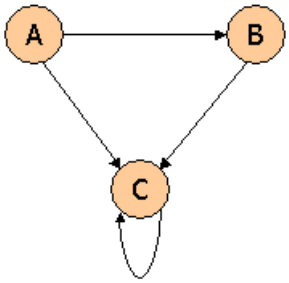
CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 5: Large Graph Processing

===Part 1===

1. Consider three Web pages with the following links:



Suppose we compute PageRank with a β of 0.7, and we introduce the additional constraint that the sum of the PageRanks of the three pages must be 3, to handle the problem that otherwise any multiple of a solution will also be a solution. Compute the PageRanks a, b, and c of the three pages A, B, and C, respectively.

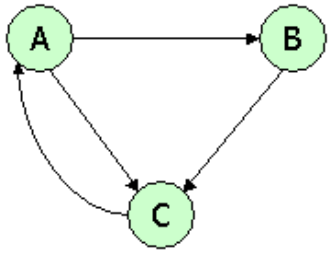
The rules for computing the next value of a, b, or c as we iterate are:

a <- .3
b <- .7(a/2) + .3
c <- .7(a/2+b+c) + .3

The reason is that a splits its PageRank between b and c, while b gives all of its to c, and c keeps all its own. However, all PageRank is multiplied by .7 before distribution (the "tax"), and .3 is then added to each new PageRank.

In the limit, the assignments become equalities. That immediately tells us a = .3. We can then use the second equation to discover b = .7*.3/2 + .3 = .405. Finally, the third equation simplifies to c = .7(.555 + c) + .3, or .3c = .6885. From this equation we get c = 2.295. It is now a simple matter to compute the subs of each two of the variables: a+b = .705, a+c = 2.595, and b+c = 2.7.

2. Consider three Web pages with the following links:



Suppose we compute PageRank with β=0.85. Write the equations for the PageRanks a, b, and c of the three pages A, B, and C, respectively.

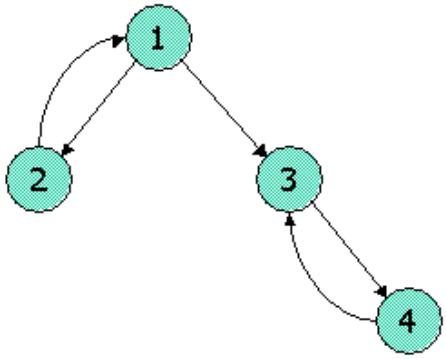
Here are the equations in the general form, where the PageRank of a node is set equal to β times the fair share of the PageRank of each predecessor of that node, plus (1-β) divided by the number of nodes (3), times the sum of the PageRanks.

a = .85c + .05a + .05b + 0.05c
b = .425a + .05a + .05b + 0.05c
c = .85b + .425a + .05a + .05b + 0.05c

If we simplify so there is only one term for each variable, we get:

.95a = .9c + .05b
.95b = .475a + .05c
.95c = .9b + .475a

3. Consider the following link topology.



Compute the Topic-Specific PageRank for the following link topology. Assume that pages selected for the teleport set are nodes 1 and 2 and that in the teleport set, the weight assigned for node 1 is twice that of node 2. Assume further that the teleport probability, (1 - beta), is 0.3.

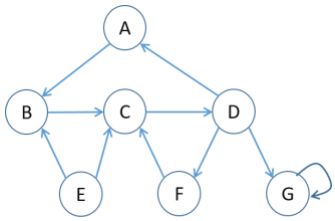
TSPR(1) = .3576
TSPR(2) = .2252
TSPR(3) = .2454
TSPR(4) = .1718

===Part 2===

1. Given the following graph,

1) how many dead ends are there in the graph? For each dead end (if any), please indicate the set of vertices forming the dead end.

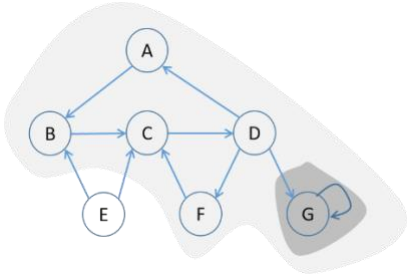
2) how many spider traps are there in the graph? For each spider trap (if any), please indicate the set of vertices forming the spider trap.



Answer:

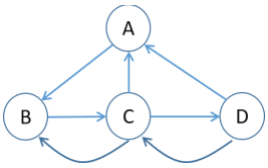
1) No dead ends

2) Two spider traps, shown below.



2. Set up the PageRank equations for the below graph, assuming β = 0.8 (jump probability = 1−β). Denote the PageRank of node *x* by *r(x)*.

Tutorial Solutions



Answer:

r(A)=0.8*(r(D)/2+r(C)/3)+0.2/4

r(B)=0.8*(r(A)+r(C)/3)+0.2/4

r(C)=0.8*(r(B)+r(D)/2)+0.2/4

r(D)=0.8*(r(C)/3)+0.2/4