Web-graphs. Lecture 2.

Popularity as a Network Phenomenon

- Popularity is a phenomenon characterized by extreme imbalances: while almost everyone goes through life known only to people in their immediate social circles, a few people achieve wider visibility, and a very, very few attain global name recognition.
- We focus on the Web as a concrete domain in which it is possible to measure popularity very accurately.
- We will start by using the number of in-links to a Web page as a measure of the page's popularity.

Popularity as a Network Phenomenon

- The Central Limit Theorem says that if we take any sequence of small independent random quantities, then in the limit their sum (or average) will be distributed according to the normal distribution.
- If we model the link structure of the Web, for example, by assuming that each page decides independently at random whether to link to any other given page, then the number of in-links to a given page is the sum of many independent random quantities (i.e. the presence or absence of a link from each other page), and hence we'd expect it to be **approximately (!)** normally distributed.
- It is the case of Erdős–Rényi model, where degree is distributed by **Poisson**.
- **BUT** it is not the case of real complex networks.

Power Laws

From Easley & Kleinberg, Chapter 18

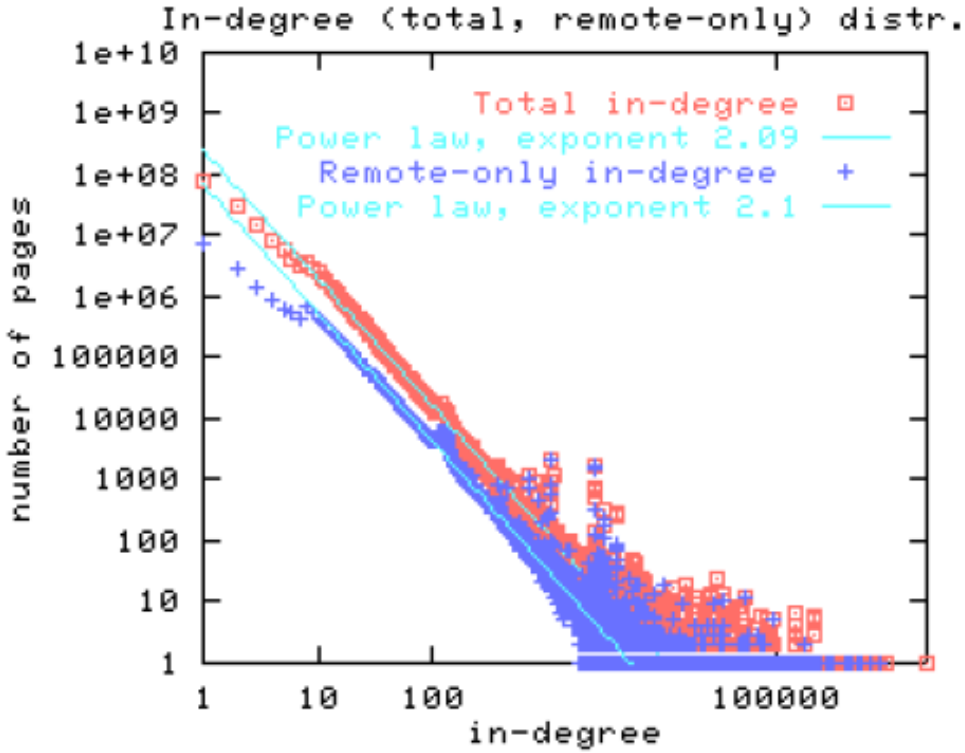


Figure 18.2: A power law distribution (such as this one for the number of Web page in-links, from Broder et al. [80]) shows up as a straight line on a log-log plot.

