

---

**HW 4: Search**

Assigned: 01/30/12

Due: 02/10/12 13:30

---

*We encourage you to discuss these problems with others, but you need to write up the actual solutions alone. At the top of your homework sheet, please list all the people with whom you discussed. Crediting help from other classmates will not take away any credit from you. Start early, especially Question 1, and come to office hours with your questions! The TAs would greatly appreciate it if, for soft copy submissions, you just email a single pdf file containing your solutions, code, figures, etc.*

**0 How long have you taken to finish this problem set? [Bonus points]****1 Rankmaniac Reloaded [100 points]**

In this exercise you'll try your hand at search engine optimization. Your goal is to create a web page featuring an image that ranks highly under Google and/or Bing image search for the query "rankmaniac 2012". In previous iterations of this course, your predecessors competed to have their web pages rank highly on Google and Bing search for the search query "rankmaniac 201X". We'll expect you guys to 'outrank' their best efforts!

**Background:** Building a powerful web presence is becoming increasingly important to companies, organizations, and even individuals. Many of these entities have an economic incentive to attract high volumes of (relevant) web traffic to their web sites. Search engine optimization (SEO) is the process of increasing the traffic to a web site by improving its ranking on search engines for certain relevant keyword searches. You can read the Wikipedia entry on SEO ([http://en.wikipedia.org/wiki/Search\\_engine\\_optimization](http://en.wikipedia.org/wiki/Search_engine_optimization)) for more details. Wikipedia also has an interesting article on the history of online SEO contests ([http://en.wikipedia.org/wiki/Nigritude\\_ultramarine#Nigritude\\_ultramarine](http://en.wikipedia.org/wiki/Nigritude_ultramarine#Nigritude_ultramarine)).

**Your task:** Working in groups of 3-4, your objective will be to create a web page containing an image that will rank as highly as possible among the search results on Google and/or Bing image search for the query "rankmaniac 2012". (The quotes are not included in the query.) **You will be evaluated based on the ranking of your image for this search query on February 15, 2012 at 12:30pm.** To check the rankings, we'll use Adam's laptop running Google Chrome in incognito mode, with safe search ON. Note that we are giving you this much time because it typically takes a little while for a new page to be crawled and indexed by Google/Bing, so you'll have to start right away!

**Important notes:**

- Your page must include the names of all group members. You can create your page inside or outside of the Caltech domain, as you prefer.
- You are *not* allowed to use any money for this entire exercise. Buying domain names for hosting the website or paying at Amazon Mechanical Turks for clicking on links to your webpage are not allowed.

## Grading:

- (a) **Report (50 points):** Each group must turn in one detailed report describing the steps taken to optimize your ranking, and why you think your approaches will work. The report must describe the contribution of each team member to your team's effort. Of course, your report should mention the url of the page you have created. **This report is due 1.30 pm, Feb. 10.**
- (b) **Get listed (20 points):** You receive 10 points if an image on your page is listed by Google image search when the query "rankmaniac 2012" is searched for, and another 10 points if an image on your page is listed by Bing image search for the same query.
- (c) **Beat the TAs (20 points):** The TAs will take a shot at this assignment along with you. You get 20 points if your image is ranked higher for the query "rankmaniac 2012" than the one on the page created by the TAs on either Google or Bing image search.
- (d) **Beat your predecessors (5 points):** You get 5 points if your image is ranked higher than the highest ranked image from the pages created by your Caltech predecessors for the search query "rankmaniac" on either Google or Bing image search.
- (e) **Image suggestions in web search (5 points):** The whole class gets 5 points if image suggestions show up for the search query "rankmaniac 2012" on the plain Google/Bing web search.
- (f) **Class competition (Bonus points):** Each team that is first for "rankmaniac 2012" on either Google or Bing image search gets 10 points. Each team that is second on either Google or Bing image search gets 8 points. Each team that is third on either gets 6 points.

Have fun! We hope to see some very creative approaches to raising PageRank...

## 2 Warmup with stationary distributions [10 points]

We have seen two ways of calculating the stationary distribution of a transition matrix  $P$ . One is the iterative method  $\pi = \pi_0 \lim_{n \rightarrow \infty} P^n$  for initial  $\pi_0$ , the other one is to solve the equation  $\pi = \pi P$ . For the following probability transition matrices, first use  $\pi = \pi P$  to get the stationary distribution, and then show whether or not  $P^n$  converges as  $n \rightarrow \infty$ . If it converges, show that  $\pi = \pi_0 \lim_{n \rightarrow \infty} P^n$  does not depend on  $\pi_0$ . (hint: you may want to diagonalize  $P$  to handle  $P^n$ .)

1. 
$$\begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0.1 & 0.7 & 0.2 \\ 0.6 & 0.1 & 0.3 \end{pmatrix}$$

2. 
$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \\ 0.4 & 0 & 0.6 & 0 \end{pmatrix}$$

### 3 Training to be a farmer [45 points]

To help prepare you to think about problem 1, we'll do a little analysis of the design of link farms, and how their structure affects the PageRank calculations.

Consider the web graph. It contains  $n$  pages, labeled 1 through  $n$ . Of course,  $n$  is very large. Let  $r_i$  denote the PageRank of page  $i$ , and  $r = (r_1, r_2, \dots, r_n)$  denote the vector of PageRanks of all pages.

- (a) You now create a new web page  $X$  (thus adding a node to the web graph).  $X$  has neither in-links, nor out-links. Let  $\tilde{r} = (\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_n)$  denote the vector of new PageRanks of the  $n$  old web pages, and  $x$  denote the new PageRank of page  $X$ .

