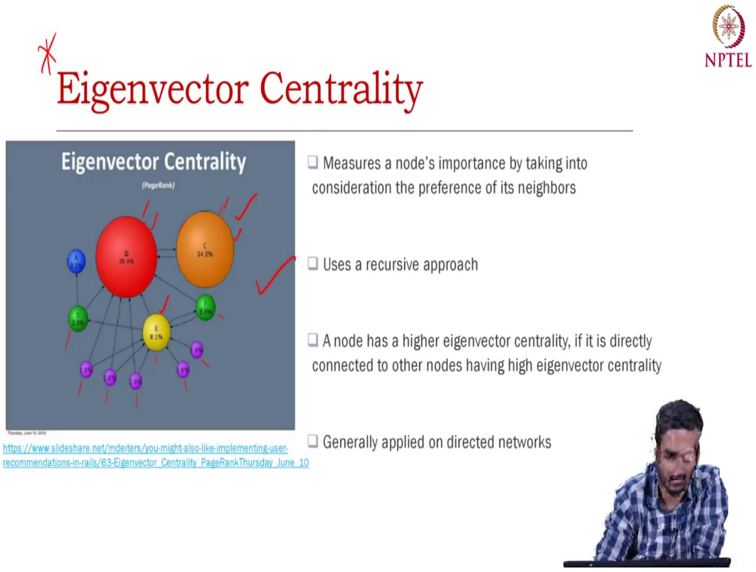
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 02
Lecture - 10
Lecture - 05

(Refer Slide Time: 00:30)

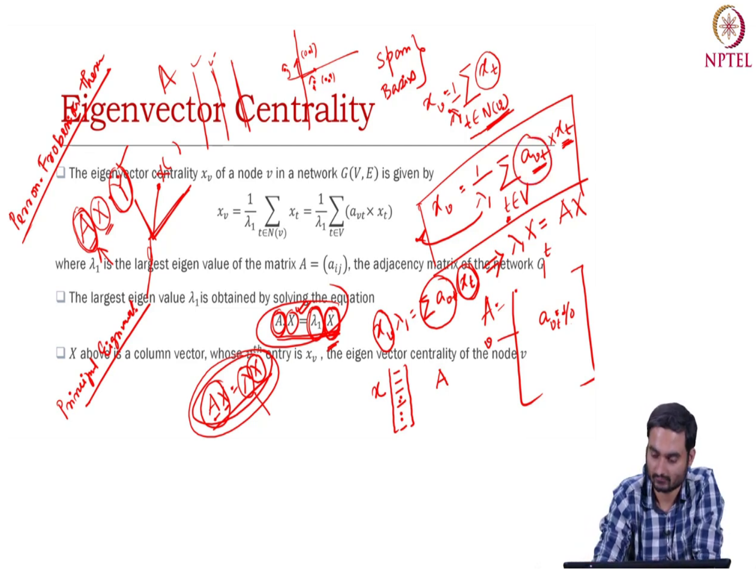


Eigenvector Centrality

https://www.slideshare.net/mdeiters/you-might-also-like-implementing-user-recommendations-in-rails/63-Eigenvector_Centrality_PageRankThursday_June_10

- ❑ Measures a node's importance by taking into consideration the preference of its neighbors
- ❑ Uses a recursive approach
- ❑ A node has a higher eigenvector centrality, if it is directly connected to other nodes having high eigenvector centrality
- ❑ Generally applied on directed networks

(Refer Slide Time: 00:35)



Eigenvector Centrality

- ❑ The eigenvector centrality x_v of a node v in a network $G(V,E)$ is given by

$$x_v = \frac{1}{\lambda_1}\sum_{t\in N(v)} x_t = \frac{1}{\lambda_1}\sum_{t\in V}(a_{vt}\times x_t)$$

where λ_1 is the largest eigen value of the matrix $A = (a_{ij})$, the adjacency matrix of the network G

- ❑ The largest eigen value λ_1 is obtained by solving the equation

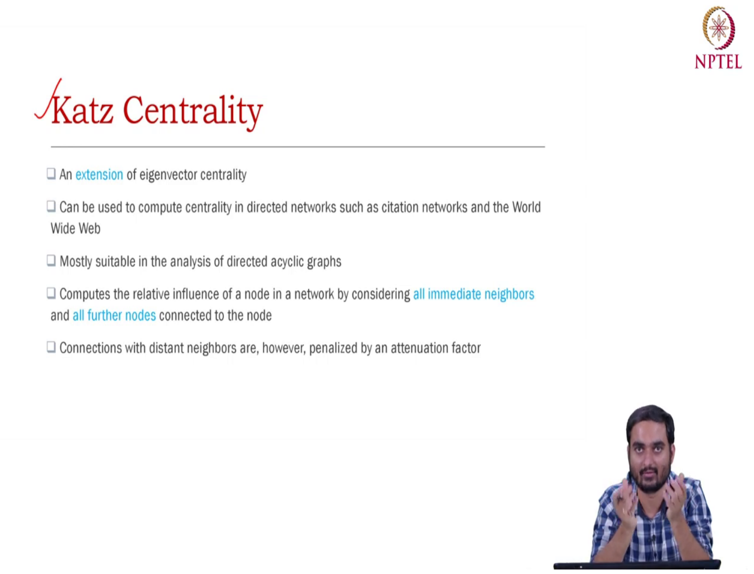
$$A.X = \lambda_1 X$$

- ❑ X above is a column vector, whose v^{th} entry is x_v , the eigen vector centrality of the node v

