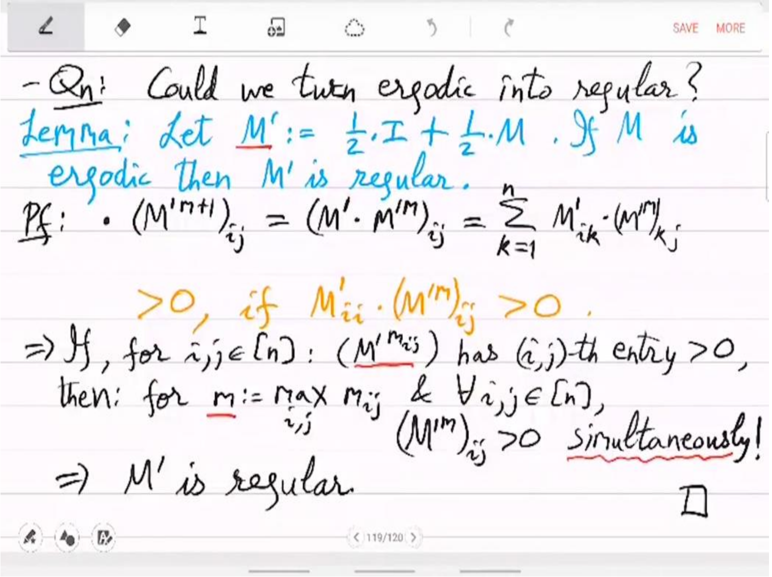
Probability for Computer Science
Prof. Nitin Saxena
Department of Computer Science and Engineering
Indian Institute of Technology - Kanpur

Module - 6
Lecture - 23
Cell Genetics

Last time we were doing Page Rank algorithm. We discussed 3 strategies to do that. And the best strategy was that you introduce this probability p that the web-surfer will either stray to a random page in the internet or it will follow one of the links on the current webpage. So, this allowed every entry in the matrix to become positive, because the surfer can go now, because every webpage can be visited now potentially.

So, that gave you stationary distribution which gave you the rank of pages on the internet. Another notion we want to study weaker than regularity is ergodic, ergodicity. So, ergodic means that for every position i comma j, some exponent m, small m, that mth power will have ijth entry positive. So, regularity implies ergodicity, but converse is not true. For example, this 0 1 1 0 matrix. So, what do you do then? Can you turn ergodic into regular? So, we do that by introducing these additive terms on the diagonal.

(Refer Slide Time: 01:41)



- Qn: Could we turn ergodic into regular?

Lemma: Let $M' := \frac{1}{2}.I + \frac{1}{2}.M$. If M is ergodic then M' is regular.

Pf: • $(M'^{m+1})_{ij} = (M' \cdot M'^{m})_{ij} = \sum_{k=1}^{n} M'_{ik} \cdot (M'^{m})_{kj}$

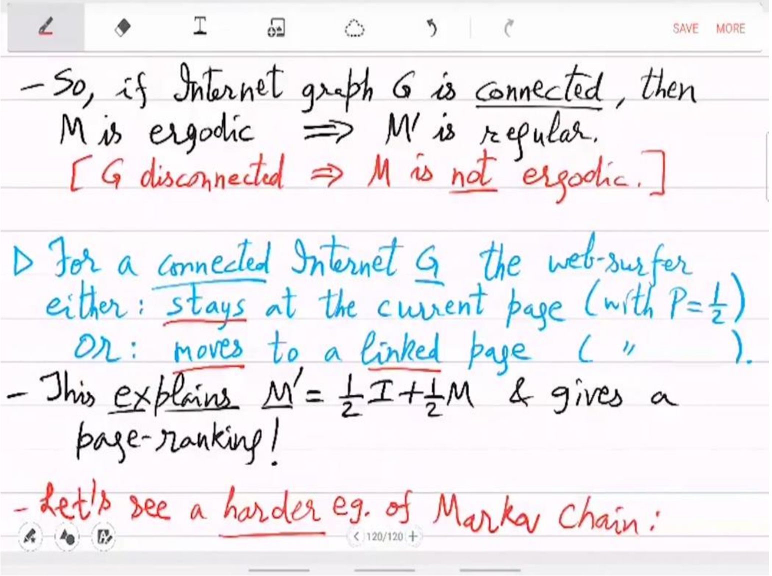
> 0, if $M'_{ii} \cdot (M'^{m})_{ij} > 0$.

