

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

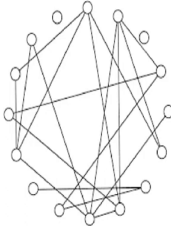
Chapter - 03

Lecture - 02

So, as we have been discussing about you know network growth models and the last lecture we have understood why a synthetic network or a network growth model is important right. So, now let us start you know discussing a simple model for network generation is called random graph model or random network model.


(Refer Slide Time: 00:44)


Synthetic Networks: Random ^N Network Model



An instance of $G(16, \frac{1}{2})$ network

- Also popularly known as **Erdős and Rényi model** (or ER model)
- A number of variants of the model
- Popular variants:
 - $G(N, K)$ model [Erdős and Rényi 1959]: From the set of all networks of N nodes and K edges, a network is chosen uniformly at random
 - $G(N, p)$ model [Gilbert 1959]: Network has N nodes, and any random pair of nodes has a probability p of being adjacent independently with any other pair of nodes in the network
- Both the variants behave identically in the limiting case
- $G(N, p)$ model considered as the standard random network model





So, this model was proposed by Erdos and Renyi Paul erdos and Albert Alfred Renyi in long time back actually and it was one of the simplest models that we can think of for generating a random network ok. This is called ER this is also called ER model your ER growth model stands for Erdos and Renyi and it basically has two realizations ok.

So, let us first try to understand how this model works. It basically says that let us assume that there are N number of nodes ok. There are N number of nodes present in a network, we want to create a network with N number of nodes. But, we do not know how the edges are formed right. So, what we will do? We will choose a pair of nodes and then we connect it based on certain probability right.

So, say for example, we toss a coin if the outcome is head, we connect the pair of nodes otherwise we do not connect ok. So, if you see there is a very simple idea see very simple idea of you know connecting a pair of connecting pairs of nodes, but it turned out to be effective you know at least in satisfying some of the properties that real world network exhibits ok.

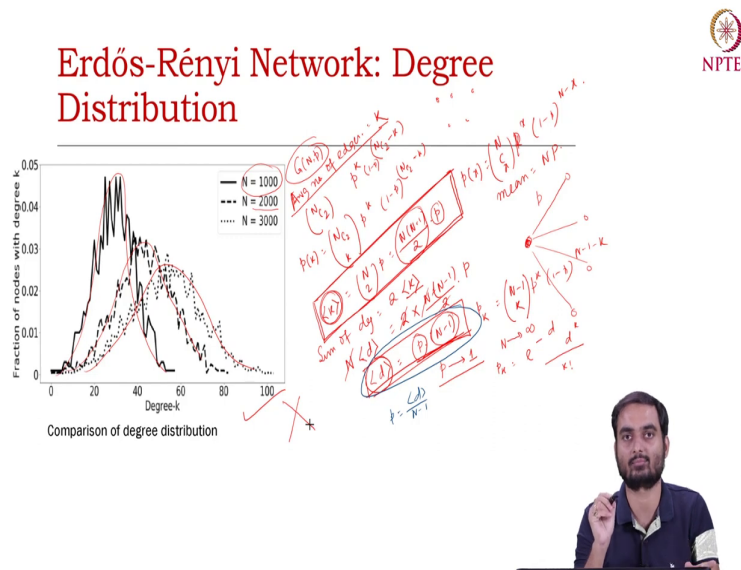
So, so far we have discussed three properties. Let me remind you one is the, I mean it should have an average high average local clustering coefficient. The second one is it should it should follow the small world property meaning that the average path length should be proportional to logarithm of \log of N , where N is the number of nodes and the third one is a scale p property which basically says that the degree distribution should follow a power law property ok.

So, there are two ways to actually explain the random graph model. One is called the $G(N, K)$ model $G(N, K)$ model $G(N, K)$ model, where N is the number of nodes and K is the number of edges that you want to create ok and this is basically ER model that Erdos and Renyi actually proposed it. Basically says that you know there are N number of nodes you and you want to create K number of edges.