Rich get Richer

- Task: to propose a model that offer probable explanation for some features of existing networks.
- The models of such a kind are **generative network models**. That is, they model the mechanisms by which networks are created. The idea behind models such as these is to explore **hypothesized generative mechanisms** to see what structures they produce.
- The power law is a somewhat unusual distribution and its occurrence in empirical data is often considered a potential indicator of interesting underlying processes. A natural question to ask therefore is how might a network come to have such a distribution? This question was first directly considered in the 1970s by Price.
- In considering the possible origins of the power law in the citation network, Price was inspired by the work of economist Herbert Simon, who noted the occurrence of power laws in a variety of (non-network) economic data, such as the distribution of people's personal wealth. $\rightarrow$ **The Rich get Richer effect**
- Price gave a name to Simon's mechanism: he called it cumulative advantage, although it is more often known today by the name **preferential attachment**, which was coined in 1999 by Barabasi and Albert.

Preferential attachment. Barabasi and Albert "model"

- Fix some initial graph G_0.
- Add vertices one by one.
- With each new node we add m new edges incident to the new vertex.
- New node is connected with m **different** existing vertices.
- Preferential attachment: the probability of connection of the new edge to some existing node is proportional to its degree.

Authors argue that we obtain the **sparse network with n nodes and mn edges, having small diameter and pow-law degree distribution**.

Barabasi and Albert "model": wording inaccuracies

- Random graph appearing in the model strongly depends on the properties of G_0
- The formulation of multivariate distribution of the new edges ends is not clear.
- Bollobash and Riordan: by appropriate choice of G_0 and edges ends distribution we can obtain **any fixed** number of triangles in the network.
- **It is not a specific model, it is a class of models.**

Bollobash - Riordan model

Our aim is to construct the random graph G_m^n having n vertices and mn edges.

Let $deg_G(v)$ be the node's v degree in graph G.

$\mathbf{m = 1}$

- Fix G_1^1 – one node and one loop
- G_1^{n-1} is a graph on $n-1$ vertices $\{v_1, \ldots, v_{n-1}\}$ having $n-1$ edge
- G_1^n is obtained from G_1^{n-1} by addition of one node and edge incident to it and to some node $i \in [1, \ldots, n]$ with the following probability:

$$P(i = s) = \begin{cases} \frac{deg_{G_1^{n-1}} v_s}{2n-1}, & \text{if } 1 \leq s \leq n-1 \\ \frac{1}{2n-1}, & \text{otherwise.} \end{cases}$$

