EE126 Lab 3 Mini-Project

Shane Barratt: 24279112

March 15, 2015

Navigating the endless number of academic papers can be overwhelming, especially when there are tens of thousands of papers for extremely specialized fields. To solve this problem, I turned to the PageRank algorithm first theorized and developed by the founders of Google, Larry Page and Sergey Brin.

The dataset, from the Stanford Network Analysis Project, consists of links between 27,770 high-energy physics theory papers published between the years from 1992-2003. When a paper cites another, a directed edge is formed between the two. In total, there are 352,807 edges in this graph. The data was retreived from *http://snap.stanford.edu/data/cit-HepTh.html.*

In order to calculate the "PageRank" of each paper, the graph was modeled as a Markov Chain with each paper being a state. With a probability c, a state will transition to any other state including itself. With a $1-c$ probability, a state will transition uniformly randomly to any other paper it is connected to. In other words,

$$p_{ij} = \begin{cases} \frac{c}{1+cN} & \text{if } x \geq 0 \\ \frac{c+\frac{N_{ij}}{N'}}{1+cN} & \text{if } x < 0 \end{cases},$$

where N_{ij} is the count of links between paper i and paper j, N is the total number of papers and N' is the number of papers that paper i is connected to. For simplicity, assume that there is at least one link going out of every page. Ideally a Probability Transition Matrix P would be constructed where $P_{ij} = p_{ij}$ and the left eigenvector would be its stationary distribution. However, this is computationally infeasible with a 27,770 x 27,770 matrix. Equally infeasible, balance equations could be set up and iteratively updated until convergence. To solve this infeasibility, a random walk was performed and the stationary distribution $\pi_i \approx \frac{K_i}{K}$, where K_i is the number of times the random walk visits page i, and K is the total number of steps taken by the random walk. We can approximate this because at the limit these are equal.

A simulation in Java was constructed to perform this random walk and with a c value of 0.15 and 100,000,000 transitions produced the following results:

PageRank	Title	Cited By (Google Scholar)	Year Published
0.02429542	Noncompact Symmetries in String Theory	537	1992
0.02354308	An Algorithm to Generate Classical Solutions for String Effective Action	20	1992
0.00725771	Monopole Condensation, And Confinement In N=2 Supersymmetric Yang-Mills Theory	3640	1994
0.00547113	String Theory Dynamics In Various Dimensions	2561	1995
0.00497628	Exact Results on the Space of Vacua of Four Dimensional SUSY Gauge Theories	780	1994

Most of the papers have a page-rank under 0.001, implying that only a select few of the papers are well-connected in the graph. It is not surprising that the papers with the highest page rank were published the earliest in the date range.