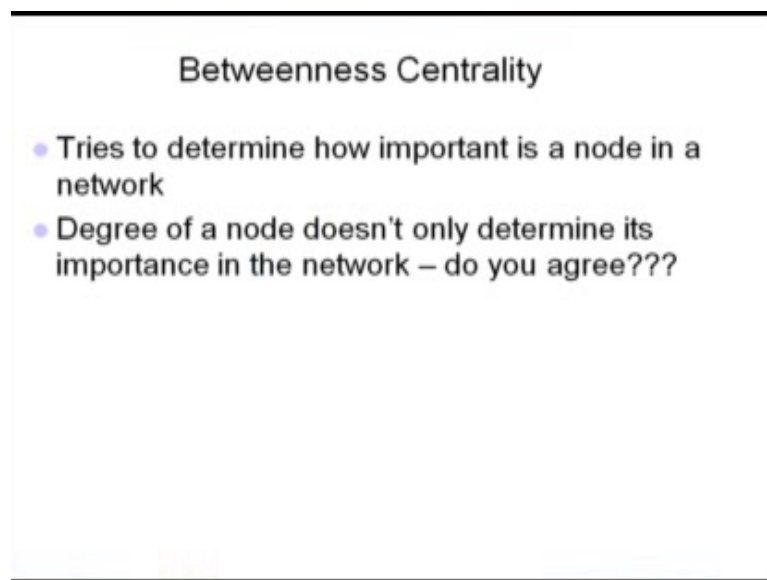


Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 04
Network Analysis – III

Welcome back. Last day we got introduced to our self to this centrality measures. You are trying to understand like, how important a node is or how prominent a node is in a network. And one of the measures that we defined was degree centrality. Now, today I start with a claim that degree centrality is not the only way to indicate the importance of a node in a network.

(Refer Slide Time: 00:49)



Betweenness Centrality

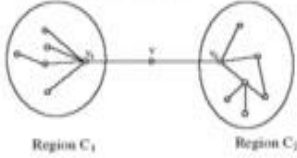
- Tries to determine how important is a node in a network
- Degree of a node doesn't only determine its importance in the network – do you agree???

So, that is what I also write down in the slides is that degree of a node does not you know actually essentially determine all the importance of a node in a network. And I would probably think that you would also tend to agree. One of the striking examples or one of the good examples to substantiate this claim is the following one.

(Refer Slide Time: 01:11)

Betweenness Centrality

- Tries to determine how important is a node
- Degree of a node doesn't only determine its importance in the network – do you agree???
- The node can be on a *bridge* centrally between two regions of the network!!

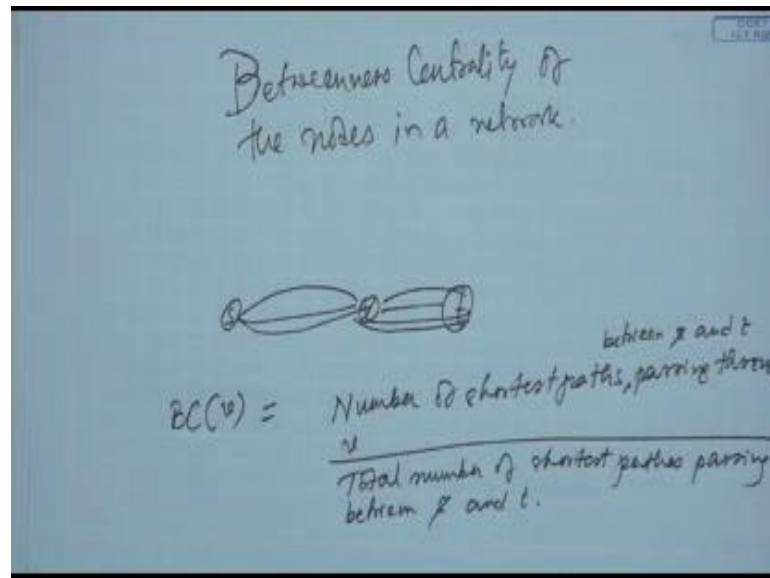


Region C₁ Region C₂

So where you see in the slides that there are these two regions C 1 and C 2 and these two regions are only connected by the vertex v sitting in the middle of the two regions. This is this node is kind of bridging between the two different regions of the networks. And definitely these nodes are very, very important, for instance if you want to pass information from region C 1 to region C 2 you have to only rely on v there is no other way.

