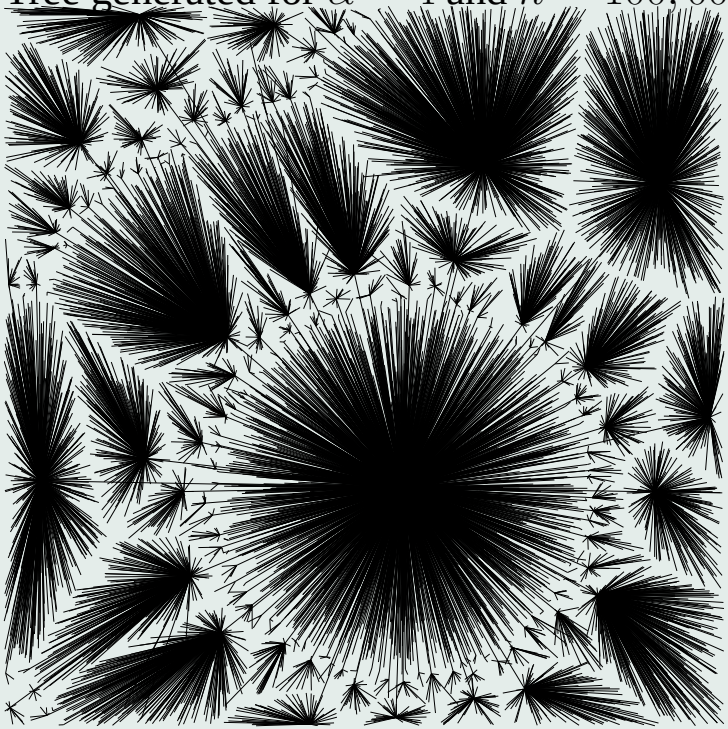
(3) If $\alpha \geq 4$ and $\alpha = o(\sqrt{n})$, then the degree distribution of T is a power law. Specifically, the expected number of nodes with degree at least D is greater than $c \cdot (D/n)^{-\beta}$ for some constants c and β:

$$E\left[\,|\{i : \text{ degree of } i \geq D\}|\,\right] > c(D/n)^{-\beta}$$

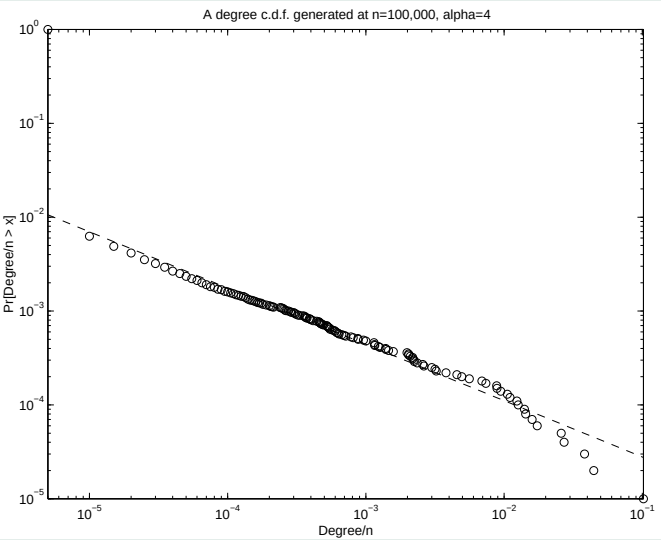
For $\alpha = o(\sqrt[3]{n})$ the constants are: $\beta \geq 1/6$ and $c = O(\alpha^{-1/2})$.

Experiments

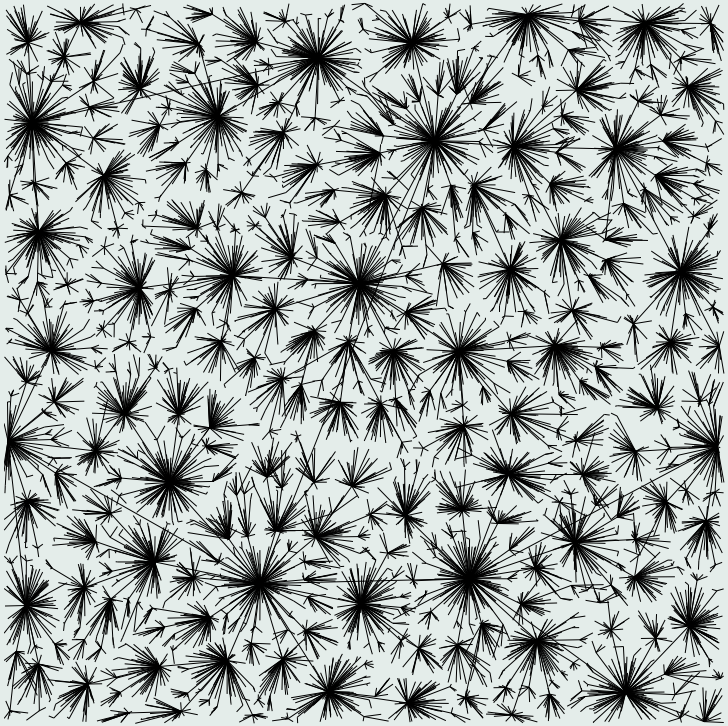
Tree generated for $\alpha = 4$ and $n = 100,000$.



c.d.f. for $\alpha = 4$ and $n = 100,000$.