$\Rightarrow$ If, for $i,j \in [n]$: $(M'^{m_{ij}})$ has (i,j)-th entry > 0,
then: for $m := \max_{i,j} m_{ij}$ & $\forall i,j \in [n]$,
$(M'^{m})_{ij} > 0$ simultaneously!

$\Rightarrow$ M' is regular. $\square$

So, if you do this i + m, then M prime will be regular, if M was ergodic. So, that is a partial converse of regularity and ergodicity, the connection between the two. So, this proof we are done, and let us now just quickly interpret this.

(Refer Slide Time: 02:09)



- So, if Internet graph G is connected, then M is ergodic $\Rightarrow$ M' is regular.
[G disconnected $\Rightarrow$ M is not ergodic.]

▷ For a connected Internet G the web-surfer either: stays at the current page (with $P=\frac{1}{2}$)
Or: moves to a linked page (").

- This explains $M' = \frac{1}{2}I + \frac{1}{2}M$ & gives a page-ranking!

- Let's see a harder eg. of Markov Chain:

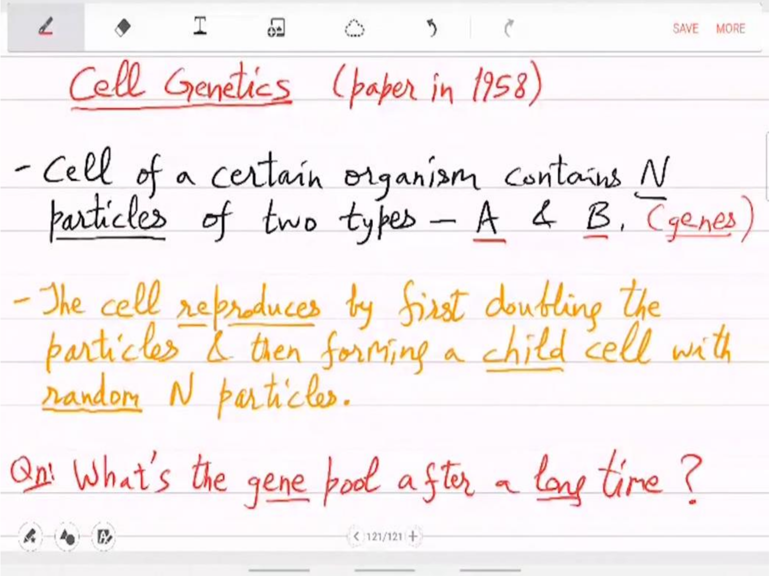
What does this mean? So, this means that; so, if internet graph G, if the graph is a connected graph, if there are no isolated vertices, because, if there are isolated vertices, then again it cannot be ergodic, because not every entry i comma j can be made positive. But in case it is connected, then you can see that at some point, there will be a path from i to j. So, ijth entry can be made positive.

So, if the graph is connected, then this M is ergodic which implies that this M prime is regular. And note here that G disconnected implies that M is not ergodic. So, then, this method ergodic idea will not help, because again there are these isolated vertices which you cannot reach. So, i comma j can never be made positive. So, for a connected internet graph, there is a different strategy, this strategy 4 which says that; so, what does it say?

So, for a connected G, the web-surfer either stays at the same page with probability half or moves to a link, which is again probability half. So, either you stay or you move to a link page. And this is what is the interpretation of M prime equal to half of i + m. So, this explains M prime equal to half of i plus half of M and gives a page-ranking. So, this is the physical interpretation of what we just did, when we went from ergodic to regular.

And this you can do if the graph, internet graph was connected; otherwise you do what you did in strategy 3. So, that finishes this example. So, where are we? So, we have finished Markov chain stationary distribution and given this Page Rank algorithm, which is a very useful, nice application. So, now, let us see a harder example. So, it will be harder because there will be no ergodicity. So, there will be actually no stationary distribution, but still somehow we will use these concepts to understand what is happening at n equal to infinity.

(Refer Slide Time: 06:43)



Cell Genetics (paper in 1958)

- Cell of a certain organism contains N particles of two types – A & B. (genes)
- The cell reproduces by first doubling the particles & then forming a child cell with random N particles.