So, there are lots of interesting and important repercussions of having nodes like v in a given network. But then the question is that how to quantify the presence of such nodes in a network. And that brings us to what we call the Betweenness Centrality of the nodes in a network.

(Refer Slide Time: 02:12)



Now if you try to question like, how does the node v differ from the other nodes; if you try to make a question like this, so what metric or what property can differentiate between the node v and the other nodes in the network. One of the immediate things that you can think of is probably the shortest paths. The number of shortest paths that are passing between any pair of points through v , if that is very high with respect to the total number of shortest paths that in general pass through the pass between those two points then you can say that v kind of sits in the middle of these two pair these nodes and is acting like a bridge.

For instance, if all the shortest paths between say two nodes s and t , if all the shortest paths get routed through v so then the betweenness of node v in the context of the pair of nodes s and t would be very high. That is what we try to define as betweenness centrality as in the next slide

So basically, Geodesic distance is nothing but the shortest distance or the shortest path length. So centrality or betweenness centrality of the node v is nothing but the number of shortest paths passing through v , number of shortest paths between s and t passing through v expressed as a fraction of total number of shortest paths passing between s and t . Basically, you try to express the number of shortest paths passing through v relative to the total number of shortest paths between s and t . And that you do for all pairs of nodes

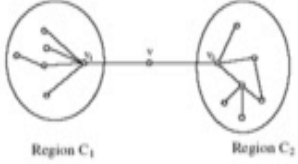
in the network in the context of a particular vertex v . And that is what defines g_v as I show in the slides.

(Refer Slide Time: 05:15)

Betweenness Centrality

- Centrality of v : Geodesic path between s and t via v expressed as a fraction of total number of geodesic paths between s and t

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



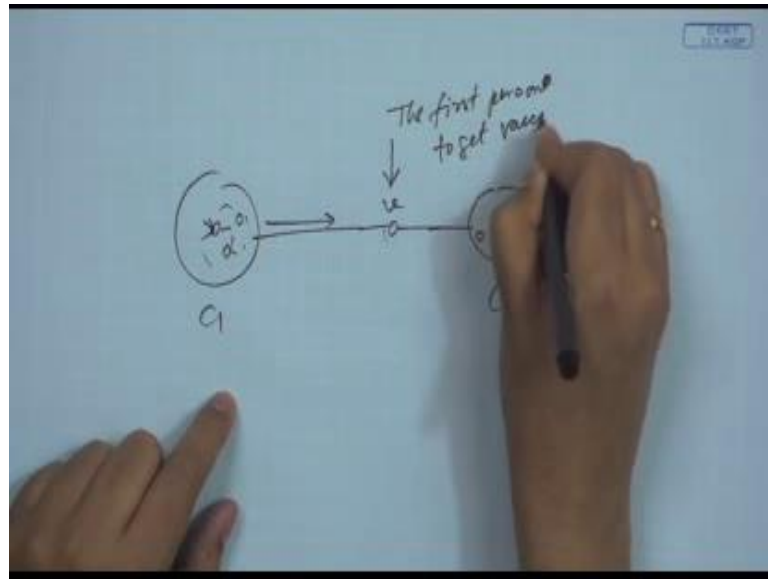
$$g(v) = 2 \sum_{s \in C_1, t \in C_2} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$$= 2 \sum_{s \in C_1, t \in C_2} 1$$

$$= 2N_1N_2$$

So g_v is nothing but the sum over all pairs, s and t where neither s or t is actually v sum of the ratio of $\sigma_{st}(v)$ which is the number of shortest paths that pass through v from s to t divided by the total number of shortest paths between s and t . So now given this formula if you try to compute the betweenness centrality of the node v in the picture on the slide then you get a number as $2n_1n_2$, where n_1 is the total number of nodes in the region C_1 and n_2 is the total number of nodes in region C_2 . That is basically, all shortest paths between all pairs of nodes in this two regions actually pass through v . And there could be very important repercussions of say removing v in the context of particular network.

(Refer Slide Time: 06:37)



For instance, say if you have an epidemic spreading network. Suppose there is an epidemic spreading over this network that I have shown you in the slides. Let us draw it for simplicity again. Suppose you have the two regions C 1 and C 2 with some nodes out here and some nodes out here and all these nodes are connected via this node v. Now suppose there is some disease outbreak here in this region and suppose this person is infected immediately the others in this region will also get infected right, and this infection will now pass on.