Experiments

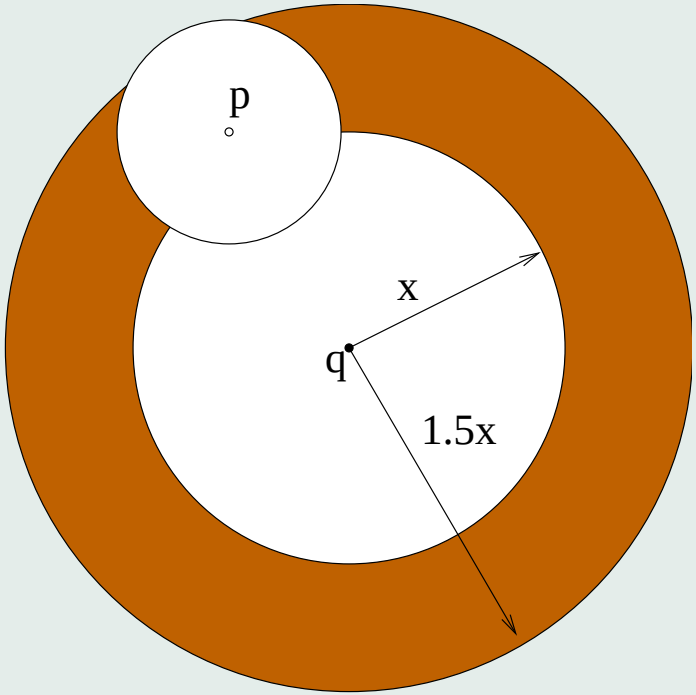


Tree generated for $\alpha = 20$ and $n = 100,000$.

Explanation — Proof

(2) If $\alpha = \Omega(\sqrt{n})$, then the degree distribution is exponential.

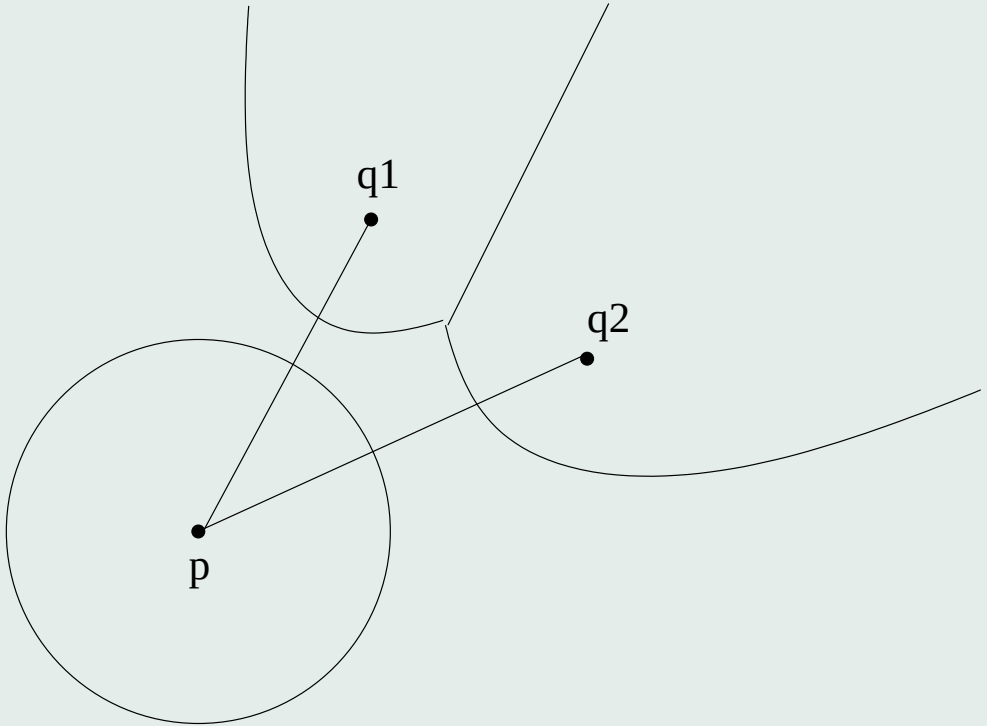
Why?



Explanation — Proof

(3) If $\alpha \geq 4$ and $\alpha = o(\sqrt{n})$, then the degree distribution of T is a power law. distribution is exponential.

Why?



So what? Why do we care about power laws?

Possible answers:

- Because we think that we have discovered some deep relation. But many times there is really nothing there. We see what we want to see.
- Because they are intriguing. They ask for an explanation. Why, for example, do the files on this particular computer follow such a regular pattern?
- Because we may use them to design better algorithms.
- Because we need the right model for simulations (and sometimes for analytical results).
- Because we can get analytical results for other problems: For example, there is some indication that a “disease” spreads easily over scale-free graphs while it needs some critical spreading rate for other graphs.