Perron-Frobenius theorem
Principal Eigenvector
Span Basis
$x_v = \frac{1}{\lambda_1}\sum_{t\in N(v)} x_t$
$x_v = \frac{1}{\lambda_1}\sum_{t\in V} a_{vt} \times x_t$
$\lambda X = AX$
$AX = \lambda X$
$x_v\lambda_1 = \sum a_{vt} x_t$
$a_{vt} = 1/0$
NPTEL

So, in the last lecture we have stopped at discussing the eigenvector centrality right, let me just quickly go back if you remember this figure right. So, there we have only considered the you know immediate neighbors, the one hop neighbors. So, therefore, we have used this adjacency matrix entry A v t right.

(Refer Slide Time: 00:45)



Katz Centrality

❑ An extension of eigenvector centrality

❑ Can be used to compute centrality in directed networks such as citation networks and the World Wide Web

❑ Mostly suitable in the analysis of directed acyclic graphs

❑ Computes the relative influence of a node in a network by considering all immediate neighbors and all further nodes connected to the node

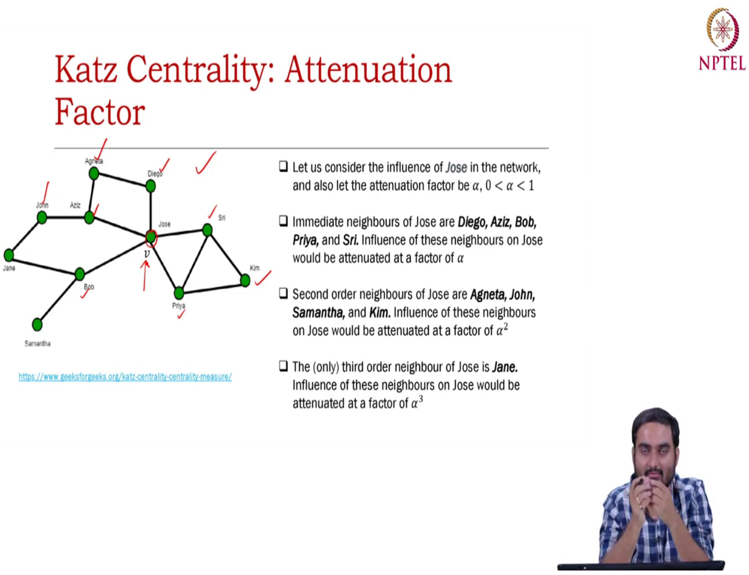
❑ Connections with distant neighbors are, however, penalized by an attenuation factor

But, what happens is that you know this is basically you can think of it as a flow of prestige from nodes which are which basically are placed in the first hop, second hop, third hop. So, from third hop you are getting some information, from second hop you are also getting some information, the first hop you are getting some prestige right. So, it should not depend on the first hop neighbors ok.

So, therefore, there was another metric proposed called Katz centrality. Katz centrality is an extension of eigenvector centrality which basically says that instead of only looking at the first hop neighbor, let us also look at the second hop neighbor, third hop neighbor. Of course, when you look at the second hop neighbor, the second hop neighbor should not contribute to the importance of the given node compared to the case in the first hop.

So, first hop neighbors will contribute more, second hop neighbors will contribute little less, third hop neighbor will contribute even less, even lesser than the second hop neighbor and so on and so forth right.

(Refer Slide Time: 01:55)



