CS 142: Network Sciences Prof. Lenore Cowen
Tufts University Spring 2025

HW 4: Pagerank
due March 13

For this assignment, submit your answers as a single pdf file named `hwk4.pdf`, alongside any code you may have written and all three requested output .txt files, using Gradescope. Submit your pdf to the assignment named ”Homework 4 PDF”. Submit your code and any generated networks to the assignment named ”Homework 4 Networks and Code”.

For this assignment, you will work with two webgraphs, Hollins and Blogs, available on the course website.

The hollins.edu webgraph is a directed graph of internal hyperlinks on the Hollins.edu web site circa 2004. It is provided to you as an edgelist. The first line gives the number of nodes and the number of edges, the next 6012 lines provide a space delimited map of node number to node web address, and the remaining lines provide space-delimited directed edges with the source node listed first and the destination node second.

The political blogs graph is a directed graph of blog links from the 2004 election. It is provided to you as an edgelist in DL format. Lines 5-1494 provide an ordered list of node row labels; lines 1496-2985 provide an ordered list of node column labels; lines 2987-223086 provide an edgelist where the first entry is a source node number, the second entry is a destination node number, and the third entry is a positive integer if that edge exists and 0 otherwise.

Please note: we say “a positive integer” but in fact, the third entry, when nonzero, is almost always 1 except for a handful of the entries that are marked as “2” rather than “1”. The meaning of “2” is completely undocumented. We believe that the “2”s are a susbset of edges that the researchers who generated this dataset flagged as interesting in their analysis, based on our own limited forensic attempts. Therefore, please treat these “2”s as the same as “1”s, i.e. just indicating the presence of an edge.

1. Implement Pagerank.
2. Run Pagerank on the hollins.edu webgraph. Provide a table of pagerank values for each node in the graph (each row should contain a tab-separated node name and Pagerank score) named `p1.txt` with your submission. What are the pages with the five highest and five lowest Pagerank values?

3. Run Pagerank on the political blogs graph. Provide a table of pagerank values for each node in the graph named `p2.txt` with your submission. What are the blogs with the five highest and five lowest Pagerank values?

4. Personalized Pagerank is identical to Pagerank, except there is some probability p of restarting the random walk from a specific source node (this is the node Personalized Pagerank is personalized for) after each step of the random walk. Run personalized pagerank for the following blogs, and provide a table of tab-separated table of personalized pagerank values named `p3.txt` with your submission. In this table, row i, column j should contain the pagerank value for blog i personalized for blog j. For each run of personalized pagerank, use 20 random walk steps and $p = 0.1$.

   - "dailykosc"
   - "atriosblo"
   - "wonkettec"
   - "talkleftc"
   - "juancolec"
   - "powerlineb"
   - "realclearp"
   - "blogsforbu"
   - "instapundi"
   - "michellema"

5. In the political blogs webgraph, the first 758 blogs are left-leaning (liberal) and the remaining 732 as right-leaning (conservative). Similarly, the first 5 blogs you produced personalized pagerank values for are left-leaning and the remaining 5 right-leaning. Can you detect any differences in behavior between left- and right- leaning blogs, either from the personalized pagerank values you produced in question 3 or from other properties of the graph?