Ordering Search Results

Q5.1 How can we solve the system of equations in the chapter?

To simplify everything, let's consider what each of the importance scores represent. They are the *probabilities* of each node being visited at each step of the random surfing process. And these probabilities must sum to 1 (since at each step in the process there is 100% chance that we are on one of the webpages):

$$w + x + y + z = 1.$$

This equation *normalizes* the sum of the importance scores, and is particularly convenient to work with. In fact, the easiest way to solve the system is to choose one importance score as a reference, get all other scores in terms of this one, and then plug the expressions back into this equation.

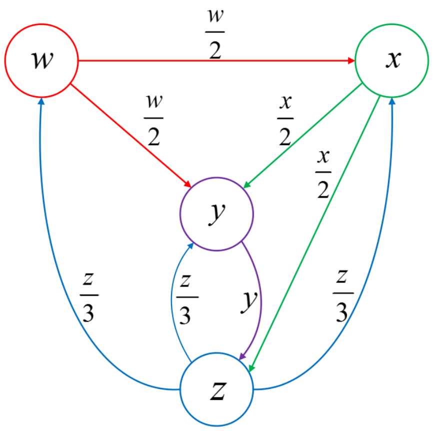

Illustration 15: The webgraph provides a convenient way of visualizing the equations for importance score.

Which variable should we choose as the reference? Well, from the first of our equations, we have w in terms of z already. As a result, it will be convenient to get everything in terms of z. We can use the second and fourth equations (the simpler ones) to do this.

Note, however, that the choice of z here is not a necessity; we could choose any of the others as the reference and get the same solution,

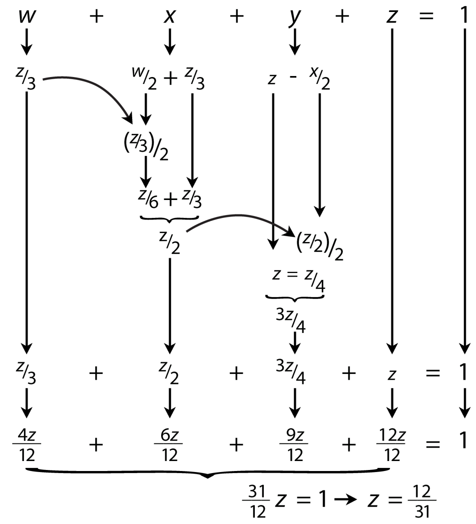

Illustration 16: The steps involved in computing the importance score of Page Z in our example.

but it would require a bit more algebra.

From the first equation, we have that

$$w = \frac{z}{3}.$$

Now, how can we get x in terms of w? We can use this relationship to substitute for w in the second equation:

$$\begin{aligned} x &= \frac{w}{2} + \frac{z}{3} \\ &= \frac{z}{6} + \frac{z}{3} \\ &= \frac{z}{2}. \end{aligned}$$

Now that we have x in terms of z, how can we use the fourth equation to do the same for y? First, we re-arrange to get y on the left:

$$y = z - \frac{x}{2}$$

Then, we can substitute for x, using our previous result that $x = z/2$:

$$\begin{aligned} y &= z - \frac{z}{4} \\ &= \frac{3z}{4}. \end{aligned}$$

With all the importance scores now in terms of z, we can plug into the normalization equation:

$$\begin{aligned} 1 &= w + x + y + z \\ &= \frac{z}{3} + \frac{z}{2} + \frac{3z}{4} + z \\ &= \frac{31z}{12}. \end{aligned}$$

With this, it follows that $z = 12/31 = 0.387$.

Now, we can backtrack and solve for the rest of the importance scores:

$$\begin{aligned} w &= \frac{z}{3} = \frac{4}{31} = 0.129, \\ x &= \frac{z}{2} = \frac{6}{31} = 0.194. \\ y &= \frac{3z}{4} = \frac{9}{31} = 0.290. \end{aligned}$$

For clarity, we illustrate the steps in computing z visually in Illustration 16.

Q5.2 What are dangling nodes, and what special treatment is needed in PageRank to account for them?

There will be no solution to the problem if the webgraph has one or more **dangling nodes**. A dangling node is one that does not point to any other webpages. For instance, if we augment our example in the book with the fifth node V as in Illustration 17, then V is a dangling node.

Thinking about the PageRank procedure from before, what are the implications of this structure on V's importance score (call it v)? Well, by definition, PageRank has each node spread a portion of its importance score to all of its outgoing neighbors. With a single outgoing

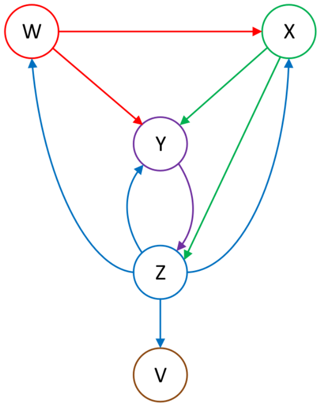

Illustration 17: Here, we have modified our original example to include a dangling Node V. It has an incoming link from Node Z, but no outgoing links, which requires its importance score to be zero. The solution to this is to assume that V has a link to every node, rather than no nodes.

link, that link gets all the score, with two, each gets half, and so on. As such, the sum of the importance of the outgoing links is equal to the node's importance. But V has *no* outgoing links: as a result, v must be 0 for the equations to work out.

Now, what does $v = 0$ imply for the rest of the graph? Well, Page Z spreads one-fourth of its importance score to V, and as a result $v = z/4$. So if v is zero, then z must be zero too! You can verify that this logic cascades through the rest of the graph, requiring the scores for the rest of the nodes to be zero as well.

PageRank's solution to this problem is to assume that each dangling node has an outgoing link to *every* node (including itself), rather than none. This is intuitive: if a random surfer lands on a page without any hyperlinks, she would have to enter some other link into the browser to keep on going.

Q5.3 What is a connected component, and what special treatment in PageRank is needed to handle these?

Even if we apply the fix for dangling nodes, there will be many solutions if the webgraph has more than one **connected component**. A

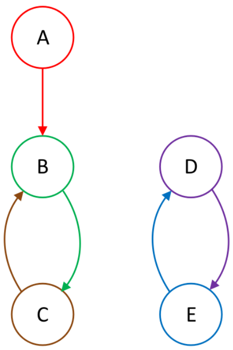

Illustration 18: Unconnected graphs are problematic, because they allow for an infinite number of solutions in the PageRank algorithm. This is dealt with by adding some randomization into the procedure, so that a surfer can get between the different components of the graph.

connected component is simply a group of nodes in which any two can reach each other (directly or indirectly), but none can reach any outside of the group. For instance, in Illustration 18, Pages A, B, and C form one connected component, and Pages D and E form another. The problem here is that a random surfer can't get from one component to the other: depending on which she starts in, she would be stuck in that **subgraph** forever. As a result, we have no way of relating the importance of nodes in one component to those in the other, which causes the problem to be mathematically underspecified (*i.e.*, many potential solutions).

The solution to this is precisely what we neglected from the random surfer concept initially: the surfer might "get bored" during the process, and enter some random website into the browser. So for a certain portion of the time, say 85%, she will follow the webgraph in deciding where to go next, and for the remaining 15% of the time, she will randomly choose from all the available webpages.