Katz Centrality: Attenuation Factor

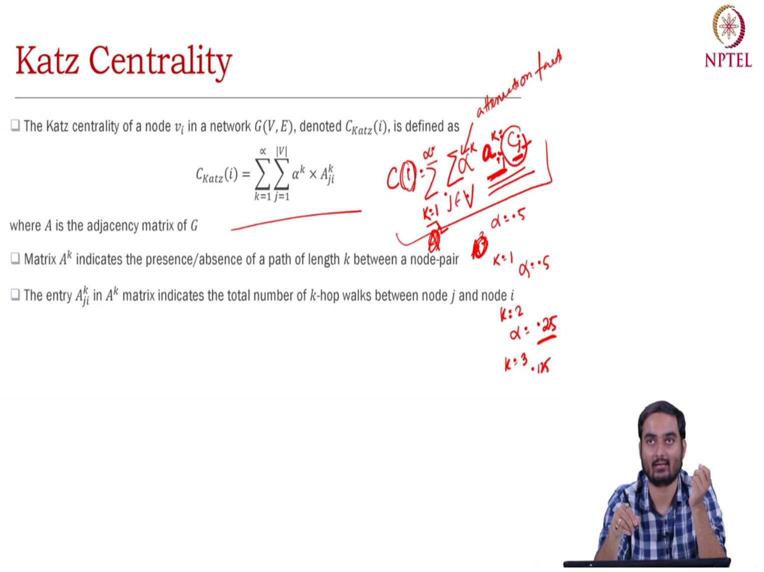
https://www.geeksforgeeks.org/katz-centrality-centrality-measure/

- ❑ Let us consider the influence of *Jose* in the network, and also let the attenuation factor be α, $0 < \alpha < 1$
- ❑ Immediate neighbours of Jose are ***Diego, Aziz, Bob, Priya,*** and ***Sri.*** Influence of these neighbours on Jose would be attenuated at a factor of α
- ❑ Second order neighbours of Jose are ***Agneta, John, Samantha,*** and ***Kim.*** Influence of these neighbours on Jose would be attenuated at a factor of α^2
- ❑ The (only) third order neighbour of Jose is ***Jane.*** Influence of these neighbours on Jose would be attenuated at a factor of α^3

So, the idea is very simple right. So, let us think of this figure, this network and you want to compute Katz centrality for node v right. Of course, node v will get more importance from the immediate neighbors like Diego, Aziz, Bob right, Priya Sri right. But, the node v will also receive some prestige from the second hop neighbors like this one, like John right, like Kim ok.

So, how do we capture this one? How do we capture in the formulation? And, remember as we go as we move away from the given node, the contribution will decrease; we also need to take this thing into account right.

(Refer Slide Time: 02:49)



Katz Centrality

- ❑ The Katz centrality of a node v_i in a network $G(V, E)$, denoted $C_{Katz}(i)$, is defined as

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{|V|} \alpha^k \times A_{ji}^k$$

where A is the adjacency matrix of G

- ❑ Matrix A^k indicates the presence/absence of a path of length k between a node-pair
- ❑ The entry A_{ji}^k in A^k matrix indicates the total number of k-hop walks between node j and node i

So, this is the formula. So, far we have we have only considered you know say this is the importance matrix for node i right. So, far we have considered A ij entry or you know if I use the same notation say small a ij entry and c j right, where j is j belongs to the vortex set right. What I am saying is now is that let us also look at the second hop neighbor. So, how do I know that a node has distance 2 and node resides at the second hob from a given node? It is very simple.

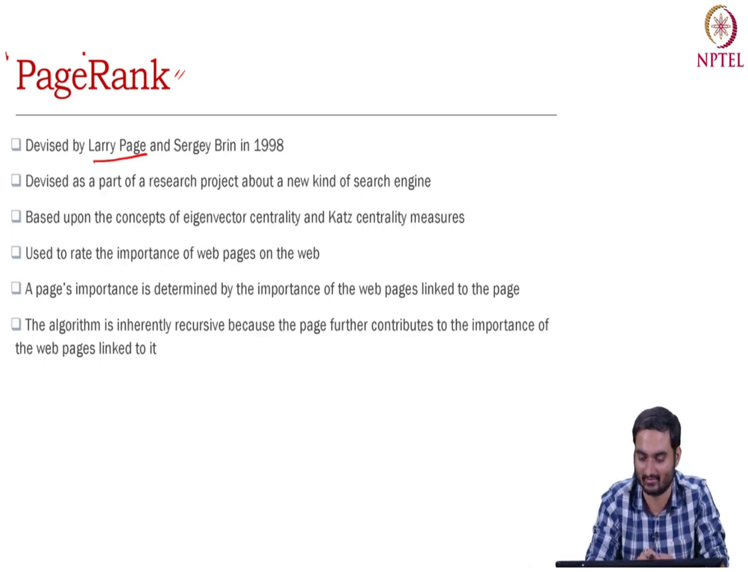
So, this adjacency matrix A will give you ideas about the first hop right, whether 2 nodes are adjacent. If you multiply A with itself, if you take the square of A right; you multiply adjacency matrix with itself; it will actually give you the second hop neighbors whether 2 nodes A and B are connected through 2 hops ok. Similarly, A 3 will basically tell you whether 2 2 3 whether 2 nodes are connected through 3 hops and so on and so forth right.

