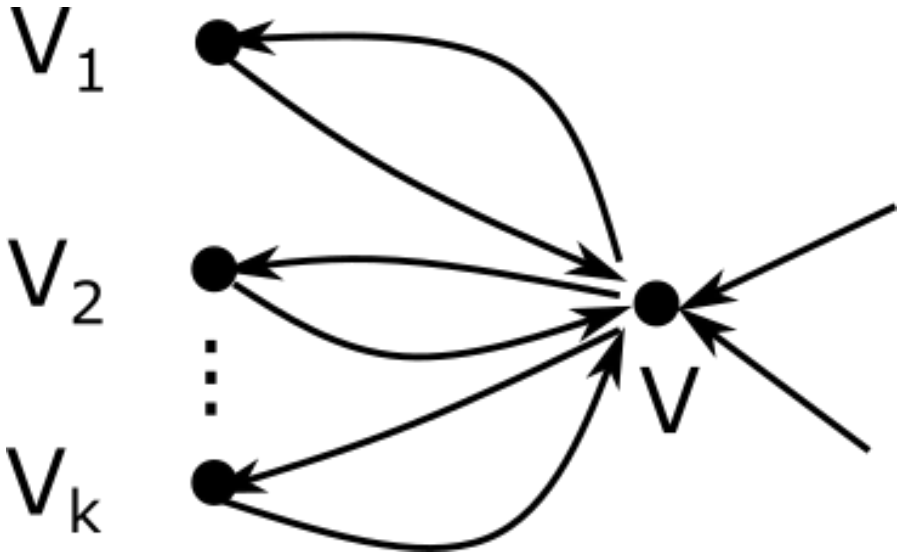
Web-graphs. Lecture 7.

PageRank Cheat

- Link farm example



- Node v wants to increase its PageRank $\longrightarrow$ it creates a series of pages co-cited pages
- Why creation of such a *link farm* is profitable for v?

PageRank Cheat

- N – number of nodes in the graph
- Simplifying assumption: there are no nodes with zero out-degree
- Node v PageRank in calculated as follows

$$x(v) = \alpha \sum_{u \to v} \frac{x(u)}{d_u^{out}} + \frac{1-\alpha}{N} \equiv \alpha p + \sum_{i=1}^{k} x(v_i) + \frac{1-\alpha}{N},$$

  where p is the contribution of all nodes except $\{v_1, \ldots, v_k\}$
- Node $v_i,\ i = 1, \ldots, k$ PageRank is calculated as follows

$$x(v_i) = \alpha \frac{x(v)}{k} + \frac{1-\alpha}{N},\ i = 1, \ldots, k$$

- Combining expressions for $x(v_i)$ and $x(v)$ we obtain

$$\begin{aligned} x(v) &= \alpha p + \alpha \sum_{i=1}^{k} \left(\alpha \frac{x(v)}{k} + \frac{1-\alpha}{N} \right) + \frac{1-\alpha}{N} \\ &= \alpha p + \alpha^2 x(v) + \frac{1-\alpha}{N}(\alpha k + 1) \end{aligned}$$

PageRank Cheat

- Resolving the last equation for $x(v)$ we obtain

$$x(v) = \frac{1}{1-\alpha^2}\left(\alpha p + (\alpha k + 1)\frac{1-\alpha}{N}\right)$$

- $k \uparrow \longrightarrow x(v) \uparrow$
- From this example it is clear that one needs to detect and remove link farms in the process of PageRank calculation to increase the quality of search engine

PageRank distribution

Examples from *Pandurangan, G., Raghavan, P., & Upfal, E. (2002, August). Using pagerank to characterize web structure. In International computing and combinatorics conference (pp. 330-339). Springer, Berlin, Heidelberg.*