And in if the node v gets infected somehow then there is a high chance that all the nodes in the region C 2 will also get infected through v. One of the important questions is like in such networks is whom to vaccinate, and possibly if you have a limited number of vaccines then the first person that one should vaccinate in such an epidemic spreading network is perhaps the first person to get vaccine. If there is a limited supply of vaccine then nodes like v should be vaccinated first.

(Refer Slide Time: 07:56)

Betweenness Centrality

- Removal – what can this lead to??
- Increase in the geodesic path – extreme case is infinity (network gets disconnected)
- Can you visualize the impact of removal of the nodes with high betweenness in the following networks??
 - Epidemic network
 - Information network
 - Traffic network

As I say its removal could have very strong repercussions in various networks like; the epidemic network example we have already taken, in the information networks say for instance that the internet router network that we were looking at in the last days lecture. If one of the nodes like v gets shuts down at some point in time then the entire internet backbone of a particular region might fall through. So it is very important to maintain the functionality, maintain the well being of such nodes which sit in between, otherwise disconnected regions. It is only through this particular node that the two regions get connected. These types of nodes are very, very important.

Similarly in the traffic network, if a particular road or is the only connecting road between say two different geographies and if that there is a blockage in that road then you can immediately understand that the entire traffic flow from one region to the other region will get disturbed. So these kinds of nodes actually are important in various different contexts.

(Refer Slide Time: 09:12)

Flow Betweenness

- What if the nodes with high betweenness behave as reluctant brokers and do not allow two other nodes (of different regions) to establish a relationship.
- They must find other ways to establish relationship (may not be cost effective)
 - Something like “wanting to propose someone via a third party (say his/her friends) who is also (kind of) your friend – but this common friend is reluctant to pursue the proposal!”
- This is the main idea of flow betweenness
- Takes into account all paths (not only the shortest ones) from s to t via v – computationally quite intractable for large networks.

So that is all about the betweenness centrality. Now, the question is that sometimes suppose in case of information spreading that is happening in a particular network and there is a node v , now these nodes can actually have a very strong advantage of keeping the information restricted to itself and not passing it or not forwarding it any further. This is a undue advantage that these types of nodes actually get. If the node wishes then it can keep the information upon to itself and not pass it on any further.

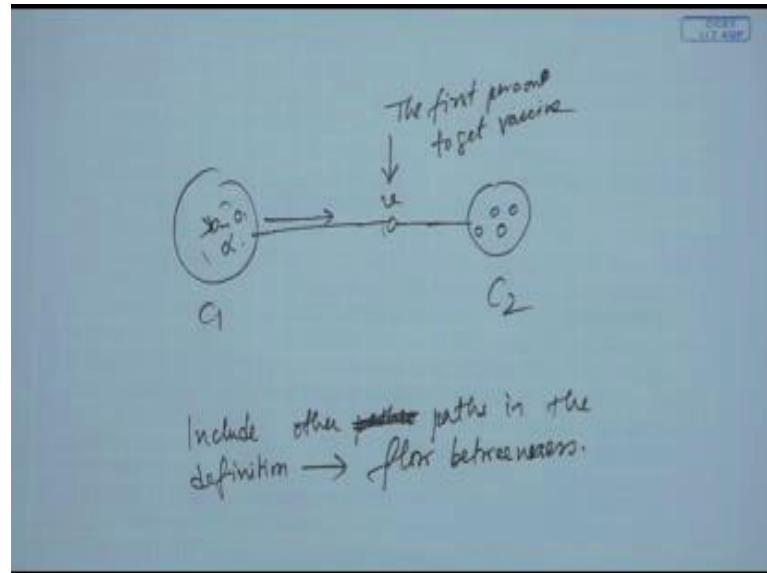
So, in such cases passage of information to the rest of the network is completely blocked. In these contexts such nodes are also often called brokers or social brokers. They keep the information and they actually do not pass it unless and until there is some extra advantage given in terms of say money or something to them, so unless or until that is done actually pass on the information any further.

So then, what is the way out? The way out is to look for not so strong brokers in the network. So the idea is in terms of quantification as I was telling you that for betweenness centrality you need to look for shortest paths. What if we could look at not all so short paths, little longer ones? If we could also include in our definition not only the shortest paths, but if there is any path between two regions through a node v not only shortest paths then we consider it.

Of course, as soon I give you this definition there is a limitation to it. For large graphs it is very difficult to compute paths of any length, this is well understood, this is the hard

problem. So paths of computation of paths of any length are a very difficult task. So, only we can do so in smaller size graphs.