So, what I am doing now, I am adding another notion which is the distance or hop and so, k ranges from 1 to infinity. So, I have 1 hop, 2 hop, 3 hop dot dot dot infinite number of hops right of course, it is not possible, but let us take the maximum right. So, when I take the first hop neighbor k would be 1, when I take the second hop neighbor this would be 2 and so on and so forth ok. Along, with this I am also using a factor alpha, this alpha factor its a constant, it is called attenuation factor right.

It is called attenuation factor, alpha ranges between 0 to 1 ok, alpha to the power k. So, when say alpha is 0.5. So, when k equals to 1, meaning you are looking at the first hop alpha will be 0.5, when k equals to 2 alpha would be 0.25 right. So, 0.25 will be multiplied with this one. Now the what is this? This is the; this is the eigenvector centrality of node j right, if j is at the second hop; the contribution would be lesser because now this is multiplied by 0.25 right. If k equals to 3, it would be 0.125 even lesser.

So, as you move away from particular node v from a particular node i, you are basically giving less weightage to the contribution ok. Again, you can easily map it in terms of matrix concrete matrix right, in terms of A x etcetera and it would also give you ideas about; I mean it would also lead to eigenvector eigen value in a different form. But, the problem here is that the matrix multiplication A, A square, A cube that is extremely time consuming right. Therefore, we do not use Katz centrality in general.

(Refer Slide Time: 06:26)



PageRank

❑ Devised by Larry Page and Sergey Brin in 1998

❑ Devised as a part of a research project about a new kind of search engine

❑ Based upon the concepts of eigenvector centrality and Katz centrality measures

❑ Used to rate the importance of web pages on the web

❑ A page's importance is determined by the importance of the web pages linked to the page

❑ The algorithm is inherently recursive because the page further contributes to the importance of the web pages linked to it

Now, we move to the very important concept called PageRank. Now, this PageRank was something, the PageRank matrix is something based on which the company like Google was formed in 1998, 1999. So, what is PageRank? The idea behind PageRank actually you know comes from eigenvector centrality, but in a different manner right. So, let me first briefly talk about the history behind the PageRank metric and then I will define what is PageRank.

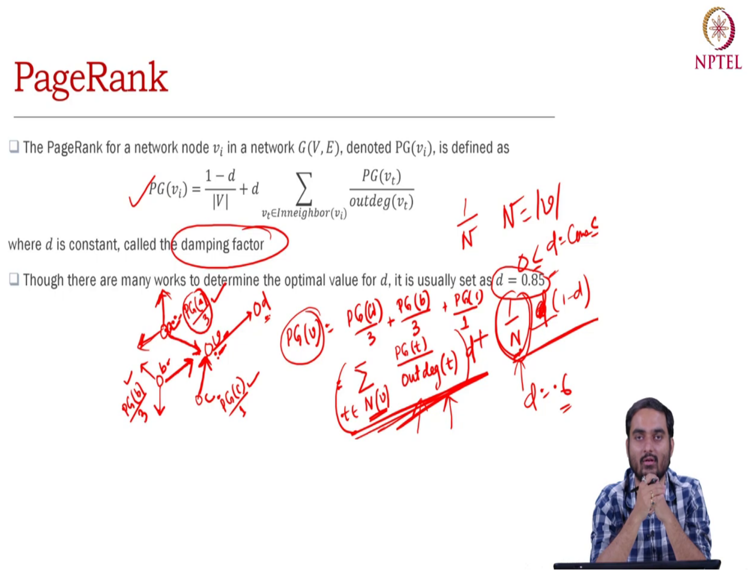
So, PageRank was proposed by Larry Page and Sergey Brin you all know right, the founders of Google. So, Larry Page and Sergey Brin; so, Sergey Brin and Larry Page they basically developed PageRank at Stanford, when they were undergrad students in 1996 and this was a part of their course project ok. And in fact, you know there are rumors I do not know whether this is true or not, but there are rumors that that project was brutally rejected by the course instructor right; by saying that this is not that exciting project.

Later they kept on developing the I mean on the ideas and in 1998, they I mean along with some other coauthors of course, one was the great late Professor Rajeev Motwani, who was the professor at Stanford. Along, with the other coauthors they published this paper in 1998 in worldwide web conference and the rest is history right.

So, based on the notion of PageRank Google was formed, initially you know the entire you know the patent who was filed by Stanford University, then later on Google bought the patent from Stanford University. And, you know and it basically Stanford sold that entire patent to Google with around 330 million US dollar right in 2004-2005.

So, this PageRank is I mean the name PageRank, I am not sure whether the name PageRank comes from Larry Page or PageRank comes from web page right. Because, the PageRank is used to measure the importance of web pages right, but you know this is what it is.

