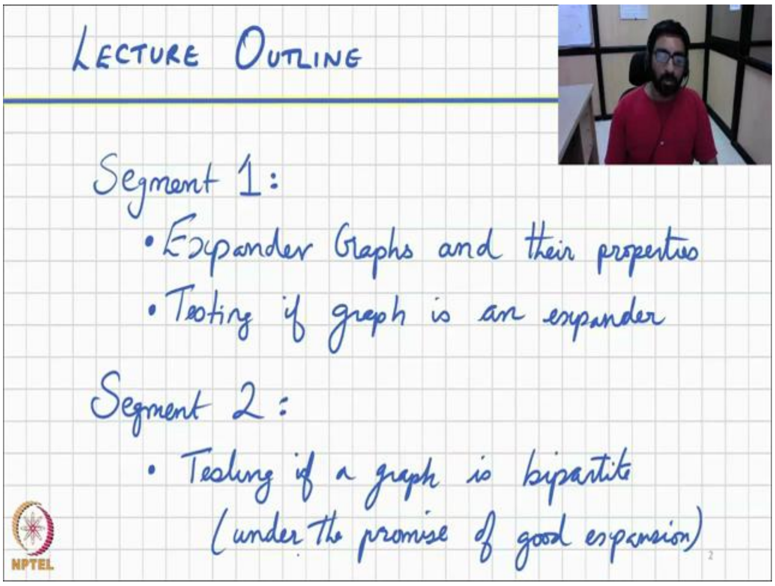
Algorithms for Big Data

Prof. John Ebenezer Augustine
Department of Computer science and Engineering
Indian Institute of Technology, Madras

Lecture - 42
Property Testing: Random Walks Algorithms

Hello everybody. In today's lecture, we are going to talk about property testing algorithms that use random walks. And many of the concepts that we are going to look at today will relate to the notions of Markov chain, random walks, stationary distribution and things to that nature we have looked at in the past.

(Refer Slide Time: 00:38)



LECTURE OUTLINE

Segment 1:

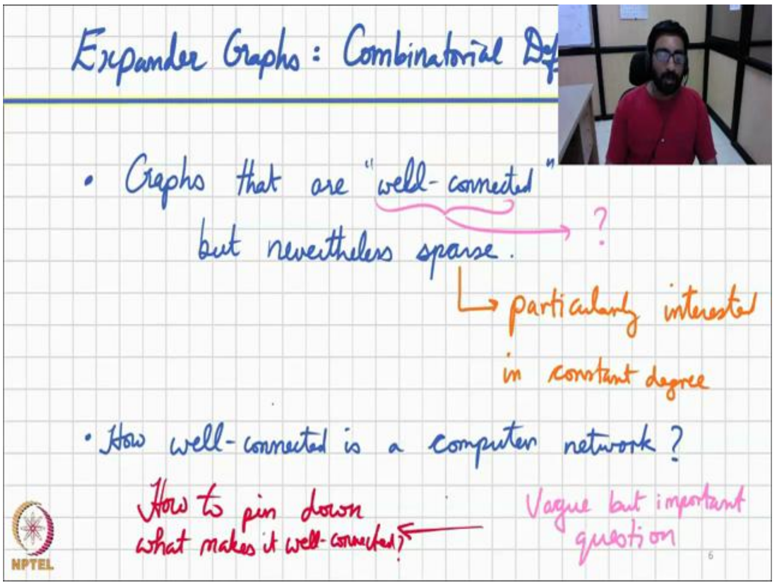
- Expander Graphs and their properties
- Testing if graph is an expander

Segment 2:

- Testing if a graph is bipartite
  (under the promise of good expansion)

Today's lecture outline will be the following. There will be two segments. In segment 1, we will be considering expander graphs and their properties and then we will briefly talk about property testing algorithms to test whether a given graph is an expander or not. And in segment 2, we will be discussing property testing algorithm to test whether a graph is a bipartite graph or not under the promise of reason be of good expansion.

(Refer Slide Time: 01:13)



Expander Graphs : Combinatorial Def

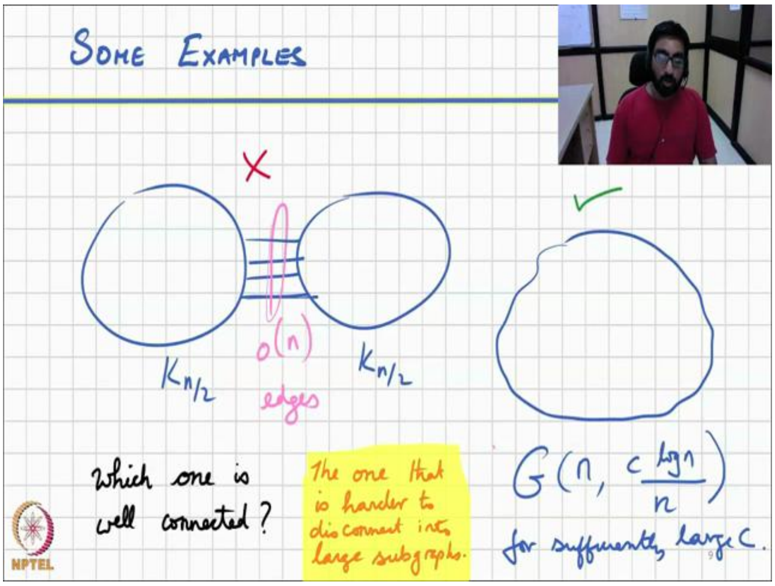
- Graphs that are "well-connected" → ?
  but nevertheless sparse.
  ↳ particularly interested in constant degree
- How well-connected is a computer network?
  How to pin down what makes it well-connected? ← Vague but important question