And how do you create? You just choose a pair of nodes you connect it and you keep on doing this thing for K number of times, there is no probability as such. In this particular case the number of edges will always remain same right. The second realization is basically $G(N, p)$ where p is the probability, probability of connecting a pair of nodes.

It basically says that you choose a pair of nodes and then with probability p you connect them with probability $1 - p$ you do not connect them ok. This this was proposed by Gilbert in 1959. So, both these models I mean in the asymptotic level in the limiting case both the models behave similarly, but in this particular you know chapter we consider $G(N, p)$ ok as the standard way to explain random graph models ok.

(Refer Slide Time: 03:58)



So, let us do some math ok. Let us first try to understand what would be the if we have $G(N, p)$ model right. What would be the expected number of edges expected or say average number of edges right that can be formed in the network G ok. So, let us assume that the average number of edges e is k ok. It means that there are you know $N \times (N-1)$ choices.

There are N number of nodes and there are $N \times (N-1)$ pairs possible pairs and out of this $N \times (N-1)$ pairs only k pairs have been successful in terms of connections ok. So, each such each of the k edges which are formed the probability is p . So, for k number of such edges it should be p to the power k and for you know $N \times (N-1) - k$ possible pairs right the probability would be $1 - p$, because $N \times (N-1) - k$. This number of pairs have not been connected right and the probability is $1 - p$ right.

Now, if you think of the you know the probability right. This basically follows the binomial distribution as you see right. So, it is basically p of you know k , k is N you know $N \times (N-1)$ choose k right p to the power k $1 - p$ to the power $N \times (N-1) - k$ ok. So, this is the probability of k number of edges in a random network ok. So, what is the expected number so this since this is exactly following the binomial distribution which is p of x right. We know in $\sum_{k=0}^N k \binom{N}{k} p^k (1-p)^{N-k} = Np$ right and what is the mean? Mean is always Np right.

So, in our case in our case if we take the mean number of edges or average number of edges of a random graph model it should be it should be $N \times (N-1) \times p$ which is N into N minus

$\frac{1}{2} \times p$ ok. So, this is the average number of edges that you can think of in this particular model $G(N, p)$ ok. So, if this is the average number of edges let us think of the degree ok. So, what is the degree? So, we know that sum of degree sum of degree is $2 \times$ number of edges, here $2 \times$ average number of edges right.

So, let us assume that the average degree is d right. So, if the average degree is d what is the sum of all degree? It is $N \times d$. There are N number of nodes and each of these nodes has on an average d degree right. So, this is $2 \times$ this one $2 \times$ this one I replace this by this.

So, this should be $N \times \frac{1}{2} \times p$ ok. This and this will cancel out this and this will cancel out. So, we will have average degree is essentially $p \times (N - 1)$ ok and this is also reasonable right. You can I mean in other ways you can also understand why it looks like this think of a node right.

So, a node can have a node can have $(N - 1)$ number of possible nodes with which it can connect right and all this I mean each of these pairs right, each of these $(N - 1)$ pairs there is a probability p of connection right. So, actual number of pairs which are going to be actual number of nodes which are going to be connected with the particular node would be this one ok. So, we have seen two things one is that. So, this is the and you please try to remember these 2 equations right because we will you will keep using these equations later on also.

So, this is the number of edges average number of edges and this is the average degree right. So, what actually is what other things you we can infer. So, as p tends to 1 I mean p goes towards 1 right p is a probability right. What would happen? So, if this is 1 then the number of edges would be $\frac{N \times (N - 1)}{2}$, which is basically a click. So, p equals to 1 means what? p equals to 1 means we take every we take a pair of edges and we connect.

Because the probability of connecting a pair of nodes is always 1 . So, we will always connect. We will it is not the case that, we will not encounter any case that the pair of nodes will not be connected ok. So, as p increases this will increase right; we will also see that as p increases the average degree also tends to be $(N - 1)$, because it is ultimately it is a click. So, every node will be connected to other $(N - 1)$ other nodes ok.