(Refer Slide Time: 09:06)



PageRank

❑ The PageRank for a network node v_i in a network $G(V,E)$, denoted PG(v_i), is defined as

$$PG(v_i) = \frac{1-d}{|V|} + d \sum_{v_t \in Inneighbor(v_i)} \frac{PG(v_t)}{outdeg(v_t)}$$

where d is constant, called the damping factor

❑ Though there are many works to determine the optimal value for d, it is usually set as $d = 0.85$

NPTEL
$\frac{1}{N}$
$N = |V|$
$0 < d$: Cons
$PG(v) = \frac{PG(a)}{3} + \frac{PG(b)}{3} + \frac{PG(c)}{1}$
$\sum_{t \in N(v)} \frac{PG(t)}{outdeg(t)}$
$\frac{1}{N}(1-d)$
$d = .6$

So, let us look at the PageRank formulation. So, remember PageRank was developed to measure the importance of web pages, you think of this the entire world wide web as complex network. I mentioned this thing in the 1st lecture, where you where you have different web pages as nodes and 2 web pages are connected through links and these links are directed links, these links are basically hyperlinks right.

So, PageRank basically is designed to mimic the way we browse you know web pages. So, how do we let us think of the browsing pattern. How do we; how do we land up to a particular web page? How do we open a web page? There are two ways to open a web page right. One way is that you randomly you know type the URL of the web page and you open it ok, or you basically follow you follow a hyperlink which was there in the previous page and through that hyperlink, you basically open the new page right. There are two ways.

So, these two factors are combined in this PageRank formulation. So, let us look at it carefully. So, let us look at the second factor first which is the you know you basically open a web page through an hyperlink ok. So, let us assume that this is your network, this is the World Wide Web network and right and say it looks like this, something like this. So, this is v, a, b, c right. So, you can open the web page v either I mean you may land up to v either from a or from b or from c.

In other words, if you think of again the eigenvector centrality, what you are saying is that when v receives when I mean since a, b and c are linking to v right, there are inward edges from a to v, b to v and c to v; as if v is receiving importance from a from b from c. So, v's importance is derived by a's importance, b's importance and c's importance right, but if you assume that importance flows through outward edges right.

So, b's entire importance will not move from b from, I mean a's entire importance will not move from a to v because, a has 3 outward edges right. And, if I assume that the importance is divided across outward edges uniformly right. So, the importance of say let us say the importance of node a is PG a some value. So, it will be divided by 3, because there are 3 outward edges right. So, PG a by 3, this much importance will basically come to v.

Similarly, if we think of b right, b also has 3 outward edges. So, PG so, PG is PageRank ok. So, PG is so, so PG b will also be divided by 3 and that quantity will come to v, but for c this would be PG c by 1, because there is only 1 outward edge. So, v will accumulate this quantity, this quantity and this quantity. So, now, PageRank of v would be PageRank of a by 3 plus PageRank of b by 3 plus PageRank of c by 1.

If I if you write it in a standard notation, what I am saying is that you are basically looking at all the neighbors or the all the immediate neighbors of v. So, neighbors of v right and let us say; let us say t is one such neighbor right. PageRank of t divided by out degree of t and then you take the sum ok. Now, remember these neighbors are basically inward neighbors, because they are inward neighbors right; a, b, c are inward neighbors.

Now, let us say there is an edge from v to d; so, d is an outward neighbor. So, that the d will not be considered while computing the PageRank of v ok. So, this is the first quantity. So, you are basically saying that I am moving through a through a hyperlink and I am actually opening a page video which is d. There is another quantity which basically says that I just randomly open node v right.

So, if I assume that all the nodes are equally likely to be opened ok, I can open any node with the same probability; say the probability is; so, say there is N number of nodes, N is mode v. So, if you so, if you think of uniform probability then all the nodes will be any node will be opened with the probability 1 by N ok.

So, the total probability; so, the total PageRank of v would be this plus 1 by N right. If you think of the two quantities this is one quantity, this is another quantity. So, the total probability would be total PageRank could be this plus this ok.

But, I do not want to give equal weightage is to both this and this ok. In fact, I want to give some weightage to this and some weightage to this ok, for that I use a weight d. So, I multiply this by say I multiply this by d and I multiply this by 1 minus d; remember d is a constant, d is a constant whose value ranges between 0 to 1 ok. So, let us say d equals to 0.6 right. So, I want to give 0.6 importance to this one and 0.1 importance to this one.

Remember PageRank always ranges between 0 to 1 ok; kind of an weighted average. Now, what is d? d is called the damping factor, d is called the damping factor. This is just a constant which tries to balance out the two components ok. In certain applications, if you want to give more weightage to; more weightage to your inward neighbors; so, d would be higher in that case right. So, it turned out again there is no concrete reason why I mean how to choose the value of d, this is basically a hyper parameter.

But, it turned out that if the PageRank will be effective, if you choose d equal to 0.85. Again, there is no reason behind choosing 0.85, but in the original paper Larry Page and Brin, they use d equals to 0.85 and people generally use d, I mean assign d between 0.7 to 0.85, 0.9 ok. So, this is the; this is the notion behind PageRank.