Write  $\tilde{r}$  and  $x$  in terms of  $r$ . Comment on how the PageRanks of the older pages changed due to the addition of the new page (remember  $n$  is a very large number). Hint: Use the stationary equations to calculate PageRank, not the iterative approach.

- (b) Unsatisfied with the PageRank of your page  $X$ , you create another page  $Y$  (with no in-links) that links to  $X$ . What are the PageRanks of all the  $n + 2$  pages now? Does the PageRank of  $X$  improve?
- (c) Still unsatisfied, you create a third page  $Z$ . How should you set up the links on your three pages so as to maximize the PageRank of  $X$ ?
- (d) You have one last idea, you add links from your page  $X$  to older, popular pages (e.g.: you add a list of "Useful links" on your page). Does this improve the PageRank of  $X$ ? Does the answer change if you add links from  $Y$  or  $Z$  to older, popular pages?
- (e) Describe what steps you might take to raise the PageRank of  $X$  further. You do not need to prove anything here, just summarize your thoughts based on the previous parts. For extra credit though, you can prove what the structure for a link farm with  $m$  nodes should be to optimize the PageRank of  $X$ .

**Note:** As mentioned in class, we use the notation  $G = \alpha P + \frac{1-\alpha}{n}(\mathbf{1}_{n \times n})$  for the transition matrix. For a page that has  $k$  outgoing links, we put  $1/k$  for the corresponding entries of  $P$ . However, when a webpage has no outgoing links, we add a 1 as the corresponding diagonal element of  $P$  for making its row-sum one. Note that this makes  $G$  a valid transition probability matrix.

### 4 Beyond PageRank [45 points]

PageRank is one example of a definition of importance, a.k.a., *centrality*, that is used to capture the relative importance of pages in the web graph. Centrality is a much more general concept however, and is useful in many contexts. For example, we may be interested in knowing how important a person is in a social network, or how important a road is in a city's transportation infrastructure.

As you might expect, there are a number of different ways of defining the centrality of a node in a network, and each might induce a different ranking of the nodes in terms of 'importance'. How appropriate any one definition is depends on the application scenario. In this exercise, we'll introduce you to the four most common measures of centrality.

Let's start with the definitions. Consider a **connected, undirected** graph  $G$  with  $n$  nodes, labeled  $1, 2, \dots, n$ .

1. **Degree centrality** Let  $d_i$  denote the degree of node  $i$ . The degree centrality of node  $i$  is defined by

$$C_D(i) = \frac{d_i}{n-1}.$$

This definition of centrality assigns ‘importance’ to a node proportional to its degree.

2. **Closeness centrality** Let  $l(i, j)$  denote the distance (length of the shortest path) between nodes  $i$  and  $j$ . The closeness centrality of node  $i$ , denoted  $C_C(i)$ , is defined as the reciprocal of the average distance between  $i$  and all other nodes, i.e.,

$$C_C(i) = \frac{n-1}{\sum_{j \neq i} l(i, j)}.$$

According to this definition, a node that has, on average, shorter paths to other nodes is ‘more important’.

3. **Betweenness centrality** Let  $P(j, k)$  denote the number of shortest paths between nodes  $j$  and  $k$ , and  $P_i(j, k)$  denote the number of those shortest paths that pass through  $i$ . One can think of  $\frac{P_i(j, k)}{P(j, k)}$  as a measure of the ‘importance’ of  $i$  with respect to connecting  $j$  and  $k$ . Betweenness centrality attempts to capture how ‘important’ a node is with respect to connecting other nodes in the graph. The betweenness centrality of node  $i$  is defined as

$$C_B(i) = \frac{\sum_{j, k: j \neq k, j, k \neq i} \frac{P_i(j, k)}{P(j, k)}}{\binom{n-1}{2}}.$$

4. **PageRank** Suppose you replace each undirected edge  $(i, j)$  in  $G$  by two directed edges  $(i \rightarrow j)$  and  $(j \rightarrow i)$ . This gives us a directed, strongly connected graph. On this graph, we define the centrality of a node to be its PageRank, obtained by using the first cut PageRank algorithm you’ve seen in class (i.e., with  $\alpha = 1$ ). Recall that the vector of PageRanks  $r = (r_1, r_2, \dots, r_n)$  obtained using this algorithm will satisfy

- (i)  $r_i \geq 0$  for all  $i$ ,
- (ii)  $\sum_{i=1}^n r_i = 1$ ,
- (iii)  $r_i = \sum_{j \in N(i)} \frac{r_j}{d_j}$  for all  $i$ .

Here,  $N(i)$  is the set of neighbors of node  $i$  and  $d(i)$  is its degree.

Phew! With the definitions out of the way, let’s get what you have to do.

**Your task:**

- (a) Contrast the four definitions of centrality described above. Specifically, for each pair of definitions, either prove that for *any* connected, undirected graph  $G$ , the two definitions rank the nodes in the same order of importance, or give a counterexample that proves otherwise.

Hint: To get you started, we’ll tell you that the PageRank  $r_i$  we’ve described turns out to be proportional to the degree centrality  $C_D(i)$ . Of course, you still have to prove this! (Don’t be confused though, what we’re claiming is that for an *undirected* graph, PageRank is equivalent to degree centrality. This is certainly not true for the web graph, which is *directed*.)

- (b) For each of the centrality measures, describe a specific application setting where it best captures the notion of “importance” for the nodes.