6

So, that we will do let us get it to segment 1. And for this, we first need to understand what it means for graphs (Refer Time: 01:23) expander. In general, expander graphs that are well connected, but nevertheless a sparse. So, this little bit of a attention between these two notions. If a graph is well connected you expect the graph to have lots of edges; at the same time, we are also looking for sparse graphs. So, this extension as it turns out can be actually resolve quite well so we have what we call a expander graphs that are really well connected, but are in nevertheless a very sparse, so these are the very interesting combinatorial objects, but there is this many interesting aspects to this notion of expander graphs.

In particular, one of the aspects is that it is not just a combinatorial object, but it also can be viewed as an algebraic notion could be viewed from probability theory and so on. So, it is very interesting notion that we just many interest diverse areas to study. So, at this point our understanding of well connectedness is somewhat vague. So one could ask try to understand this notion from the point of view computer networks, and so what is this notion of well connectedness, so what exactly how do we pin it down, how do we pin down the definition of well connectedness.

(Refer Slide Time: 03:03)



Some Examples

Which one is well connected?

The one that is harder to disconnect into large subgraphs.

$G(n, c\frac{\log n}{n})$

for sufficiently large c.

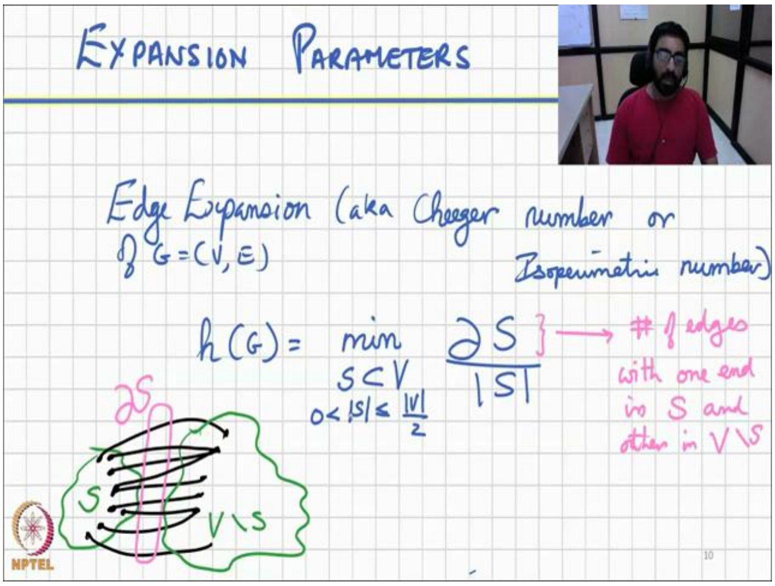
So, look at the couple of examples. So, here two examples one is a graph whether a two the completely connected sub graphs and joint together by some number of edges. Say let say some little o of n edges or even o of an edges for that matter. And then there is other graph that is basically random graph this is a an (Refer Time: 03:33) graph on n vertices, and each edge in this graph each pair of vertices in this graph will have an edge between them with some probability c log n over n for some sufficiently large constant c.

Now the question is which one of them is well connected. As it turns out the answer is that the (Refer Time: 04:05) graph is the well connected graph, whereas the other graph two complete sub graphs connected by some o of n or little o of n edges is not very well connected. And the reason for this is the following, if you look at the addition graph, the any which way we try to partition into two large sub graphs, the number of edges connecting those twos sub graph is going to be big omega of n log n.

Whereas if you look at this other graph with two complete sub graphs, the number of edges connecting those two sub graphs is only o of n or even little o of n. So, it is very easy to disconnect the graph on the left, you only have to remove the edges connecting those two complete sub graphs. It is harder to disconnect the graph on the right, while braking up the graph into two equal sub graphs. So, this gives us a sense of what we are looking for in terms of what is the well connected graph versus what is not well connected graph.

Now this for example, that the graph on the right it is easy to isolate a few vertices, because it is essentially a sparse of graph. So by deleting few edges, we can disconnect the graphs that is really not what we are talking about, if you look at the graph on left disconnecting few removing a few edges will not disconnect the graph. So, you really need to remove all the edges in the middle before you can disconnect the graph. So there is the circle t here where in even though it is easy to disconnect the graph on the right, it is still considerable connected because it is hard to disconnect the graphs so that the two sub graphs are large.