(Refer Slide Time: 11:35)



So that idea where you can include other paths in the definition gives us a new definition which is called Flow Betweenness. In flow betweenness, the major difference between betweenness centrality and flow betweenness is that in case of flow betweenness you can consider not only the shortest path but also any other path, path of any length that passes through v between a pair of nodes s and t . And like a motivating example is what I have given in the slides in green. So this is the simple broker example that I was talking of, but in the probably in a context which is more relevant to you people you can read it up yourself.

(Refer Slide Time: 12:46)

Eigenvector Centrality (Bonacich 1972)

- In context of HIV transmission – A person x with one sex partner is less prone to the disease than a person y with multiple partners

Now the question is like we have so far learned at least two different ways of measuring the importance of a node in a network; one of them was the degree centrality and the other one of was betweenness centrality. Now the question is that all is there any other way in which the prominence or the importance of a node can be defined. The answer to this is, if you ask me the answer to this is yes. Another example that I will put forward in front of you now is the following. Suppose, you have a sexual contact network like the Swedish sex web that I have showed you earlier, so in that context this particular example is very relevant.

(Refer Slide Time: 13:38)

Who among A and B is more prone to an STD?

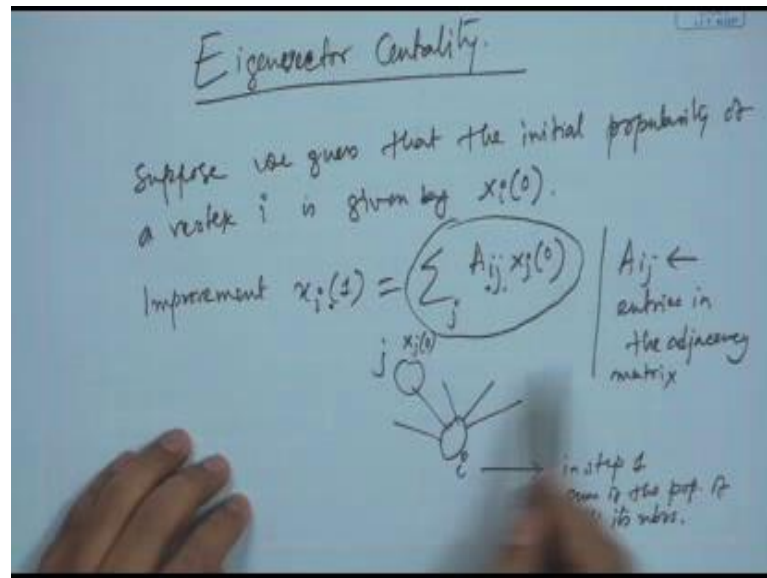
— Recursive Defn of centrality

Say there is one node A and there is another node B out here, and the node A has many sexual contacts whereas, the node B has only one sexual contact C. Now given this piece of information our question is whether from this information we can say which among these two nodes A and B is more prone to say a Sexually Transmitted Disease. Who among A and B is more prone to an STD? Now in order to answer such questions, if you just look into this particular snapshot or this much information from the graph probably that is not enough.

And some of you probably have already guessed why, because it might be the case that although B has one partner C, but C itself has many, many, many, many, many sexual partners. In that case the chances that B gets infected by a STD are way higher than an itself. So, many means I really mean many, many. The idea therefore, is that your propensity to get infected depends on your neighbors' propensity to get infected and that in turn depends on their neighbors' propensity to get infected and so on and so forth. So this is a Recursive Definition of centrality.

In other words you can also think of it in terms of prominence or importance. So your popularity or prominence is actually determined by the popularity of your neighbors or of your friends, their popularity is determined by their friends and so on and so forth. So your popularity is actually a share from your friends, your friends of friends, your friendses friendses friends and so on and so forth. So it is a recursive definition. And the best way to encode this recursive definition is in the form of what we will now study the Eigenvectors.

