PROFESSOR: PageRank is a measure of the importance of a web page. But let me immediately correct my own confusion that I suffered from for some time until very recently, which is that even though PageRank is used for ranking pages, it's called PageRank after its discoverer, developer, Larry Page, was one of the co-founders along with Serg Brin of Google.

So the motivation is that when you-- at least before Google, when you did a standard retrieval on a web page using keyword Search and similar kinds of criteria, you'd get back millions of hits, most of which were really low quality and you weren't interested in, and with a few useful pages buried in the millions. And the question was, all of these documents are indistinguishable in terms of keyword search and textual patterns, how do you figure out which are the important ones. And the idea that Page came up with was to use the web structure itself, the structure of the worldwide web, to identify the important documents.

So we can think of the whole internet as a graph where a user is on a page, and we think of a URL as a link to another page, as a directed edge. And users are kind of randomly traveling around in the worldwide web. They're at a page, they randomly click a link to get to another page, and they keep doing a walk on the web graph. And every once in a while, they're going to find that the thread that they're on is kind of losing steam, where they find themselves in some kind of a cycle and they will randomly start over again at some other page. And we want to argue or hypothesize that a page is more important when it's viewed a large fraction of the time by these random browsers and random users.

So to be formal, we're going to take the entire worldwide web, trillions of vertices, as a digraph. And there's going to be an edge from one URL to another, from V to W, if there's a link from the page V to the page W, or the URL W. W might not even be a page, it might be a document, which means it doesn't have any links on it. But for the real vertices are the web pages that have links on them.

OK, that's the model. And we're going to make it into a random walk graph by saying that if you look at a URL V, at a vertex V, all of the edges out of it are equally likely. It's a simple model, and it might or might not work. But in fact, it did work pretty well. That is the model of the worldwide web as a random walk graph.

So to be more precise, the probability of the edge that goes from V to W is 1 over the out

degree of V. That is, all of the out degree of V edges leaving vertex V get equal weight. Now to model this aspect that the users start over again if they get bored or they get stuck, we can formally add to the digraph a hypothetical super-node, which-- and with the property that there's an edge from the super-node to every other node with equally likelihood. So once you hit the super-node then following an edge is tantamount to saying, pick a random page and start over again.

To get to the super-node, we have edges back from other nodes in the graph back to the super-node. In the reading, we said that we were going to have nodes back from terminal nodes that had no edges out. For example, a document or something like that. That's actually not sufficient, because-- for the PageRank to work in the theoretical way that we want it to because even if there is no dead nodes, you might be in a clump of nodes which you can't get out of. And you'd want to be able to-- and even though none of them was a dead end, because they all had arrows going out to each other.

And so you'd really want a node from a-- an edge from a clump like that back to the super-node to model starting over there. The simplest way to do it really is to simply say that there's an edge to the super-node from every vertex. So wherever you are, you can randomly decide to start over. And Page and Brin and their co-authors in the original paper on PageRank suggested that the edge back from a vertex to the super vertex might get a special probability. It might be customized, as opposed to being equally likely with all of the other edges leading a vertex. In fact, I think they decided that there should be a 0.15 probability from each vertex of jumping at random to the super-node.

OK. Let's just illustrate this with an example. This is a random walk graph that we've seen before modeling coin flipping. And when I add the super-node, there's this one new vertex super, and there's an edge from the super vertex to every other one of the vertices in the graph. And from each vertex in the graph, there is an edge going back. I've illustrated that with two-way arrows. So this is really an arrow with two arrowheads. It represents an arrow in each direction.

Now in the original paper, actually, Page didn't talk about a super vertex. Instead, he talked about each vertex randomly jumping to another vertex. But that would just get the whole state diagram completely clogged up with edges, so it's more economical to have everybody jump to the super vertex and the super vertex jump back to everybody. And that saves a significant number of edges.

So PageRank, then, is obtained by computing a stationary distribution for the worldwide web. So s bar is a vector of length trillions that the coordinates are indexed by the web pages. And we want to calculate the stable distribution. And then we'll simply define the page rank of a page is its probability of being there in the stationary distribution, the v component of the stable-- stationary distribution, s. And of course, we'll rank v above s when the probability of being in v is higher than the probability of being in w.

By the way, I don't have the latest figures, but there were-- I guess I've heard people who've worked for Google say, and in some of the Wikipedia articles, that it takes a few weeks for the crawlers to create a new map of the web, to create the new graph. And then it takes some number of hours, I think under days, to calculate the stationary distribution on the graph, doing a lot of parallel computation.

So a useful feature about using the stationary distribution is that ways to hack the links in the worldwide web to make a page look important are-- will not work very well against PageRank. So for example, one way to look more important is to create a lot of nodes pointing to yourself, fake nodes. But that's not going to matter, because the fake nodes are not going to have much weight since they're fake and nobody's pointing to them. So even though a large number of fake nodes point to you, their cumulative weight is low, and they're not adding a lot to your own probability. Likewise, you could try taking links to important pages and try to make yourself look important that way, but PageRank won't make you look important at all if none of those important nodes are pointing back. So both of these simple-minded ways to try to look important by manipulating links won't improve your page rank.

The super-node is playing a technical role in making sure that the stationary distribution exists. So it guarantees that there's a unique stationary distribution, s bar. By the way, I sometimes use the word stable and sometimes stationary. They're kind of synonyms, although I think officially we should stick to the word stationary distribution. As I've mentioned before, when a digraph is strongly connected, that is a sufficient condition for there to be a unique stable distribution. That's actually proved in one of the exercises in the text at the end of the chapter.

The super-node mechanism also ensures something even stronger, that every initial distribution p converges to the stationary distribution, to that unique stationary distribution. Stated precisely mathematically, if you start off at an arbitrary distribution of probabilities of being in different states, p, and you look at what happens to p after t steps-- remember, that

you get by multiplying the vector p by the matrix M raised to the power t-- and you take the limit as t approaches infinity, that is to say, what distribution do you approach as you do more and more updates. And it turns out that that limit exists, and it is that stationary distribution. So it doesn't matter where you start, you're going to wind up stable. And as a matter of fact, the convergence is rapid. What that means is that you can actually calculate the stable distribution reasonably quickly, because you don't need a very large t in order to arrive at a very good approximation to the stable distribution.

Now the actual Google rank and ranking is more complicated than just PageRank. PageRank was the original idea that got a lot of attention. And in fact, the latest information from Google is that they think it gets overattention today in the modern world by too many commentators and people trying to simulate ranking. So the actual rank rules are a closely-held trade secret for Google-- by Google. They use text, they use location, they use payments, because advertisers can pay to have their search results listed more prominently, and lots of other criteria that have evolved over 15 years. And they continue to evolve. As people find ways to manipulate the ranking, Google revises its ranking criteria and algorithms. But nevertheless, PageRank continues to play a significant role in the whole story.