(Refer Slide Time: 06:15)



EXPANSION PARAMETERS

Edge Expansion (aka Cheeger number or Isoperimetric number) of G=(V,E)

$$h(G) = \min_{\substack{S \subset V \\ 0 < |S| \le \frac{|V|}{2}}} \frac{\partial S}{|S|}$$

$\partial S \} \longrightarrow$ # of edges with one end in S and other in V \ S

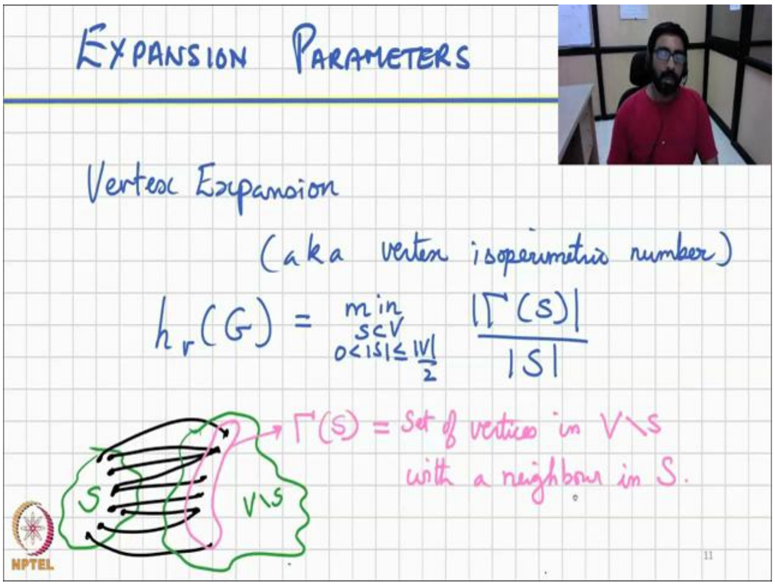
So, with that motivation, we are ready to formally define expander graphs. And as it turns out this notion expanded graphs is captured by not just one parameter, but actually several different parameters and intuitively and they all end up being this capturing the same family of graphs, but each one of them comes with its own flavor. The first parameter is edge expansions. So well known notion it also goes by the name of Cheeger number or isoperimetric number.

So, given this graph G that the edge expansion is as follows to compute h G, you consider the all the possible subsets of the vertex set of cardinality at most half the total number of vertices. And for each one of these subsets, we compute this ratio the number of edges with one end in S and the other end in V minus S.

So, here is the picture that we have over here so this is the set S and the set of vertices outside of S, we this dou S (Refer Time: 07:39) is the number of edges that are going from S to V minus S. And this ratio that we are interested in is dou S over the cardinality of S; and this quantity is minimized, we are considering the least such ratio over all subsets of the vertex set. And this parameter this lowest ratio is called the edge expansion.

So, if you intuitively think about the lowest ratio is obtained in when you consider a way to partition the vertex set into S and V minus S, so that the number of edges going out of S in proportion to the cardinality of S is the least. And every other way to perform this partition this ratio is significantly is more, so that this captures the worst case the cut such that if you dwell strength of the cut is the poorest.

(Refer Slide Time: 08:54)



EXPANSION PARAMETERS

Vertex Expansion

(aka vertex isoperimetric number)

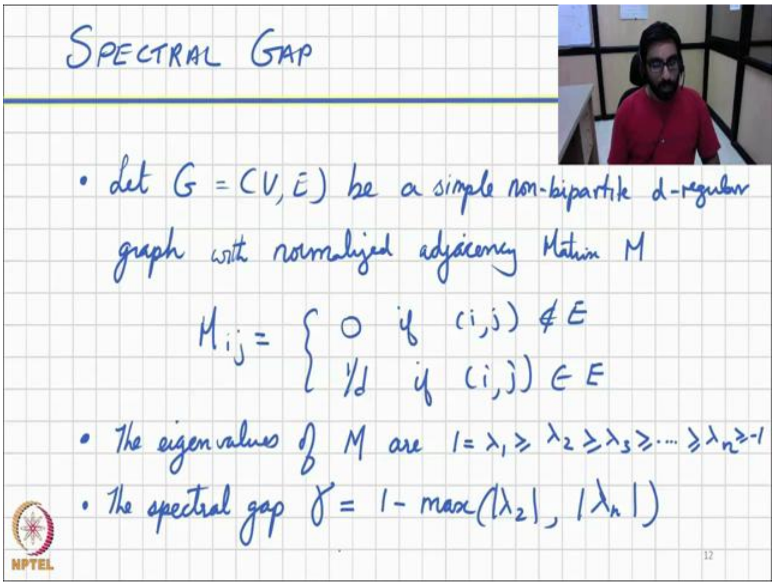
$$h_v(G) = \min_{\substack{S \subset V \\ 0<|S|\le \frac{|V|}{2}}} \frac{|\Gamma(S)|}{|S|}$$

$\Gamma(S)$ = Set of vertices in $V \setminus S$ with a neighbour in S.

In a similar fashion, we can also define the notion of vertex expansion, and this is very, very similar. It is also called the vertex isoperimetric number is just like an edge expansion, we are going to consider all the subsets of the vertex set of cardinality at most half the number of vertices. And here we are going to consider the ratio, the size of the neighborhood of S over the cardinality of S. So, here this gamma of S is the set of vertices in V minus S with a neighbor n S. So, if you going back to this picture, so if you consider the set S over here and V minus S over here previously, we counted the number of edges leaving s, but now we are not counting the number of edges, but rather the number of neighboring vertices of the set S.

And so the notion is quite similar, in fact these two parameters the edge expansion and vertex expansion for graphs that are bounded degree are related to each other by a constant.