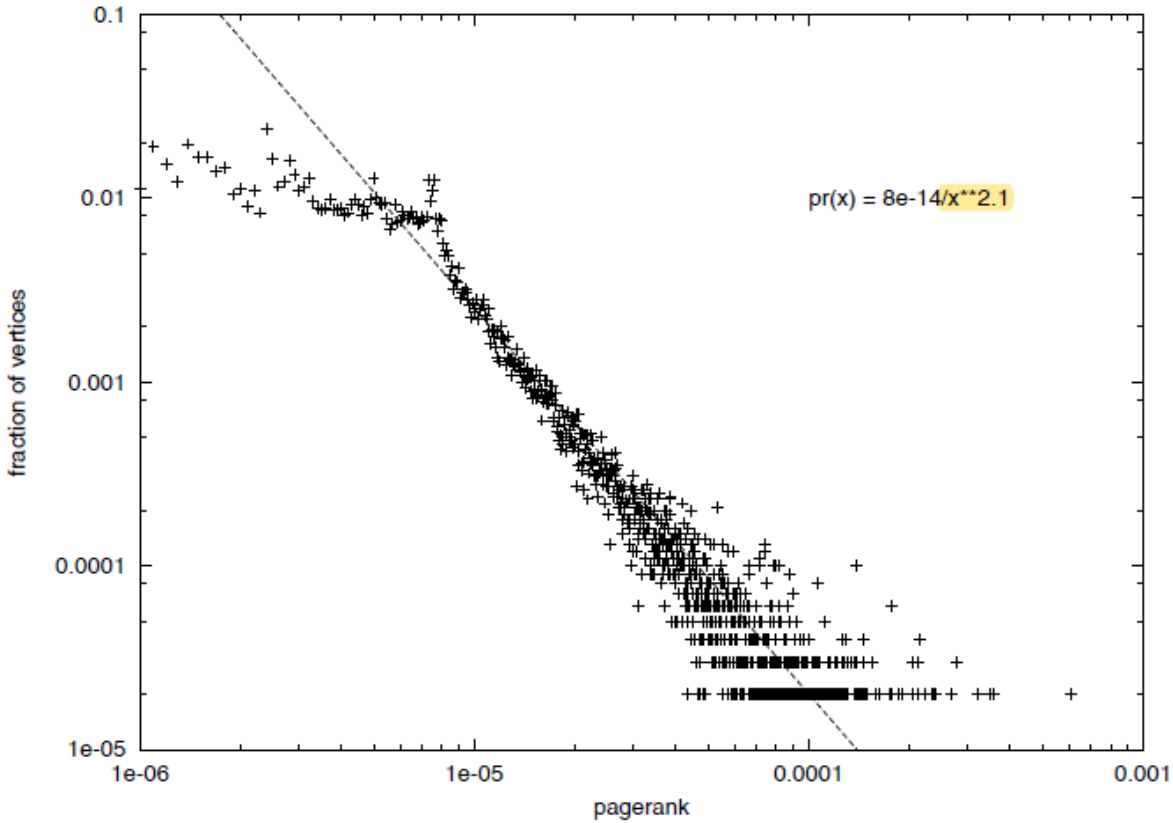


Figure 3. Log-log plot of the PageRank distribution of the Brown domain (`*.brown.edu`). A vast majority of the pages (except those with very low PageRank) follow a power law with exponent close to 2.1. The plot almost flattens out for pages with very low PageRank.

PageRank distribution

Examples from *Pandurangan, G., Raghavan, P., & Upfal, E. (2002, August). Using pagerank to characterize web structure. In International computing and combinatorics conference (pp. 330-339). Springer, Berlin, Heidelberg.*