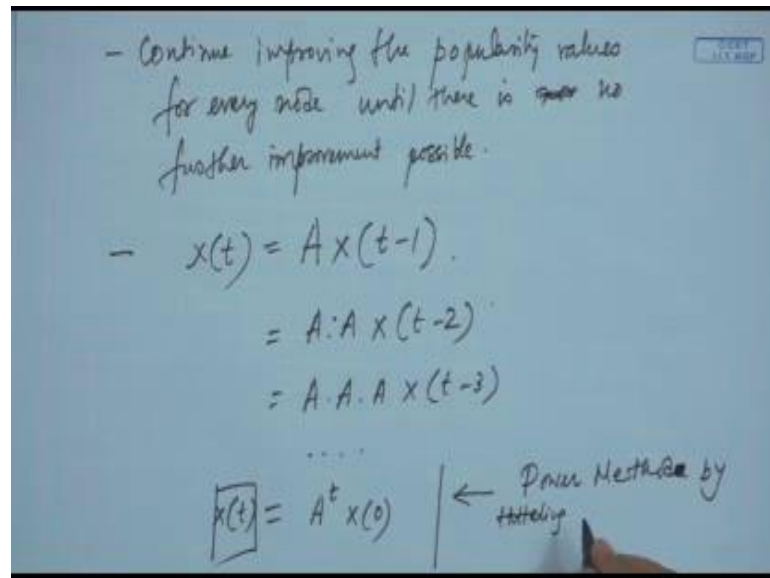
(Refer Slide Time: 16:36)



Eigenvector Centrality, as the name suggests this will borrow concepts from your school days where you have been introduced to the idea of eigenvectors and simple linear algebra. So the idea is very simple. You have to express the popularity of yourself in terms of the popularity of your neighbors and so on and so forth. Suppose you guess, so we make a guess that the initial popularity of a vertex or a node i is given by say $x_i(0)$. Now in every step we look for improvements of x_i . Now the improvement $x_i(1)$, that is improvement at step 1 can be written as $\sum_j A_{ij} x_j(0)$. That means, suppose you have a node i out here and it is connected to some j other nodes and j 's popularity is say $x_j(0)$, then the popularity of i in step 1 is the sum of the popularity of all its neighbors. Here A_{ij} is nothing but the entries in the adjacency matrix which holds the neighborhood information.

So, if i is a neighbor of j then in the adjacency matrix there will be an entry 1, otherwise the entry will be 0. This is a 0, 1 matrix, a binary matrix which actually expresses the neighborhood relationship of i . So now, as you can see from this expression out here what you do is you look at all neighbors of i , that is why A_{ij} if j is a neighbor of i then the popularity of j counts into the new improved popularity value of the node i . And in this way you keep on doing this.

(Refer Slide Time: 19:43)



So, you continue improving the popularity values, so this you do for every node popularity values. Continue improving the popularity values for every node until there is no further improvement possible. This can be very nicely encoded in a functional form. So, x_t is x_t which is kind of a vector of all the popularity values of all the nodes in the network is nothing, but the adjacency matrix times the popularity value at t minus 1, and this continues. This you can write as A multiplied by $A x_{t-2}$ this can be written as A into A into $A x_{t-3}$ and so on and so forth.

Finally, x_t can be expressed as a power t , so the adjacency matrix is multiplied with itself t times a power $t x_0$. So you start with any arbitrary x_0 values some initial guess the final t values will be the point at which there is no further improvement possible you will have a situation like this, where this t is like very large so after that point there is no further improvement. The more and more powers you raise A to there is no further changes in the matrix.

And finally, you get the resultant x_t values for each node. This is the score or the popularity score of each of the nodes. This x_t contains finally the popularity score of each of the nodes. So this is actually referred to as the Power Method by Hotelling. This is what I have written in this slide.

(Refer Slide Time: 22:30)

Eigenvector Centrality

- Idea is to define centrality of vertex as sum of centralities of neighbors.
- Suppose we guess initially vertex i has centrality $x_i(0)$
- Improvement is $x_i(1) = \sum_j A_{ij} x_j(0)$
- Continue until there is no more improvement observed
- So, $x(t) = \mathbf{A}x(t-1) \Rightarrow x(t) = \mathbf{A}^t x(0)$ [Power iteration method proposed by Hotelling]

So this is called the power method by Hotelling, and actually this is the method which is usually used to compute the eigenvectors of a given matrix.

Now, the question is like we have so far been able to give a nice expression to this x t 's which is nothing but the popularity scores of each of the individual nodes after some say sufficient time steps t after which there is no improvement in the scores. Now does this value get related or in some way is it related to the eigenvectors, that is what we will try to see next. In order to do so we will start with a very simplistic assumption.