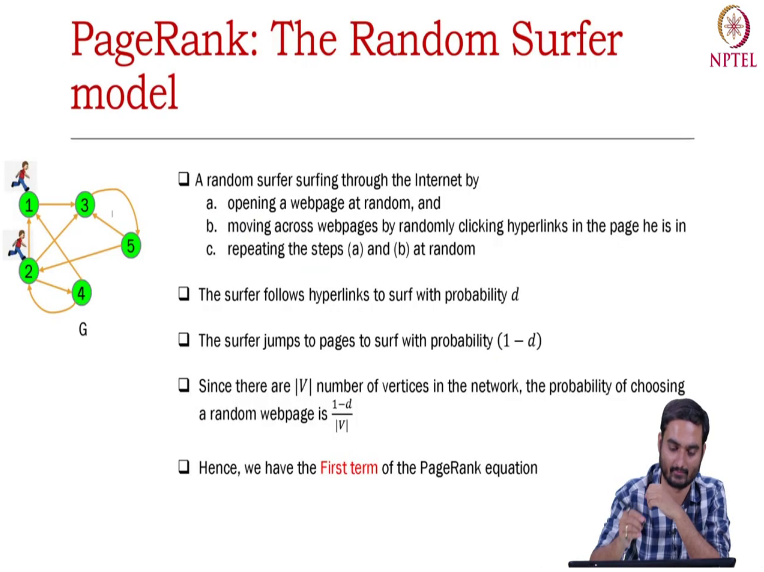
So, of course, there is another notion, you can relate PageRank to something called random walk. I will discuss you know that notion in another chapter called link analysis later, I think 4th or 5th chapter. But, you know you can actually explain PageRank in different way. This is one explanation random walk based approach explanation is another way of explaining PageRank.

Remember PageRank is PageRank the metric PageRank was highly influenced by several other matrix proposed earlier. For example, you know the citation metric right. So, Garfield who is very I mean one of the prominent researchers in the area of scientometrics, bibliographic analysis, bibliometrics and so on. So, Garfield proposed similar kind of matrix in 1950s ok.

So, PageRank was actually proposed by I mean was influenced by different citation analysis, because ultimately you see that whenever you see a directed network like a World Wide Web network or a citation network, you can use PageRank. You can also use PageRank in undirected network, while you can assume that are an undirected edge is basically a bidirectional; an undirected edge can be formed by two bidirectional edges, from a to b and b to a right.

So in fact, around the same time in 1998, 1999 another important metric was proposed, it is called HITS centrality, I will discuss about HITS right. But, you know you think of that time period 1996 till 2000, 2002 whole bunch of metrics you know have been proposed for quantifying the importance of nodes ok.

(Refer Slide Time: 18:48)

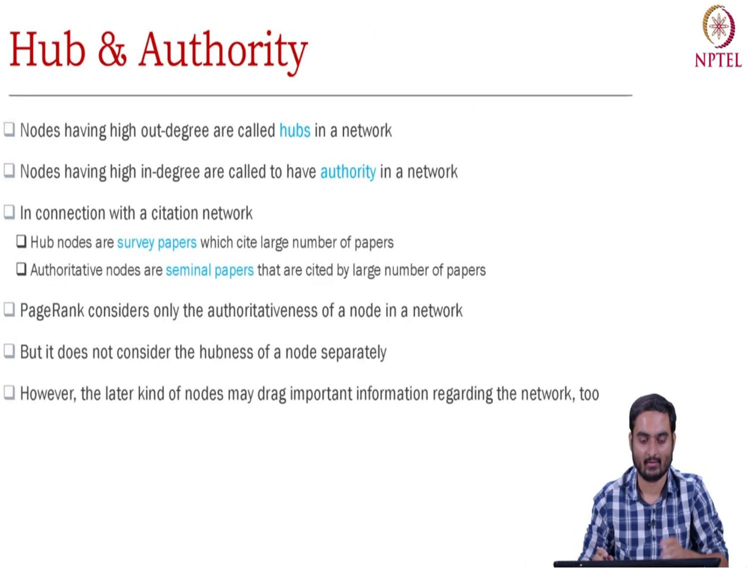


PageRank: The Random Surfer model

- ❑ A random surfer surfing through the Internet by
  a. opening a webpage at random, and
  b. moving across webpages by randomly clicking hyperlinks in the page he is in
  c. repeating the steps (a) and (b) at random
- ❑ The surfer follows hyperlinks to surf with probability d
- ❑ The surfer jumps to pages to surf with probability $(1-d)$
- ❑ Since there are $|V|$ number of vertices in the network, the probability of choosing a random webpage is $\frac{1-d}{|V|}$
- ❑ Hence, we have the First term of the PageRank equation

So, this random surfer model, we will random walk model, we will discuss in the later chapter.