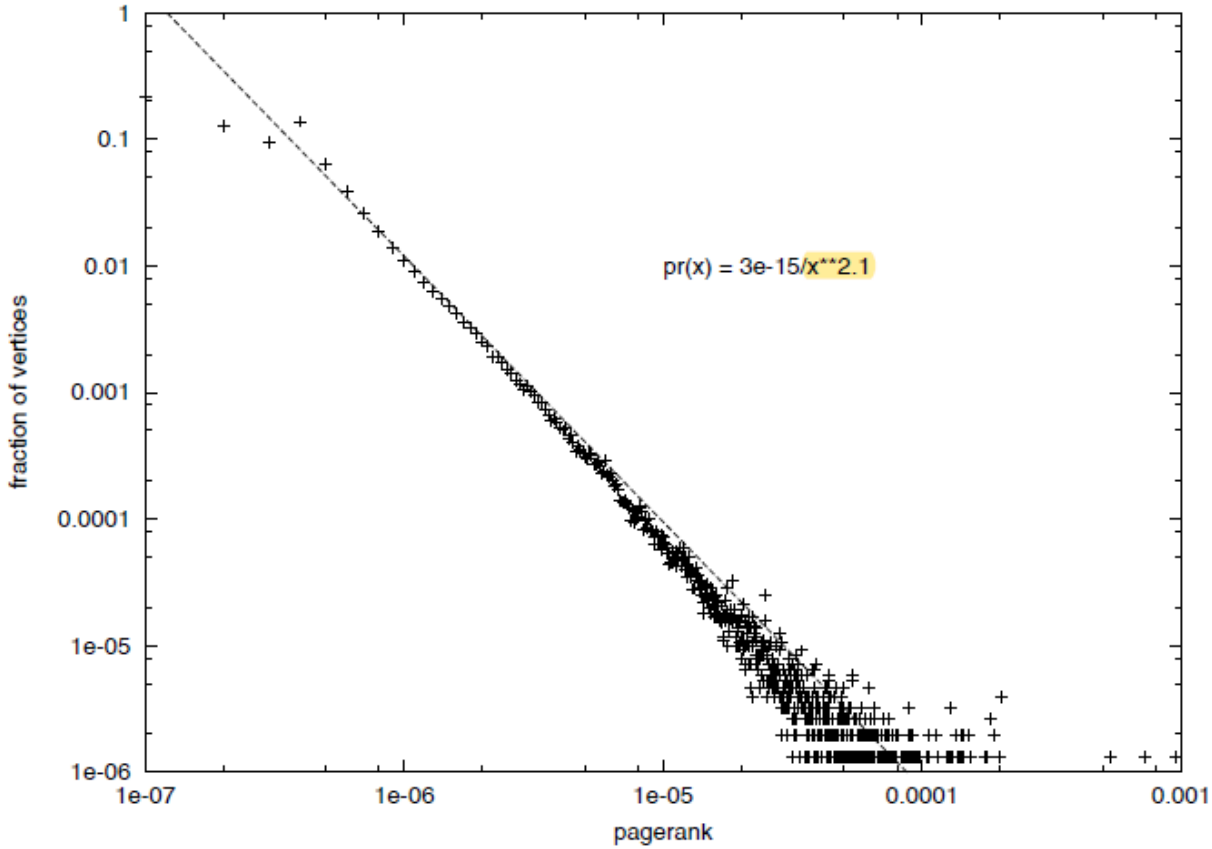


Figure 4. Log-log plot of the PageRank distribution of the WT10g corpus. The slope is close to 2.1. Note that the plot looks much sharper than the corresponding plot for the Brown web. Also, the tapering at the top is much less pronounced.

PageRank distribution

- In real complex networks PageRank distribution is scale-free (= described by the power-law)
- It is of interest to test if the property holds in the networks growing models like Backley - Ostgus and Bollodas - Riordan ones.
- However, it is difficult to obtain **analytical** estimates for these models because of loops.
- While calculating PageRank Google removes the loops
- Analytical results can be obtained in the framework of another model of growing network with preferential attachment
- **Avrachenkov model**: *Avrachenkov, K., & Lebedev, D. (2006). PageRank of scale-free growing networks. Internet Mathematics, 3(2), 207-231.*

Avrachenkov model

- The network grows at the speed of one node per time step
- m – the number of outgoing links from each node
- At each time step a new node creates m links to the existing nodes.
- Let us denote the growing network at arbitrary time step n by G_n^m
- We denote by $d_v(n)$ the *in*-degree of node v at time step n

Avrachenkov model

- At step 0 the initial node 0 is created and it has no links. The initial node has weight m by definition.
- Then, at the next time step 1 a new node has no other choice but to connect its m links to the initial node. Node 1 receives the weight m and the weight of node 0 becomes $2m$.
- A new node that appears after time step 1 connects each of its m edges independently with probability proportional to the existing nodes' weights equal to in-degrees plus m. Namely, the probability that node n connects to node v, $v < n$, is given by

$$\mathbb{P}\left[n \to v | G_{n-1}^{m}\right] = \frac{d_v(n-1)+m}{\sum_{k=0}^{n-1}(d_k(n-1)+m)} = \frac{d_v(n-1)+m}{2m(n-1)+m}$$

For instance, node 2 connects with probability $2/3$ to the initial node 0 and with probability $1/3$ to node 1

PageRank distribution in the Avrachenkov model

- $x_i(n) = \alpha \sum_{j \to i} \frac{x_j(n)}{d_j^{out}} + \frac{1-c}{n+1}$
- **Theorem (Avrachenkov, Lebedev)**
  If $i > 0$

$$\mathbb{E}x_i(n) \approx \frac{1-\alpha}{1+n}\left(\frac{1}{1+\alpha} + \frac{\alpha}{1+\alpha}\left(i+\frac{1}{2}\right)^{-\frac{1+\alpha}{2}}\left(n+\frac{1}{2}\right)^{\frac{1+\alpha}{2}}\right)$$

- PageRank values $x_i(n)$ are not discrete, they are absolutely continuous. Let $p_n(x)$ be the PageRank density function

$$p_n(x) \approx \Theta\left(\frac{1}{x^{\kappa}}\right),$$

where $\kappa = \frac{3+\alpha}{1+\alpha}$. If $\alpha = 0.85$ $\kappa \approx 2.1$