(Refer Slide Time: 23:40)

Let $x(0)$ be a linear combination of ^{some of} the eigenvectors of the matrix A

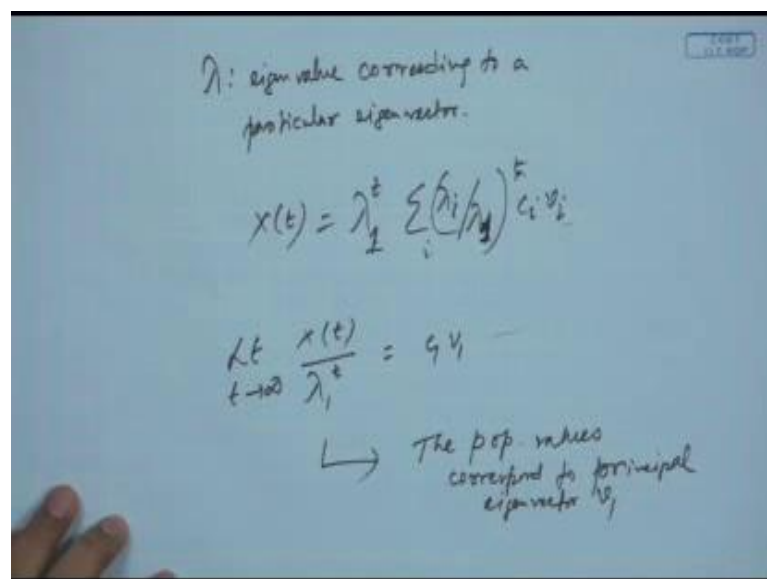
$$x(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_n(0) \end{bmatrix} = c_1 \begin{bmatrix} v_1^1 \\ v_1^2 \\ v_1^3 \\ \vdots \\ v_1^n \end{bmatrix} + c_2 \begin{bmatrix} v_2^1 \\ v_2^2 \\ v_2^3 \\ \vdots \\ v_2^n \end{bmatrix}$$
$$x(0) = \sum_i c_i v_i$$
$$\Rightarrow x(t) = A^t \sum_i c_i v_i = \sum_i \lambda_i^t c_i v_i$$

$\left\{ \begin{array}{l} Av = \lambda v \\ A^2 v = \lambda^2 v \\ \dots \end{array} \right.$

We will say let our x_0 values that is the values of the popularity the initial guessed values of the popularity that you start off with for all the individual nodes. So, at time t equals 0 you have an initial guess for each of the individual nodes. So, let that x_t be a linear combination some of the eigenvectors of the matrix A , that is the adjacency matrix. Let us express x_0 , so let us choose x_0 in such a way that it is actually x_0 that is a vector like x_1^0, x_2^0 until x_n^0 if there are n nodes in the system is equal to a linear combination of say some constant C_1 into one of the eigenvectors say v_1 or say $v_{i_1}, v_{i_2}, v_{i_3}, v_{i_4}$ and so on and so forth up to v_{i_n} . So this is the n th eigenvector. So, this class C_2 sums v_{j_1}, v_{j_2} and so on and so forth.

So, if you express this x_0 as a linear combination of this eigenvectors then you can immediately write x_0 equal to sum of $c_i v_i$ for all i 's. From this since you have already have an expression for x_t you can write x_t as $A^t x_0$, we saw that x_t is equal to nothing but $A^t x_0$. So, we can write that x_t is equal to $A^t \sum c_i v_i$. This can be written as $\sum \lambda_i^t c_i v_i$, where you know from the eigenvector equation that $A v_i$ is equal to $\lambda_i v_i$ or $A v_i = \lambda_i v_i$. where you know that $A v_i$ is equal to $\lambda_i v_i$. That means, $A^2 v_i$ is equal to $\lambda_i^2 v_i$ and so on and so forth. From there you get this particular equation. Here, in this case λ_i is the eigen value of the, let us write later on.

(Refer Slide Time: 27:07)



So, λ_1 is the eigen value corresponding to a particular eigenvector. From there you can write $x(t)$, if you see some simplification will tell you that you can bring out the principle eigen value outside and you can write the rest as λ_i by λ_1 power t for all $i < n$. Now, in the limit of t to infinity $x(t)$ by λ_1^t is nothing but $c_1 v_1$, because everything else all other values goes to 0.

So, t tends to infinity means this λ_1 is larger than all the other λ values so this fraction is always lower than 1. so this power always tends to 0, so everything becomes 0 except for the first one except for the case where i is equal to exactly 1. When i is equal to 1 this is λ_1 by λ_1 that becomes 1 power t that is $1 c_1 v_1$ plus the rest x everything goes out due to in the limit t tends to infinity.

Basically, what you have is that the popularity values correspond to the principle eigenvector v_1 . If you just calculate the principle eigenvector of the adjacency matrix you immediately get the popularity scores that are why this is called the Eigenvector Centrality.

Thank you.