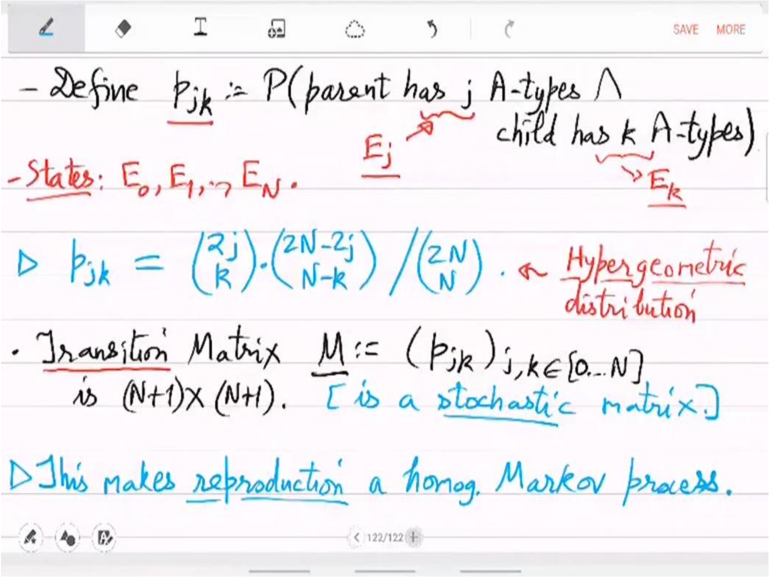
Qn: What's the gene pool after a long time?

So, let us see this cell genetics example. So, this is a paper from 1958; based on that. So, what happens here is, cell of a certain organism contains, let us say big N particles of 2 types, A and B. So, you can call them genes. So, particles are of 2 types, there are these A particles and there are B particles. So, they may be, A maybe half of the particles, N by 2; or it may be 0; or it may be N; it can be anything. And remaining are type B.

So, think of this as gene A and gene B. Now, when reproduction happens, this cell reproduces by first doubling the particles, then forming a child cell with random N particles. So, first these big N many particles are made 2 times N. And then a random N subset of particles is chosen, which becomes child cell. So, that is a reproduction or replication. So, we want to understand what happens?

So, now the child will also reproduce and then the grandchild will reproduce and so on. So, we want to see what happens in the end. Then, this is repeated infinitely many times. Then, what happens to the gene pool? So, that is the question. So, what is the gene pool after a long time? This is what we want to understand. What is the distribution of big A and big B in the end?

(Refer Slide Time: 09:52)



- Define $\underline{p_{jk}} := P(\text{parent has } j \text{ A-types} \wedge \text{child has } k \text{ A-types})$ $\underline{E_j}$ E_k

-States: $E_0, E_1, \ldots, E_N$.

▷ $p_{jk} = \binom{2j}{k} \cdot \binom{2N-2j}{N-k} \Big/ \binom{2N}{N}$. ← Hypergeometric distribution

• Transition Matrix $\underline{M} := (p_{jk})_{j,k \in [0..N]}$ is $(N+1) \times (N+1)$. [is a stochastic matrix]

▷ This makes reproduction a homog. Markov process.

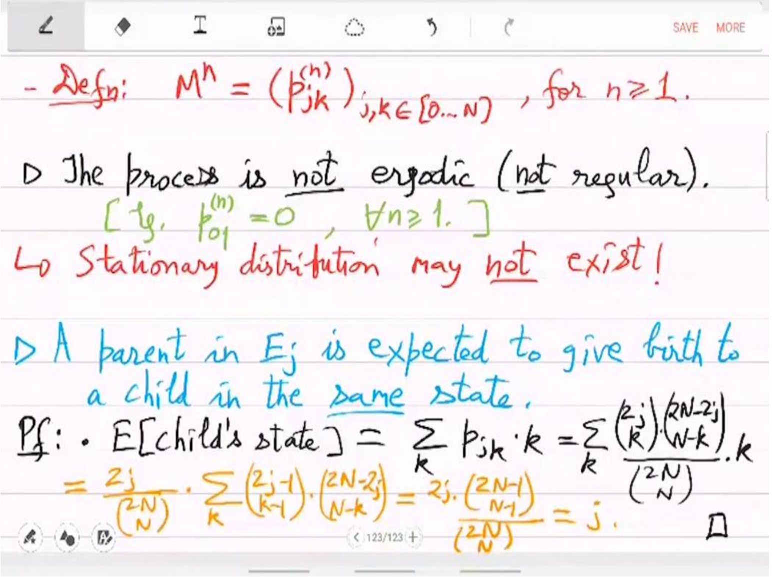
So, let us define P jk as the probability; that parent has j A-types, while the child has k A-types. So, this is the probability that from a parent of distribution j comma N - j. So, j many A-type particles and N - j many B-types. When the reproduction happens, the child has k A-type particles and N - k B-types. So, let us call this event E j, and let us call this event E k. So, from E j to E k, that is the state change. So, the states are E 0, E 1 to E N.