(Refer Slide Time: 10:18)



SPECTRAL GAP

- Let $G = (V, E)$ be a simple non-bipartite d-regular graph with normalized adjacency Matrix M

$$M_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ 1/d & \text{if } (i,j) \in E \end{cases}$$

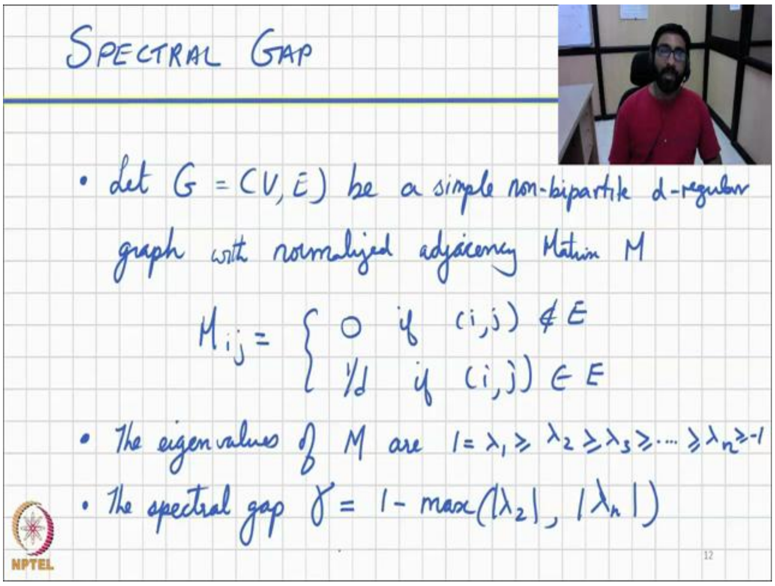
- The eigenvalues of M are $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n \geq -1$
- The spectral gap $\gamma = 1 - \max(|\lambda_2|, |\lambda_n|)$

12

Now we come to completely different way of looking at expander graphs are completely different parameter, call this spectral gap. So, let us consider the graph G; and for simplicity, we are going to assume the graph is simple; it is a non bipartite d regular graph. And let us look at the normalized adjacency matrix m, so here the entries of m are either zero if the edge i j does not exist over the entry i jth entry is 1 over d, if i j belongs to E. So, 1 over d here can be thought of as the as the probability so consider a random walk that is at the vertex i is 1 over d is the probability with which it would transition to vertex j. So, this is called the normalized adjacency matrix, and it also sometimes called the random walks matrix.

So, now let us consider the Eigen values of this matrix m. This is being a symmetric matrix. It is with rows adding up to 1, what we are going to have is Eigen values ranging from 1 to minus 1, the highest Eigen value is going to be 1 itself, and the remaining Eigen values are going to fall between 1 and minus 1. And in fact, now because this is a non-bipartite graph, the least Eigen value is actually going to be strictly larger than minus 1. Now the spectral gap is this gamma is given by 1 minus the maximum of either lambda 2 absolute value of lambda 2 or the absolute value of lambda n.

(Refer Slide Time: 12:29)



SPECTRAL GAP

- Let $G = (V, E)$ be a simple non-bipartite d-regular graph with normalized adjacency Matrix M

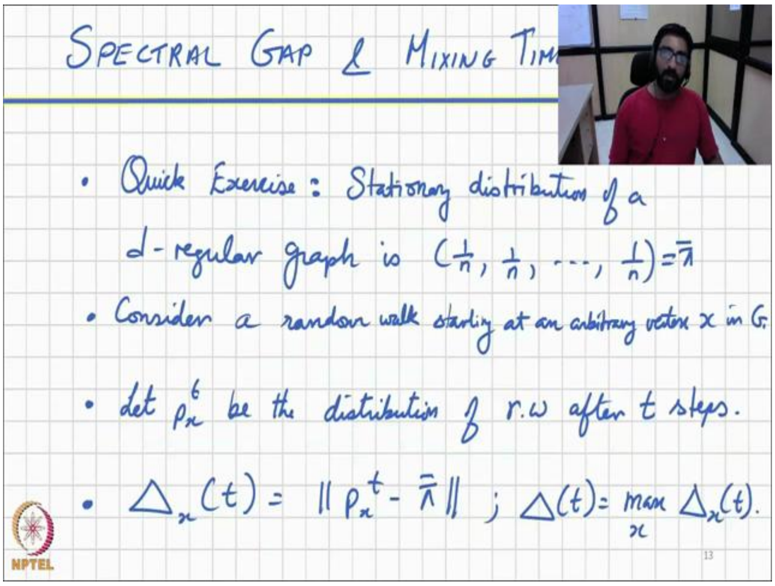
$$M_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ 1/d & \text{if } (i,j) \in E \end{cases}$$

- The eigenvalues of M are $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n \geq -1$
- The spectral gap $\gamma = 1 - \max(|\lambda_2|, |\lambda_n|)$

12

When the spectral gap is large as an at least a constant bounded away from 0 then this graph typically has good expansion and is typically called an expander graph.