So, these two observations are important to understand. Now if you look at the degree distribution right. So, in case of degree distribution also we will follow the same thing right.

So, we basically say that look we have you know $N - 1$ number of for every node for every node we have $N - 1$ number of possible cases right and let us say the degree is k . The degree of every, I mean we basically want to understand p^k right. Probability of a node having degree k is essentially $\binom{N-1}{k} p^k (1-p)^{N-1-k}$ ok.

So, this is essentially binomial distribution right and we also know that as N tends to infinity what happens is that p^k starts behaving starts behaving as a Poisson distribution right. So, $e^{-p} \frac{p^k}{k!}$ to the power minus say d right d to the power k by factorial of k ok. So, essentially as you see here in this particular figure right. This is the degree distribution of a random graph model ok with N , N equals to 1000 we have this kind of graph. When N equals to 2000 we have you know little bit spreaded graph and when N equals to 3000 we have this kind of graph right.

(Refer Slide Time: 12:17)

Erdős-Rényi Network: Emergence of Giant Component



Theorem: A giant component emerges in a random network when the average degree of the network is greater than or equal to unity, i.e., $\langle k \rangle \geq 1$.

- For emergence of a giant component, only one link per node on-an-average is sufficient!! The above condition necessary, too
- The emergence is not a smooth, gradual process; it follows a **second-order phase transition**
- Regimes of evolution:
 - **Subcritical Regime ($0 < \langle k \rangle < 1$)** \Rightarrow A number of small isolated clusters in the network, as the number of links is much less than the number of nodes
 - **Critical point ($\langle k \rangle = 1$)** \Rightarrow A distinguishable giant component emerges
 - **Supercritical Regime ($\langle k \rangle > 1$)** \Rightarrow A growing giant component, and less and less smaller isolated clusters and nodes
 - **Connected Regime ($\langle k \rangle > \ln N$)** \Rightarrow The giant component absorbs all nodes and components, the network becomes connected



So, now the very interesting point that we discuss ok and this is basically called the you know the relationship of this random graph model with a giant component right. So, what is giant component? Now giant component is something that we discussed in the last class I guess or with the first lecture. We see giant component is a component if you think of a graph as a disconnected right graph right, there are multiple components. A giant component is a component which actually accommodates 90 percent or 95 percent of the nodes right and other components actually have a small number of nodes ok.

So, when you know when you have N number of nodes right and you basically choose a pair of nodes and connect you keep connecting right. When we will see that a giant component emerges ok? In other words so since there is no restriction on the number of edges. We have the we have only one parameter which is p right.

Since there is no restriction on the number of edges right, can we come up with some relation either in terms of degree or number of edges right based on which we shall say that ok after this many edges or after this many average degree we will see a giant component. Because initially all the nodes are disconnected. They are the N nodes they are disconnected, then you choose a pair of nodes and with probability p you connect otherwise you do not connect.

So, you basically start creating components right. So, when would be the case that a giant component will emerge? We will see that majority of the nodes will basically form a giant component right. So, we will try to come up with a theoretical solution for this right and it interestingly we will see that in order to you know in order to create a giant component or form a giant component the average degree of a node should be 1.

What I am trying to say is that if the average degree of a node is 1 you can see a giant component emerging which is very you know non intuitive. Because I mean degree 1 is a very simple thing I mean you can always say degree 1 right, but remember this on an average right. Some nodes may have higher degree some nodes may have lower degree, but the average degree should be 1 and we will prove that you know when a giant component emerges the average degree actually needs to be 1 ok. So, let us do it you know mathematically.

So, let us you know calculate the probability that a random node a random node v is a part of the giant component. What is the probability? So, straightforward this is the size of the giant component divided by the total number of nodes present in the graph right. This is the probability. Let us define and say this is p_v belongs to G_c ok.