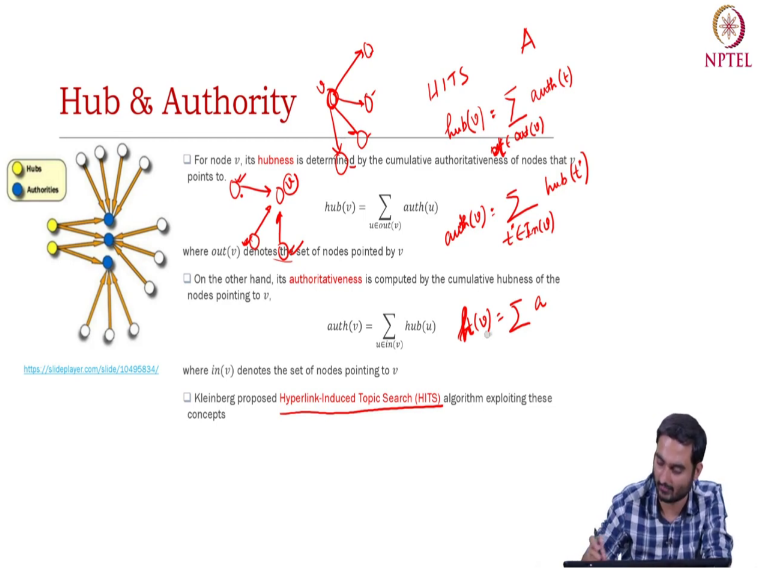
(Refer Slide Time: 18:57)



Hub & Authority

- ❑ Nodes having high out-degree are called hubs in a network
- ❑ Nodes having high in-degree are called to have authority in a network
- ❑ In connection with a citation network
  - ❑ Hub nodes are survey papers which cite large number of papers
  - ❑ Authoritative nodes are seminal papers that are cited by large number of papers
- ❑ PageRank considers only the authoritativeness of a node in a network
- ❑ But it does not consider the hubness of a node separately
- ❑ However, the later kind of nodes may drag important information regarding the network, too

(Refer Slide Time: 19:06)



Hub & Authority

□ For node v, its hubness is determined by the cumulative authoritativeness of nodes that v points to.

$$hub(v) = \sum_{u \in out(v)} auth(u)$$

where $out(v)$ denotes the set of nodes pointed by v

□ On the other hand, its authoritativeness is computed by the cumulative hubness of the nodes pointing to v,

$$auth(v) = \sum_{u \in in(v)} hub(u)$$

where $in(v)$ denotes the set of nodes pointing to v

□ Kleinberg proposed Hyperlink-Induced Topic Search (HITS) algorithm exploiting these concepts

https://slideplayer.com/slide/10495834/

$hub(v) = \sum_{t \in out(v)} auth(t)$

$auth(v) = \sum_{t' \in In(v)} hub(t')$

$A(v) = \sum a$

So, let us keep it for the time being. So, now, we will discuss a similar metric called HITS, H I T S: Hyperlink Induced Topic Search. So, this metric proposed by Jon Kleinberg from Cornell. So, again around the same time 1998, I guess. So, what is the idea here and how it is different from PageRank for example?

So, in PageRank we are only looking at the inward neighbors right. So, what Kleinberg told that look in a web page in a web page network right, there are nodes which have very high which have a lot of outward edges, high out degree right. And, there are nodes which have lot of inward edges. In the PageRank computation, you will always give more weightage to nodes which have high inward edges, but you should also give more weightage to nodes which have high outward edges.

Why? These outward edges think of a node which has a lot of out-degree right, you can think of this node as a hub node which is pointing to many outward edges, many outward nodes right. In a citation network you can think of this kind of nodes as say a survey paper or a book. In a book you see thousands of references, in a survey paper you see you know 200, 300 references. These references are links by the way in the citation network ok.

So, these books, survey papers they are also important to understand the overall area right in general whereas, of course, if a node, if a paper has a lot of inward citations right, that papers are of course important. So, we should actually look at these two quantities side by side. So, Kleinberg said that let us quantify two notions: one is called the hubness and the other is called the authoritativeness right, hubness and authoritativeness.

So, these two notions, let us assume that these two notions are there for every for every node. So, far we have only considered the notion of authority of a node based on the inward edges, now you consider another notion called hubness of a node ok. So, for every node, we will get two quantities. One is called the hubness, another is called authoritativeness ok.

So, he said that survey papers are also important which has a lot of outward edges. And of course, seminal papers are very important which also have which have lot of inward edges right.