(Refer Slide Time: 12:49)



SPECTRAL GAP & MIXING TIM

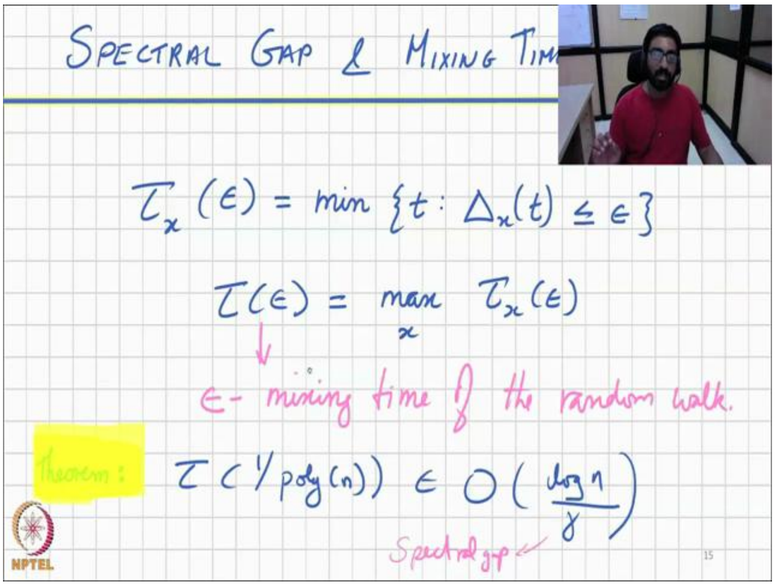
- Quick Exercise: Stationary distribution of a d-regular graph is $(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = \bar{\pi}$
- Consider a random walk starting at an arbitrary vertex x in G.
- Let p_x^t be the distribution of r.w after t steps.
- $\Delta_x(t) = \| p_x^t - \bar{\pi} \|$; $\Delta(t) = \max_x \Delta_x(t)$.

13

Now, having an understanding of this spectral gap, let us try to gain an appreciation for how random walks behave in expander graphs. As a quick exercise consider a random walk on this graph d regular non-bipartite graph, convince yourself that this graph is going to have a stationary distribution this random walk is going to have a stationary distribution of the form 1 by n comma 1 by n and so on, essentially the uniform distribution.

Now having a convinced yourself you may want pause to convince yourself of this. Having convinced yourself of that now consider a random walk starting at some arbitrary vertex x and g. We define p x of t to be the distribution of the random walk, after it has walked for t steps. And we define delta x of t to be the norm of the difference between p x of t and the stationary distribution pi bar. And this norm typically will be the l 2 norm, but other norms are have also been used in this context, but let us strict with l 2 norm. Now delta of t is simply the max over all starting vertices x delta x of t.

(Refer Slide Time: 14:35)



SPECTRAL GAP & MIXING TIM

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon\}$$

$$\underset{\downarrow}{\tau(\epsilon)} = \max_x \tau_x(\epsilon)$$

ϵ- mixing time of the random walk.

Theorem: $\tau(1/\text{poly}(n)) \in O\left(\frac{\log n}{\gamma}\right)$

Spectral gap ←

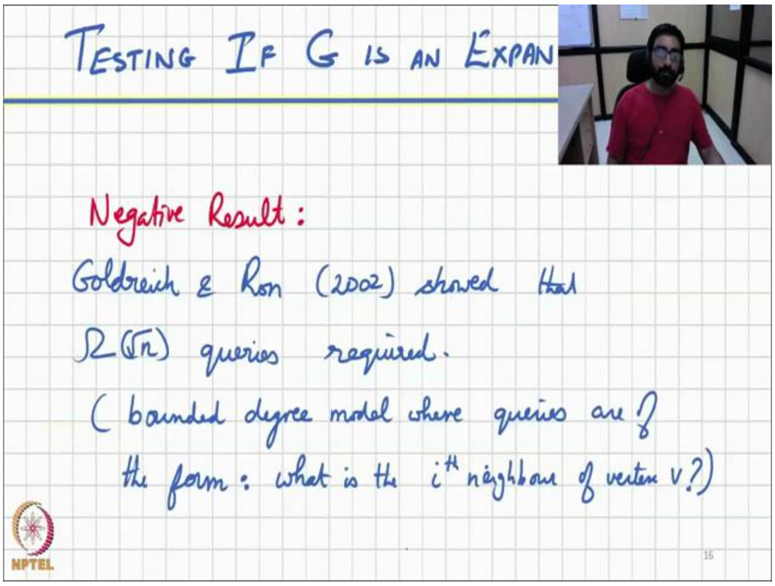
Tau x of E just make sure we understand what we are talking about. This delta t denotes how close the distribution p x of t is with respect to the stationary distribution after some t steps regardless of where we started that is where we are taking max over x. Now let us look at a tau x of epsilon, this is the smallest number of steps needed. Such that you started x and your delta x of t gets to within an epsilon onto less than epsilon which means that the current distribution of the random walk minus the stationary distribution when we take this the l 2 norm of this difference it is within this parameter epsilon.

And this is with respect to a particular starting location, and when we want this to be a (Refer Time: 15:45) to the starting location, we consider to be a this, this parameter tau of epsilon which is the maximum over all possible starting locations or worst a difference possible over all starting locations x tau x of epsilon. And these parameters call the epsilon mixing time of the random walk.

And here is an important theorem that mixing time to get to within 1 over polynomial in n difference with a stationary distribution is at most o of log n over gamma, where gamma is spectral gap. So, just to make sure we understand what we are talking about here this tau of remember tau of epsilon is the smallest this is the time required to ensure that the distribution of the random walk is within an epsilon of very stationary distribution in terms of the l 2 norm.