Let us define another notation. Let us define m which is $1 - G_c / N$ ok. It basically says that what is the probability that a node is not a part of giant component, it is basically 1 minus the first equation ok this is m . So, then if we know these two things then let us try to understand the probabilities of these 2 scenarios. So, what is the probability that the first scenario will occur. The first scenario is saying that v_i, v_j are not connected right.

What is the probability that v_i, v_j are not connected? This is $1 - p$. Because p is a probability because this is a random graph model, we have N comma p , p is the probability that nodes are connected a pair of nodes are connected. So, $1 - p$ is this one. What is the probability of the second scenario?

Which basically says that the edge exist, but nodes are not part of the I mean the edge exist, but v_j is not a part of the giant component. So, the edge exist the probability is p , but v_j is not a part of the giant component the probability is m . So, this is p times m ok. So, these are the two probabilities.

So, now we can combine. So, combinedly we define that what is the probability what is the total probability that the scenario 1 and 2 two will happen right? p of s_1 plus p of s_2 . This is essentially what? $1 - p$ plus right; what does it mean? It basically means that for a particular node this is the total probability that it will not be a part of the giant component ok for a particular node. How many such nodes are there? Because we fixed a we have already fixed a node right, we have already fixed v_i . So, how many other nodes are there?

There are $N - 1$ number of other nodes. So, what is the you know what is the total probability that $N - 1$ number of nodes? You know are not a part of the giant component this is basically $1 - p$ plus $p m$ to the power $N - 1$ ok. So, it basically says that that the giant component will never emerge if this happens right. So, none of the nodes are part of the giant component, this is the probability ok. Can we say that this is exactly same as m ? Because m also says that that the probability that a node does not belong to the giant component this one is m , so this is m exactly ok.

So, now let us do some tweaks right. Let us take the log in both the sides. So, we will have a log of m , we will basically have \log of m which is $N - 1 \log(1 - p) + p m$ ok and we already have seen earlier if you remember in our last discussion we have already seen that this one right that average degree d is $p(N - 1)$. So, p is what p is d by $N - 1$ ok. I will use this one here. So, I know that p is d by $N - 1$ ok. So, let us replace it here.

So, we will have $N - 1 \log(1 - p)$ right let us take this common $\log(1 - p)$ right $1 - m$ ok. So, this is basically $\ln(1 - x)$ right. So, when $\ln(1 - x)$ so we can approximate it as $-x$ when x tends to 0 right, if you remember the log curve right. So, as x tends to 0. So, we have I mean $\ln(x)$ is basically same as $-x$ right. So, I mean these are all these tricks are all I mean these kind of tricks are needed to simplify these complicated things right. So, we will have so $\ln m$. So, this is our x ok including the negative symbol.

So, we will have what? We will have this $N - 1$ and this $N - 1$ will vanish. We will have this one ok this one. Now let us assume another notation h ok. What is h ? h is the fraction of nodes in the giant component right essentially h is G_c / N right. This is also same as the probability of a node belonging to a giant component. I mean this is same as h equals to this is same as $1 - m$, m is the probability of not belonging to the giant component. So, this is same as $1 - m$ ok.

So, now let us come back here. So, when we move \log here. So, this will be exponential this will be m equals to exponential of right $-(N - 1)m$ ok. So, so what we will have we will basically have right. So, we will have now you basically subtract both a left side and right side from 1. So, we will have $1 - m$ equals to $1 - \exp(-(N - 1)m)$. So, what is $1 - m$ this is h . So, h equals to $1 - \exp(-d h)$, this is also h .

So, this is our equation h equals to. So, what is our equation? h equals to $1 - \exp(-d h)$ of $-(N - 1)m$ times h . Now we need to come up with the value of h right. So, I mean how do we solve this equation? So, there is no closed form solution ok. Left hand side you see h right hand side you also see h . So, how do we solve this kind of equation? So, we basically solve this equation using graphical format ok.

So, let us assume that this is y ok y equals to h equals to this one. So, here you basically have 2 things. One is y equals to h other is y equals to $1 - \exp(-d h)$ ok and this is your h and this is y . So, y equals to h is this one ok. And how do you get the y value

from here? So, this curve is already plotted. What about this curve? So, here you see there are 3 unknowns h , $h y$ and $\text{mod } d$.