So, A-type can be from 0 to big N. So, these are the N + 1 states. So, parent is in one state. Then, child is in some other state. And then, the grandchild is in some other state and so on. So, what is this probability? This is easy to compute. So, this is, parent had j A-types. So, that becomes 2j on doubling. And out of 2j, child will need k. So, these many options. Now, remaining is 2N - 2j. And from that, child will need N - k.

Those are the favourable possibilities. And what is the total number of possibilities? So, that is like choosing out of 2N particles, N particles; that is 2N choose N. So, this is exactly the probability of going from E j to E k. By the way, this is called hypergeometric distribution. This is the expression. And now, what we have just done is, we have set up a Markov chain. So, let us look at the transition matrix.

So, the transition matrix M is these probabilities, transition probabilities. This is an N + 1 cross N + 1 matrix. And you can check that for the jth row; so, from state E j, all possible E k's, these are the probabilities; so, the sum has to be 1, because the child has to get to some state, right? So, the probability of the sum is 1, and it is also a partition. So, this is stochastic. In fact, this is a nice physical interpretation of the identity that Sigma P jk is 1, for a fixed j, for any j, Sigma P jk is 1. So, you get that identity from here. So, this makes cell replication or reproduction a homogeneous Markov process.

(Refer Slide Time: 14:35)



- Defn: $M^n = \left(p_{jk}^{(n)}\right)_{j,k\in[0..N]}$, for $n\geq 1$.

▷ The process is not ergodic (not regular).
[Eg. $p_{01}^{(n)} = 0$, $\forall n \geq 1$.]
↳ Stationary distribution may not exist!

▷ A parent in E_j is expected to give birth to a child in the same state.

Pf: • $E[\text{child's state}] = \sum_k p_{jk}\cdot k = \sum_k \frac{\binom{2j}{k}\cdot\binom{2N-2j}{N-k}}{\binom{2N}{N}}\cdot k$
$= \frac{2j}{\binom{2N}{N}}\cdot\sum_k \binom{2j-1}{k-1}\cdot\binom{2N-2j}{N-k} = \frac{2j\cdot\binom{2N-1}{N-1}}{\binom{2N}{N}} = j.$ □

So, define M to the little n to have entries P jk to the, just symbolically; for nth power, these are the entries, P jk n. So, you can now study the evolution of the Markov chain. Question here is, is the process ergodic or regular? So, this process is non-ergodic. And hence, it is also not regular. Why is that? Well, because you can see that there are these entries jk which can never be made positive.

For example, this P 0k is 0, or P 01 is 0 for all n; because state E 0, there was no A-type, there was no gene A, so, all the descendants will be missing this gene. That is what it says. So, this is clearly not ergodic and it is irregular. So, there is no stationary distribution. So, basically, then, this stationary distribution may not exist. We are not sure, but it may not exist. It is a non-ergodic, it is a bad process in terms of studying the limiting case.