Now, here in this case we are specified that epsilon to be the polynomially small, so we are able to be get a get over distribution is very close to the stationary distribution within o of log n over lambda or rather gamma and this is a spectral gap. And notice that for expander this spectral gap is a constant bounded away from 0, which means that this right hand side is simply just o of log n. So, what that means is for an expander graph, if you start a random walk at any location, and allow that random walk to walk for o of log n time steps. The distribution of that random walk is going to be polynomially close to the stationary distribution. So, this is very powerful and property of expander graphs.

(Refer Slide Time: 18:01)



TESTING IF G IS AN EXPAN

Negative Result:

Goldreich & Ron (2002) showed that

$\Omega(\sqrt{n})$ queries required.

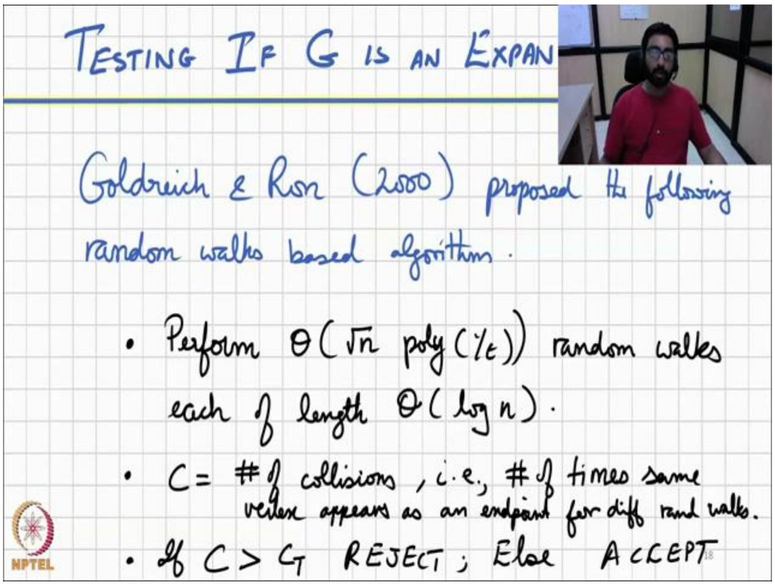
(bounded degree model where queries are of
the form: what is the i^{th} neighbour of vertex v?)

16

So, now let us actually come to the problem of testing whether the graph G is an expander or not. So, just to make sure we understand what we are talking about the input is a graph g and your allowed query access to the graph and this is in the bounded degree model, so the queries are of the form what is the ith neighbor of given vertex V.

So, you can use these queries to perform random walks, because you can go from 1 vertex V to a neighboring vertex uniformly at random and then so on so forth. So, it is perfect for random walks based algorithms, but even under this model, the two people who introduce this problem was Goldreich and Ron and they showed that at you need at least big omega of square root of n queries in order to we are able to separate graphs that are good expanders versus graphs that are not so good expanders.

(Refer Slide Time: 19:10)



TESTING IF G IS AN EXPAN

Goldreich & Ron (2000) proposed the following random walks based algorithm.

- Perform $\Theta(\sqrt{n}\ \mathrm{poly}(1/\varepsilon))$ random walks each of length $\Theta(\log n)$.
- $C =$ # of collisions, i.e., # of times same vertex appears as an endpoint for diff rand walks.
- If $C > C_T$ REJECT ; Else ACCEPT

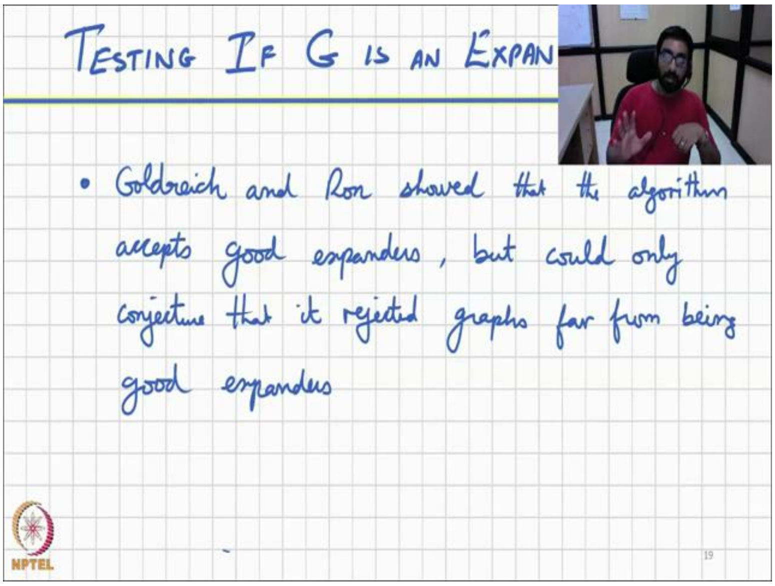
And as it turns out they proposed the very interesting simple algorithm to test whether the graph is a about expander or not. The algorithm works as follows. We perform some theta of square root of n times a polynomial n 1 over epsilon number of random walks. And each of these random walks is of length theta of log n; remember this theta of log n means that these if the graph is a good expander, it is going to these random walks are going to completely mixed, which means that their distribution is going to be very close to stationary distribution.

And now after all of these random walks have walked this distance, you count the number of collisions that is the number of times the same vertex appears as an endpoint for different random walks. Now if the number of collisions is more than some threshold values some C t which is of a well defined threshold value, then you reject the graph, you say that the graph is not a good expander; otherwise you accept.