So, let us fix a value say $\text{mod } d$ say let us say $\text{mod } d$ is average degree is say 0.5. If average degree is 0.5 then we vary h and we will get y . So, we will vary h and we get y and if you do that you will see that it basically looks like this. Here this is $\text{mod } d$ 0.5 ok. We will see a curve like this when $\text{mod } d$ equals to 1 right and when $\text{mod } d$ goes to 1.5 we will see like this.

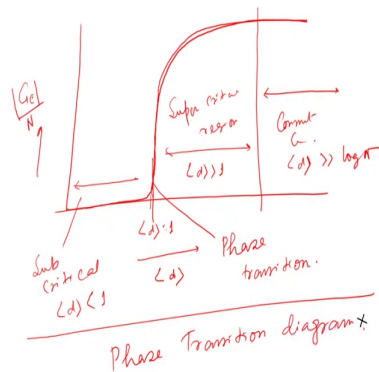
So, this line the straight line is y equals to h and the this other curves are this one. So, what are the solutions? The solutions are those points where these curves actually intersect. So, this is one point and this is another point. And how do you know the points? So, we know that this this kind of cases we take the gradient right at this point and then we at this point the gradients are same ok. And how do we get the gradient?

So, we basically take the derivative. So, this is our equation ok. So, I take the derivative with respect to h , but I will get I will get 1 here and I will get exponential of right minus right. This is the this is the gradient, this this is the derivative at I mean with respect to h right and so since we know that at this point we have a solution. So, we let us put h equals to 0, because at this point h equals to 0.

So, if I put h equals to 0 here, what we will have? We will have $\text{mod } d$ equals to 1 right. So, it basically says that in order to form. So, so this is what? This is average degree average degree is 1. So, it basically says that in order to form or in order to in order for a giant components to emerge right, you basically need you basically need the average degree as 1 ok.

And if you think of it carefully right average degree is 1 means what? It means that nodes are connected to each other right. Say something like this right average degree is 1 and if we have a line graph like this you will also have a giant component ok. So, the theorem is that a giant component emerges in a random network when the average degree of a network is greater than or equal to 1 ok. In fact, you can actually I mean based on these properties right you can actually think of different phases of a graph right.

(Refer Slide Time: 30:56)



If you think of it you can actually plot a graph like this, where this is your degree right and this is the size of the giant component or the fraction right. You will see that it to actually look like this. At mod d equals to 1, you see the giant component emerging ok and this is called phase transition ok.

So, this part of the region is called sub critical region, sub critical region where mod d is less than 1. At d equals to one is called phase transition giant component starts emerging, then this part is called supercritical region super critical region where mod d is greater than 1 and this part is called connected component when we will see that mod d is greater than greater than log of n ok. And this kind of diagram is called phase transition diagram phase transition diagram ok.

(Refer Slide Time: 32:32)

Erdős-Rényi Network: Average Path Length



A depiction of random network in tree format

- When l_{max} represent the maximum path length of $G(N, p)$,

$$1 + (d) + (d)^2 + (d)^3 + \dots + (d)^{l_{max}} = N$$
- When $(d) \gg 1$, the above yields,

$$l_{max} \approx \frac{\log N}{\log(d)}$$
- Further approximation yields,

$$(l) \propto \log N$$

Theorem: Erdős-Rényi Networks follow small-world property.



Let us look at the average path length because average path length has relations with the small world property ok. So, and this is very simple. Now so let us start with a node v ok and what is the degree? Degree is mod d average degree is mod d. It means that in the first hop you will have mod d number of nodes right and let us assume that the maximum path length meaning the diameter is d max ok.

So, at the first hop we have d square d at the second hop again at if every node has mod d number of nodes. So, we have d square number of nodes here. Then the next hop d q number of nodes here. Last in the last hop we have d to the power the I mean the l max whatever. So, this if you think of the path length is l max the diameter is l max. So, the last layer the last hop we have mod d to the power l max number of nodes ok. So, and that constitute the all the nodes.