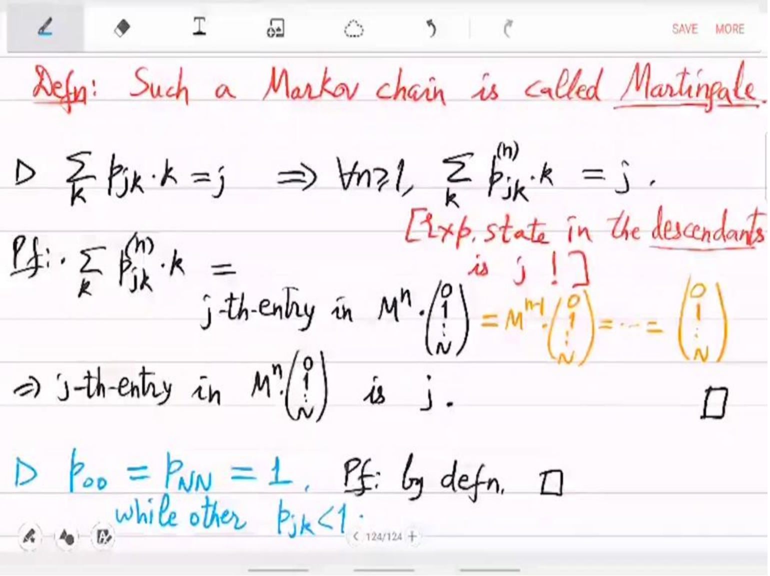
So, what do we do? So, we will keep studying properties of this. It has some nice properties. And from that, we will try to deduce something interesting. So, first property is a parent in state E j is expected to give birth to a child in the same state. This is one interesting property that a parent who is in state E j, if you look at the expectation over the states of the child, it is exactly j. So, the expected state of the child is the same as that of the parent.

So, let us prove this. It is an interesting fact. So, expectation over child's state. This is equal to probability of moving from j to k and times k. And do this over all k. So, what is this? This is equal to, look at the probability P jk that we calculated, this expression. Let us just put that expression here. So, you will get 2j and N - 2j times k. What is this? So, this is equal to, basically this 2j choose k times k, you can simplify; you can bring out 2j.

So, let us bring out 2j and similarly 2N choose N outside the summation. And what remains is 2j - 1, k - 1, N - 2j, N - k; that is it. So, this is like; made a mistake; let me change. So, initially, 2N was distributed in N. And now what you have is 2N - 1 distributed in N - 1. So, 2N - 1 is distributed into N - 1 via this, you started with 2j - 1 A-genes, A-type cells going into k - 1 A-type cells.

This sum will be everything, all possibilities, which is 2N - 1 choose N - 1. So, you get 2j times 2N - 1, N - 1 by 2N choose N, which is equal to what? So, you get basically 2j by 2. So, that is j. That is what we claimed in the beginning. So, what this is saying is that, expectation of the child's state is j when you started with a parent in state j. This will be a very useful equality. Now, such a Markov chain which satisfies this property, kind of preserving the state in the expectation, this is called martingale.

(Refer Slide Time: 21:04)



Defn: Such a Markov chain is called Martingale.

▷ $\sum_k p_{jk} \cdot k = j \Rightarrow \forall n \geq 1, \sum_k p_{jk}^{(n)} \cdot k = j.$

[Exp. state in the descendants is j !]

Pf: $\cdot \sum_k p_{jk}^{(n)} \cdot k =$ j-th-entry in $M^n \cdot \begin{pmatrix} 0 \\ 1 \\ \vdots \\ N \end{pmatrix} = M^{n-1} \cdot \begin{pmatrix} 0 \\ 1 \\ \vdots \\ N \end{pmatrix} = \cdots = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ N \end{pmatrix}$

$\Rightarrow$ j-th-entry in $M^n \cdot \begin{pmatrix} 0 \\ 1 \\ \vdots \\ N \end{pmatrix}$ is j. □

▷ $p_{00} = p_{NN} = 1$, Pf: by defn. □

while other $p_{jk} < 1$.

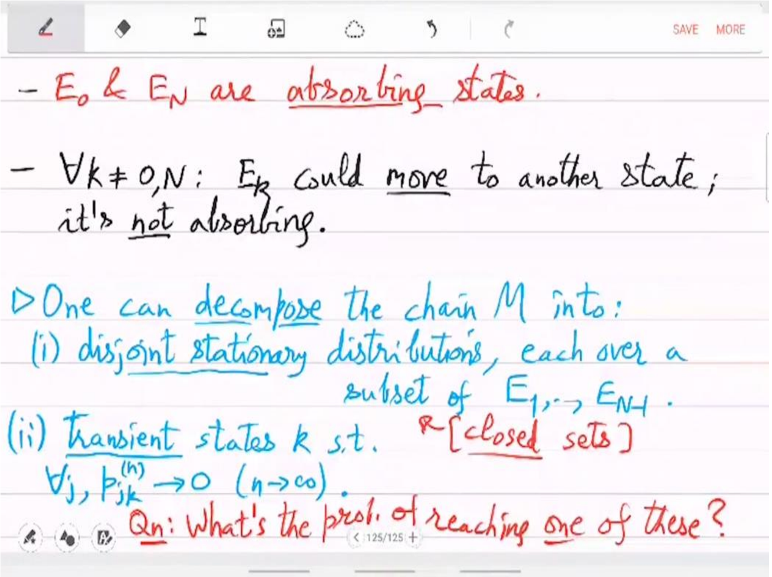
So, such a Markov chain is called martingale. This is a very special property that in expectation, the state remains the same, as the Markov chain evolves. Now, from this, what you can deduce is summation P jk times k over all k equal to j. This implies that for all N, if you look at the evolution at any time N over all k, that is also j. So, not only in the first step, but also in the descendants, expected state is remains j.