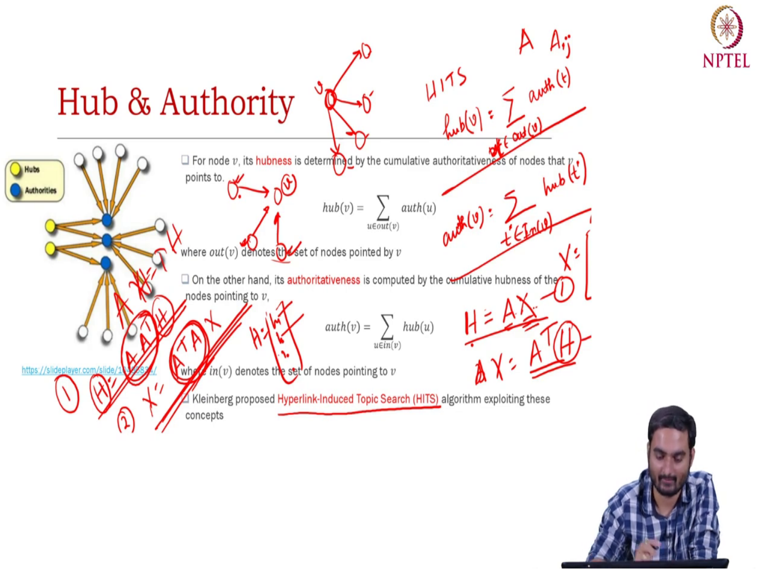
Now, how do we capture and he also said that look these two things should be interrelated. Meaning, that if a node is pointed by a highly hub node right, a node which is which has hub which has high hubsness value, that node would be would may have high authority. Similarly, if a hub node is pointed by if a hub node points to an authority node right, the hub score of this node will also increase ok.

So, what he is saying is as follows. So, the what he is saying is that say let us think of a score called hub, hubness of a node v. This is so, let us think of this right, say this is v and these are the nodes which are pointed by v. So, you look at the neighbors, the outward neighbors, the outward neighbors of v and let us assume that u is one of such, say t is one such outward neighbors. You look at the authority score, authority score of t and then you sum them up.

So, the hub score of v is the sum of authority scores of all the nodes which are pointed by v. Similarly, the authority score of a node v is the sum of say t dash hub of t dash. So, say let us say this is v and right these 3 nodes are pointing to v right. So, you have hub score of this node, you also have hub score of this node and hub score of this node. The authority score of v would be sum of the hub score of all these nodes.

So, if a paper is pointed by a book, if a paper is referred, if a paper is cited by a seminal, if a paper is cited by a survey paper; automatically that papers importance will increase. You see that things are interrelated. How do we combine them? Ok. So, let us say that you have this adjacency matrix A right. So, hub score of v small h, hub score of v would be what? Would be say a right.

(Refer Slide Time: 24:42)



Hub & Authority

- □ For node v, its hubness is determined by the cumulative authoritativeness of nodes that v points to.

$$hub(v) = \sum_{u \in out(v)} auth(u)$$

where $out(v)$ denotes the set of nodes pointed by v

- □ On the other hand, its authoritativeness is computed by the cumulative hubness of the nodes pointing to v,

$$auth(v) = \sum_{u \in in(v)} hub(u)$$

where $in(v)$ denotes the set of nodes pointing to v

- □ Kleinberg proposed Hyperlink-Induced Topic Search (HITS) algorithm exploiting these concepts

https://slideplayer.com/slide/10...

So, if I write it in a matrix form, this would be; this would be like this. So, let us say we have this vector H which is h 1 h 2 h 3. So, h 1 corresponds to the hub node hubness score of node 1, h 2 corresponds to the hubness score of node 2 and so on and so forth. So, it would be A times let us say X is the so, H is the hubness metric, X is the authoritativeness metric right, X 1 is the authority score of node 1, X 2 is the authority score of node 2 and so on and so forth; but, then A is the adjacency matrix.

So, this would be A this would be H equals to A times X and this would be X equals to right A transpose H, why transpose? Because, if A of i j indicates a link from i to j then a of j i will indicate the link from j to i right. So now, we have this two formula. So, if I replace so, it is 1, this is 2. So, in 1 if I replace X by this one, this would be H equals to right A A transpose H and if I replace H in equation 2 by equation 1.

So, equation 2 would be X equals to A transpose A X. Now, what is this? This is the equation of eigenvector eigenvalue right. Now, this is a matrix symmetric, this is also a matrix. So, this is and this is a vector. So, this is A X equals to I mean A H equals to lambda H right. This is also; this is also an equation of eigenvector eigenvalue right.

So, to get the hubness score of nodes, you basically need to look at the eigenvectors of A A transpose. Again, principle eigenvectors right, principle eigenvector; why? Because, of same reason, the one that I discussed in the Katz centrality lecture and if you want to get the authority score, you basically take the you basically derive the principle eigenvector of A transpose A right. So, such a nice equation right. So, I stop here. We will take another lecture to finish the remaining part of this chapter ok.

Thank you.