So what is the intuition behind this, now if the graph is a good expander, these square root of n polynomial 1 over epsilon number of random walks are going to completely mix. And if you recall the (Refer Time: 20:44) paradox, but these square root of n random walks, we are going to be such that they there is not too many collision if they are completely well mixed.

However, if the graph is not an expander there are going to be these random walks are not going to we are able to mix well, which means the distribution of these random walks are going to be non-uniform. And there are going to pockets of the graph, where they are more likely to be found, and therefore, more collisions are likely within these random walks and they are likely terminate lot more collisions. And so this is the key intuition behind this algorithm. So, it is a very, very interesting algorithm, but Goldreich and Ron were not able to say much about this algorithm.

(Refer Slide Time: 21:26)



TESTING IF G IS AN EXPAN

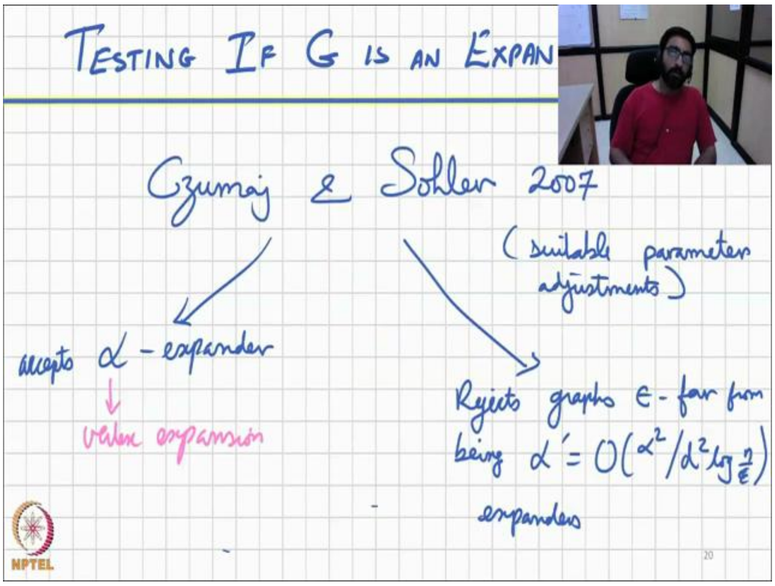
- Goldreich and Ron showed that the algorithm accepts good expanders, but could only conjecture that it rejected graphs far from being good expanders

19

We are able to say that the algorithm accepts a good expanders and this is fairly straightforward because these random walks if a good expander these random walks going to be completely mixed, which means that their distributions going to be very close to you know the uniform distribution. And this means that the number of random walks is chosen so that it is just (Refer Time: 22:08) of the threshold for significant collision to take place.

So, you will not find too many collisions. Remember for example, when we looked at the birthday paradox when we set the parameter to 2 square root of n balls thrown into n bins uniformly at random then we start to see collisions, but here the threshold is such that the a the number of collisions will be well the number of collisions would be small.

(Refer Slide Time: 22:44)

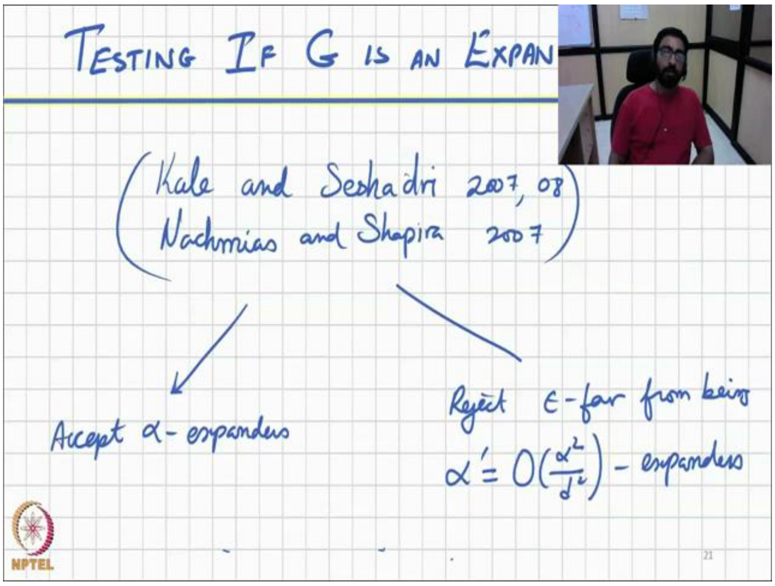

TESTING IF G IS AN EXPAN
Czumaj & Sohler 2007
(suitable parameter adjustments)
accepts α - expander
vertex expansion
Rejects graphs ϵ - far from being $\alpha' = O(\alpha^2/d^2 \log \frac{n}{\epsilon})$ expanders
NPTEL
20

Now first part of 2007 Czumai and Sohler presented analysis of this algorithm, not for they had to choose suitable parameters in order to get the analysis to work. And what they showed is that the algorithm, the same algorithm as a proposed by Goldreich and Ron, but we know that just a parameter adjustments is able to accept alpha expanders this is here we are using vertex expansion.

Moreover, they also showed that this algorithm rejects graphs that are far, in particular epsilon far from being alpha prime expanders. And here the alpha prime is at most o of alpha square divided by d square log n over epsilon. Now notice that this rejection happens only if the given graph is not only far from being an expander actually far from being a worst expander than this alpha over here. So, keep in mind that this alpha is going to be a parameter that is strictly less than 1.