So, we have node I mean 1 node in the zeroth hop. The first hop we have mod d number of nodes, in the second hop we have d mod d square number of nodes dot dot dot; last hop we have mod d to the power l max number of nodes and that is the total number of nodes N ok. So, this is basically a geometric series and this is right. So, the sum ok and you basically I mean you ignore this part right minus at N tends to infinity 1 d when d becomes so large then this does not matter.

Then we will cancel this and this one. So, we will have mod d to the power l max equals to N. You take log in both the sides right you will get this equation ok. And what does it mean? It

says that it basically says that the maximum distance meaning the diameter is proportional to log of N right, which is exactly small world property right. It is proportional to log of N; so, it if the maximum one is this one, you can also say that approximately average distance between any pairs should also be log of N proportional to log of N.

So, it means that the random walk the random graph model actually follows the small world property right. So, it follows the small world property. So, one property satisfied, but it does not follow the scale free property why? Because we have already seen earlier that it basically follows the Poisson distribution right or the binomial distribution which is not a power law distribution. So, it does not follow the power law property. So, one properties kick ticked other properties crossed right.

(Refer Slide Time: 35:49)

Erdős-Rényi Network: Clustering Coefficient



□ In $G(N, p)$

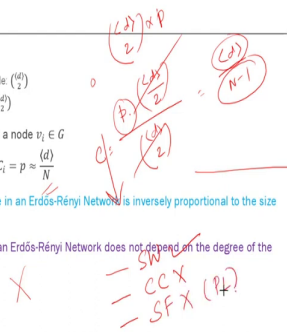
- Number of possible edges between neighbours of a node: $\binom{d}{2}$
- Expected number of edges between these nodes: $p \times \binom{d}{2}$

□ The above yields, the local clustering coefficient for a node $v_i \in G$

$$C_i = p \approx \frac{d}{N}$$

Theorem: The local clustering coefficient for any node in an Erdős-Rényi Network is inversely proportional to the size of the network

Note: The local clustering coefficient for any node in an Erdős-Rényi Network does not depend on the degree of the node



Now, let us look at the clustering coefficient the last property ok. So, what is the local clustering coefficient of a node? Ok now think about it carefully. So, it basically says that lets look at a node and there are mod d number of neighbours because the degree is mod d. So, what are the what are the possible number of connections between mod d number of neighbours this is mod d c 2 right.

And what is the what is the actual number of edges? It is should be d times p, because the same logic we have mod dc 2 number of pairs and then you connect these pairs with probability p right. So, what is the clustering coefficient then? C clustering coefficient is

possible number of edges which is $\frac{d(d-1)}{2}N$ and actual number of edges which is $\frac{d}{2}N$. So, this and this will cancel out.

So, the clustering coefficient is basically $\frac{d}{2(d-1)}$ and what is d ? We have earlier shown that d is $\frac{2E}{N}$ we have earlier shown that d is $\frac{2E}{N}$ by N minus 1 right. So, this is $\frac{d}{2(d-1)}$ or $\frac{d}{2N}$ whatever in the limiting case. So, it means that as N grows if we fix d if N grows this $\frac{d}{2N}$ decreasing the average clustering coefficient. If N is 1 million for example, the local clustering coefficient would be extremely low right.

Now this is opposite to the real world observation. In real world observation we showed earlier that for a gigantic network also the clustering coefficient is pretty high 0.14, 0.15 and so on. But it basically says, but in this case it says that with the increase of N c will decrease right. It means that the ER model does not satisfy the clustering coefficient property right.

So, it satisfies the small world property, it does not satisfy the clustering coefficient property, it does not satisfy the scale free property, which is the power law property. So, we stop here. In the next lecture we will discuss another model. We will see that how we can actually satisfy other two properties which have not been satisfied in this particular model ok.

Thank you.