So, expected state not only in the child, but in all the descendants; that is what it is saying. How do you prove this? Let us just show this for N = 2. That is, or okay, we can give a direct proof also. So, look at this. P jk n times k. So, this is equal to jth entry in M to the n times action on the vector 0 1 ... N. So, M to the n has these elements P jk n. And then, for the expectation, you are basically multiplying it with this vector where the jth entry is just j, going from 0 to N.

And what is that? So, that is the same as the jth entry. So, I can write this matrix; let us just focus on the matrix. So, from the hypothesis, M's action on this vector is itself. So, you get M to the N - 1 times the same vector. You do this many times. You end up with the same vector. So, because, it is basically, the vector does not change its invariant. So, which means that the jth entry in M to the n times this vector is just j. That is it. That is the proof.

So, that shows Sigma P jk n times k is also j. Let us write down one more property. So, probability of going from 0 to itself and probability of going from N to N; so, the case where A is absent and the case where B is absent, obviously, you cannot go outside this; so, the probabilities are actually equal. This you can also see from the definition. This or the expression for P jk, right? The proof is just by definition. So, 0 and N states are very special, while other P jk's, they are less than 1. So, this is why we will call E 0 and E N absorbing states.

(Refer Slide Time: 25:29)



- E_0 & E_N are absorbing states.
- $\forall k \neq 0, N$: E_k could move to another state; it's not absorbing.

▷ One can decompose the chain M into:
(i) disjoint stationary distributions, each over a subset of $E_1, \ldots, E_{N-1}$. [closed sets]
(ii) transient states k s.t. $\forall j, p_{jk}^{(n)} \to 0$ $(n \to \infty)$.
Qn: What's the prob. of reaching one of these?

So, if you reach E 0 or E N, then you cannot go out. Then you are absorbed here. What about other possibilities? So, let us look at k other than these absorbing states. So, E k, this nonabsorbing state, this may change to something else. So, E k could move to another state. So, it is not absorbing. If k is from 1 to n - 1, basically the F k many A-type particles, then it will actually increase when you double.

And then, the child has many possibilities, 0 to 2k possibilities are there. So, this is not absorbing. So, it is called transient. So, you have these absorbing states; you have transient states. So, what can we say about small n equal to infinity. What happens in the end, right? That question is still outstanding. So, what you can do is, one can decompose the chain M into stationary distributions, each over a subset of E 1 to E N - 1.

So, it is possible that there is not one, but many stationary distributions; like E 1, E 2 may be a stationary distribution; then E N by 2, E N by 3 or E N by 3 + 1; those three may form another stationary distribution. So, once the process reaches there, then that is where it is stuck. So, those are called closed sets; or what may happen, others are the transient states; so, let me remove transient from here.

So, transient states, k such that P jk n tends to 0 as n tends to infinity, for all j. So, these transient states k are such that, no matter which j you start from, in the limit, you will not reach them. So, you will only be able to reach the stationary distributions. And these will become kind of unreachable, because they are basically, when you reach them, there is still a probability of going out of them, and they are not part of a closed set.

So, these are the 2 possibilities. This is how the chain will evolve after infinitely many steps; either it goes into a stationary distribution, which may be of many types; it is kind of, you can think of it as a partition of the states, you go in some part; or you may go into a bad part which is these transient states, but actually in the limit, you will not reach there, because the probability becomes extremely small.

So, either you reach there with extremely small probability or with good probability you reach some stationary distribution. So, the question here is, what is the probability of reaching one of these? So, we will actually try to do this next, that what is the chance that you reached some closed state and how many closed sets?