- $\sum_{s=1}^{n} P(i = s) = 1$

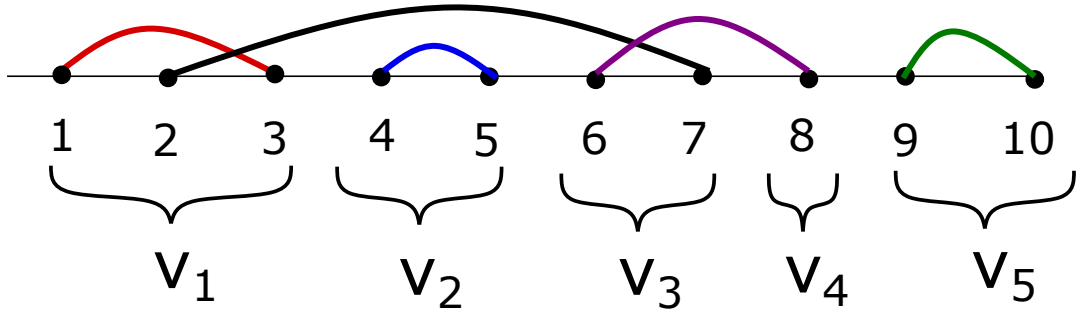
Bollobash - Riordan model

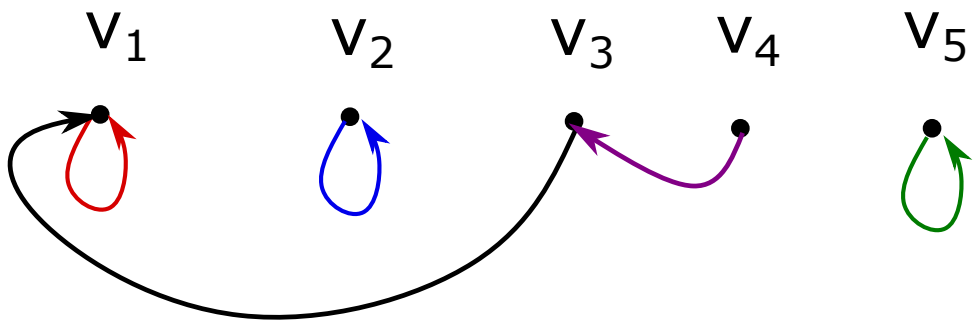
$\mathbf{m > 1}$

- Graph G_m^n is obtained from G_1^{mn} by gluing together the following nodes:
  - $\{v_1, \ldots, v_m\} \longrightarrow v'_1$
  - $\{v_{m+1}, \ldots, v_{2m}\} \longrightarrow v'_2$
  - $\ldots$
  - $\{v_{(n-1)m+1}, \ldots, v_{nm}\} \longrightarrow v'_n$
- Let $v'(v_i)$ is the name of vertex in G_m^n corresponding to the group in which v_i falls.
- Edge $(v_i, v_j) \longrightarrow (v'(v_i), v'(v_j))$

Bollobash - Riordan model. Static definition

- Line chord diagram: put $2n$ points and connect pairs (randomly) without intersections





- How many different chord diagrams exists? The number of different chord diagrams = the number of unordered pairs from $2n$ elements = $\frac{(2n)!}{n!2^n}$
- The probability to obtain some specific graph in the Bollobash-Riordan model with $m = 1$ equals to the probability to choose the line chord diagram generating this graph by taking it from the set of all line chord diagrams on $2n$ point uniformly.

Bollobash - Riordan model. Properties

Diameter

Let $m \geq 2$ and G_m^n is the random graph in the Bollobash-Riordan model.

$$diam(G_m^n) \sim \frac{\ln n}{\ln \ln n}, \quad n \to \infty$$

Let $n = 2 \cdot 10^7 \quad \longrightarrow \quad \frac{\ln n}{\ln \ln n} = 5.96$

Bollobash - Riordan model. Properties

Stability to random attacks

Let

- $m \geq 2$,
- G_m^n is the random graph in the Bollobash-Riordan model,
- $p \in [0, 1)$ some fixed constant,
- $G_{m,p}^n$ is the random graph obtained from G_m^n by removal of each node with probability p.

There exists $c = c(m, p)$, such as with probability tending to 1 while $n \to \infty$ $G_{m,p}^n$ contains only one weakly connected component of size $(c + o(1))n$

Bollobash - Riordan model. Properties

Vulnerability to attacks on hubs

Let

- $m \geq 2$,
- G_m^n is the random graph in the Bollobash-Riordan model,
- $c \in (0, 1)$ some fixed constant,
- $G_{m,c}^n$ is the random graph obtained from G_m^n by removal of the largest $[cn]$ hubs,
- $c_m^* = \frac{m-1}{m+1}$.

While $c < c_m^*$ with probability tending to 1 while $n \to \infty$ $G_{m,p}^n$ contains giant weakly connected component and does not otherwise (when $c > c_m^*$).

Bollobash - Riordan model. Properties

Degree distribution

Theorem (Bollobas, Riordan, Spencer, Tusnady)
Let $m \geq 1$, G_m^n is the random graph in the Bollobash-Riordan model. Then for each $d \leq n^{1/15}$ with probability tending to 1 and $n\ to\infty$

$$\frac{|\{v \in G_m^n : degv = d\}|}{n} \sim \frac{2m(m+1)}{d(d+1)(d+2)}$$

Restriction on d ($d \leq n^{1/15}$) can be omitted (the result of Grechnikov).
Therefore, in the Bollobash-Riordan model we have power law with $\gamma = 3$.

In real web-graphs $\gamma \in (2, 3) \longrightarrow$ some modifications should be introduced.