Weighted "Yandex" PageRank

- Let $G(V, E)$ be the web-graph
- To each node $i \in V$ we match vector of features v_i having dimension l.
- To each edge $i \to j$ we match vector of features e_{ij} having dimension m.
- Two key (new) parameters of the model:
  - Vector ω having dimension l (weights for nodes features)

$$f(\omega, v_i) = (\omega, v_i) -$$

    scalar product
  - Vector ϕ having dimension m (weights for edges features)

$$g(\phi, e_{ij}) = (\phi, e_{ij})$$

- The third parameter is α as in classic algorithm

Weighted "Yandex" PageRank

- Weighted "Yandex" PageRank is the vector $\mathbf{x}$ solving the system of linear equations $\mathbf{x} = \mathbf{A}\mathbf{x}$, where matrix $\mathbf{A}$ element A_{ij} is as follows

$$A_{ij} = (1-\alpha)\frac{f(\omega, v_j)}{\sum_{k=1}^{N} f(\omega, v_k)} + \begin{cases} \alpha \frac{g(\phi, e_{ij})}{\sum_{h:j\to h} g(\phi, e_{jh})} \mathbb{I}(j \to i), & \text{if } d_j^{out} > 0, \\ \alpha \frac{f(\omega, v_j)}{\sum_{k=1}^{N} f(\omega, v_k)}, & \text{otherwise} \end{cases}$$

- To fit the model, first, the relatively small subset of web-pages are manually ranked and, second, parameter vectors ω and ϕ and parameter α are chosen, such as ranking provided by weighted "Yandex" PageRank is close to the manual one.

Bollobas - Borgs - Riordan - Chayes model

- In the preferential attachment models described previously (Bollobas - Riordan, Backley - Ostgus, Avrachenkov models) only in-degrees were distributed by the power-law
- In these models directions are more or less artificial: there exist only links from newer nodes to the older ones. There are no directed cycles and, as follows, no SCC.
- In real graphs both in- and out-degrees distributions are described by the power law
- Bollobas - Borgs - Riordan - Chaies model is focused on the realistic directions description.

Bollobas - Borgs - Riordan - Chayes model

- We will build $G(t)$ – growing directed scale-free graph
- Let us introduce the following notations:
  - $\delta_{out} \geq 0$ – the initial traction to link creation
  - $\delta_{in} \geq 0$ – initial attractiveness of the node
  - $\alpha,\ \beta,\ \gamma > 0$ – parameters such as:

$$\alpha + \beta + \gamma = 1$$

- $G(1)$ – graph having one node v_1 and one loop

Bollobas - Borgs - Riordan - Chayes model

- $G(t)$ can be obtained from the graph $G(t-1)$ on $n(t-1)$ nodes (it is now the random value!) using one of the following three alternatives:
  - with probability α we add the new node and an *outgoing* edge going from this node to one of the existing $n(t-1)$ nodes (let it be node v) with the following probability
$$\frac{d_v^{in} + \delta_{in}}{t - 1 + \delta_{in} n(t-1)} \tag{1}$$
  - with probability γ we add the new node and one *incoming* edge from one of the existing $n(t-1)$ nodes (let it be node v) with the following probability
$$\frac{d_v^{out} + \delta_{out}}{t - 1 + \delta_{out} n(t-1)} \tag{2}$$
  - with probability β no new nodes appear but one new edge from the existing node v to the existing node ω appears. Node v is chosen with probability (2) and node ω – with probability (1).

Bollobas - Borgs - Riordan - Chayes model

Theorem (Bollobas, Borgs, Riordan, Chayes)

Let

- $i \geq 1$ be some fixed natural number,
- $x_i(t)$ be the number of $G(t)$ nodes having *in-degree* i
- $y_i(t)$ be the number of $G(t)$ nodes having *out-degree* i
- $c_1 = \frac{\alpha+\beta}{1+\delta_{in}(\alpha+\gamma)}$
- $c_2 = \frac{\beta+\gamma}{1+\delta_{out}(\alpha+\gamma)}$

Then there exist $p_i,\ q_i$ constant for each i and $\phi_i(t),\ \psi_i(t) = o(t)$ such as with probability tending to 1 for every t the following holds

$$x_i(t) = p_i t + \phi_i(t)$$
$$y_i(t) = q_i t + \psi_i(t)$$

Moreover, if $\alpha\delta_{in} + \gamma > 0,\ \gamma < 1$ there exists $c_{in} > 0$: while $i \to \infty$

$$p_i \sim c_{in} i^{-1-\frac{1}{c_1}}$$

And if $\gamma\delta_{out} + \alpha > 0,\ \alpha < 1$ there exists $c_{out} > 0$: while $i \to \infty$

$$q_i \sim c_{out} i^{-1-\frac{1}{c_2}}$$