Markov Chains

Website Traffic Prediction

Tanay Biradar

October 2, 2021

Website Traffic and PageRank

- Search engines use popularity to rank pages
- Popularity can be quantified by links to a page
  - Works in theory, can be abused in practice
- Google used a Markov Model (PageRank) to rank popularity

Website Traffic and PageRank

"PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites." [1]

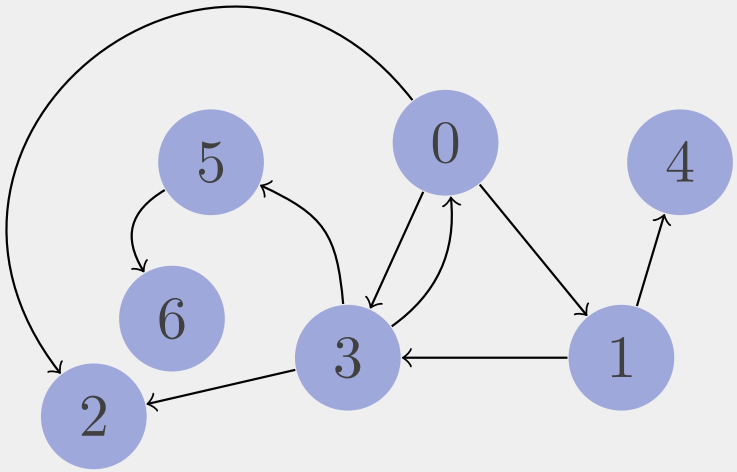
[1] https://en.wikipedia.org/wiki/PageRank

Central Question

Which page is a user most likely to land on after starting on a given page?

Model

- Represent the internet as a directed graph
  - We're looking at a small slice of the web
- Edges are links, vertices are web pages
  - Assume equal probability of traversing every link such that [2] $\Sigma w_{out} = 1$, where w is the edge weight
    - The probabilities coming out of every website must sum to 1

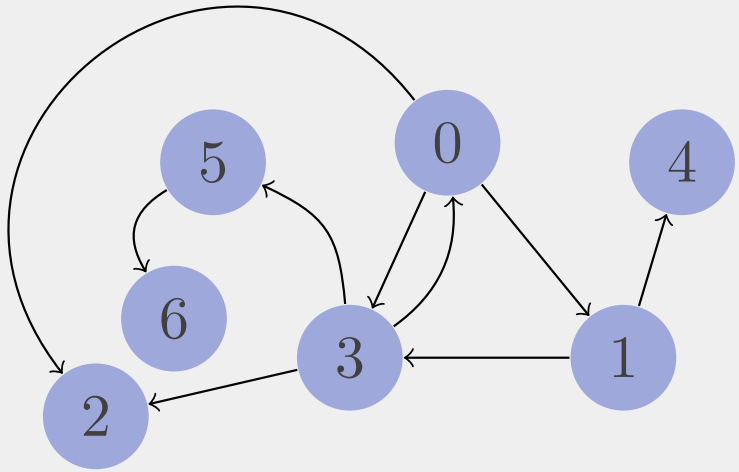


[2] https://en.wikipedia.org/wiki/PageRank

Data

The most complex part is by far the data collection

- Start at an arbitrary website, perform BFS to create an adjacency list that represents the internet
  - Keep track of visited nodes to avoid duplicate processing
- Stop after storing __ thousand links
  - I don't have the computing power of Google



Data (cont'd)

Adjacency list A stores links between pages
If $A_{ij} = 1$, there is a link from page i to page j

$$A = \begin{bmatrix} a_{00} & \cdots & a_{0n} \\ \vdots & \ddots & \vdots \\ a_{n0} & \cdots & a_{nn} \end{bmatrix}$$

Normalize the adjacency list to satisfy $\forall i \ \Sigma w_{out} = \Sigma w_i = 1$
We now have a transition matrix!

Start at a given page