So, this as a result this alpha prime a when you square the alpha here this alpha prime is going to be a worse the expansion parameter. And so the rejection happens with constant probability only if the graph is both of an epsilon far from being a worse expander. So, the result is in that sense it is not very impressive, but even getting this result was quite a challenge.

(Refer Slide Time: 24:52)



TESTING IF G IS AN EXPAN

(Kale and Seshadri 2007, 08
Nachmias and Shapira 2007)

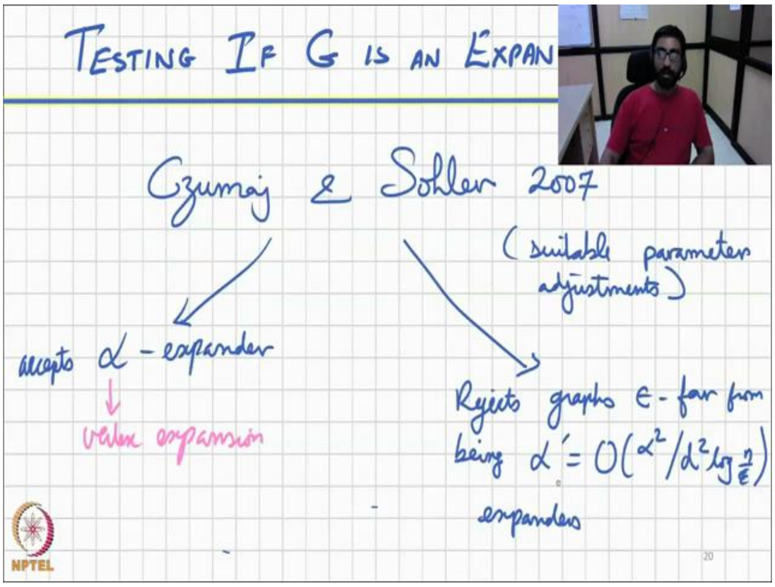
Accept α- expanders

Reject ϵ-far from being $\alpha' = O(\frac{\alpha^2}{d^2})$ - expanders

21

And then soon thereafter there are been some improvements, but again nevertheless of a same general form. So, the improvements made by Kale and Seshadri, and then Nachmias and Shapira and they were able to show that the same algorithm can accept alpha expanders, and also with some constant probability reject graphs that are epsilon far from being alpha prime expanders, where here they improve the alpha prime to be o of alpha square divided by d square.

(Refer Slide Time: 25:28)



TESTING IF G IS AN EXPAN

Czumaj & Sohler 2007

(suitable parameter adjustments)

accepts α - expander

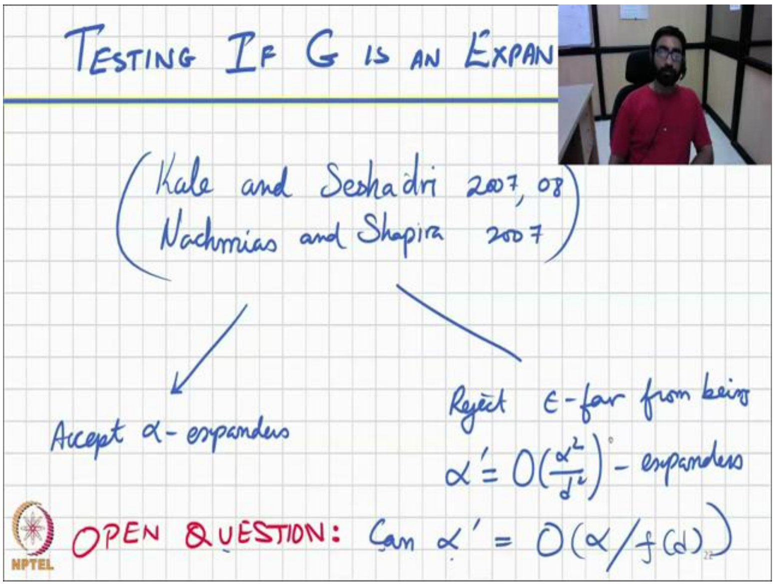
vertex expansion

Rejects graphs ϵ- far from being $\alpha' = O(\alpha^2/d^2 \log \frac{n}{\epsilon})$ expanders

20

So, they got rid of these this log n over epsilon in the denominator over here.

(Refer Slide Time: 25:34)



TESTING IF G IS AN EXPAN

(Kale and Seshadri 2007, 08
Nachmias and Shapira 2007)

Accept α- expanders

Reject ϵ-far from being $\alpha' = O(\frac{\alpha^2}{d^2})$ - expanders

OPEN QUESTION: Can $\alpha' = O(\alpha/f(d))$

But, still this alpha prime is now still remains as a function of alpha square. So, the open question is can this alpha prime can be improved to alpha over some function of d and not be not be dependent on the square of alpha. So, this is the big open question, so this is actually been a very interesting question, mainly because the algorithm is so simple and elegant, but the analysis turns out to be quite nontrivial. So, we are for that the analysis beyond the scope of our course, but the algorithm itself is a something that we should be able to appreciate a lot.

So, with that, we will we come to the end of segment 1. In segment 2, we will be talking about testing bipartite (Refer Time: 26:44).