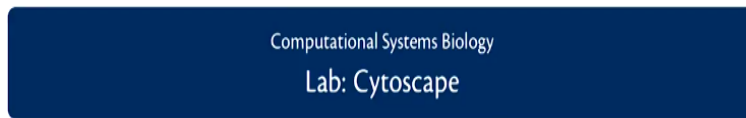


Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology – Madras

Lecture - 25
Lab: Cytoscape

So today, we will have a lab video.

(Refer Slide Time: 00:13)



- ▶ Cytoscape Introduction
- ▶ STRING Database
- ▶ Loading and Visualising Networks
- ▶ NetworkAnalyzer

Karthik Raman

Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences
Initiative for Biological Systems Engineering (IBSE)
Robert Bosch Centre for Data Science and Artificial Intelligence (RBC DSAI)
INDIAN INSTITUTE OF TECHNOLOGY MADRAS



Wherein we demonstrate this very interesting software tool called Cytoscape. So I will introduce you to Cytoscape as well as the STRING database which is a very good repository of different protein-protein functional association networks and I will also talk you through how you will load and visualize networks using Cytoscape and a very interesting plug-in for doing a lot of network analysis in terms of identifying centrality measures or even visualization which is known as network analyzer.

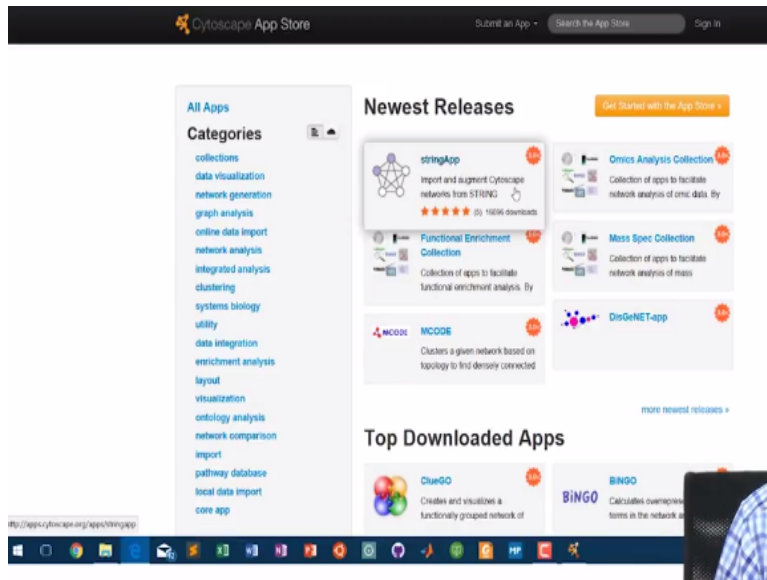
(Refer Slide Time: 00:41)



So, today let us look at cytoscape, this is the very powerful tool for doing network analysis. We previously studied how we can do network analyze with MATLAB BGL that is a little programmatic and there is also another tool known as there is other python package called network X which is very powerful for doing network analysis. Today, let us see how we can do it more simply through a graphical user interface with a cytoscape.

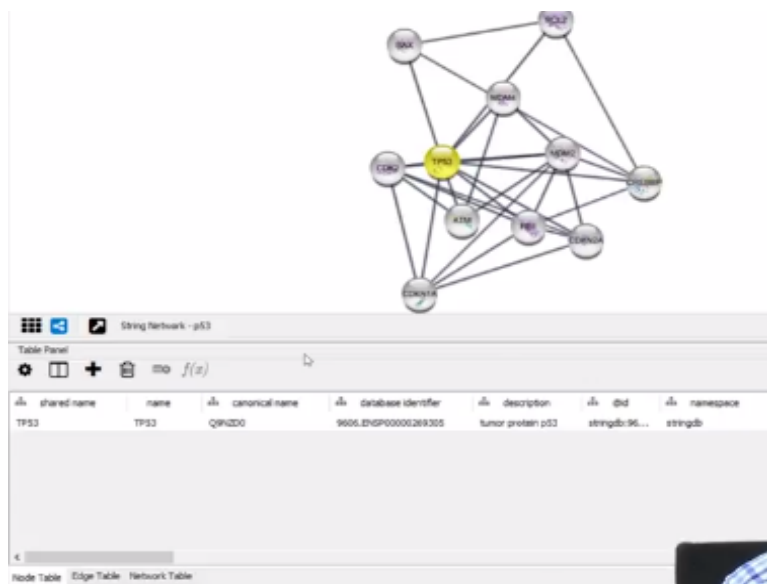
And cytoscape can actually be used to prepare really beautiful illustrations and so on. So cytoscape is available from cytoscape.org and it has many features. We will probably just try to look at a few simple features. There are several interesting useful plug-ins and so on which are known as apps.

(Refer Slide Time: 01:22)



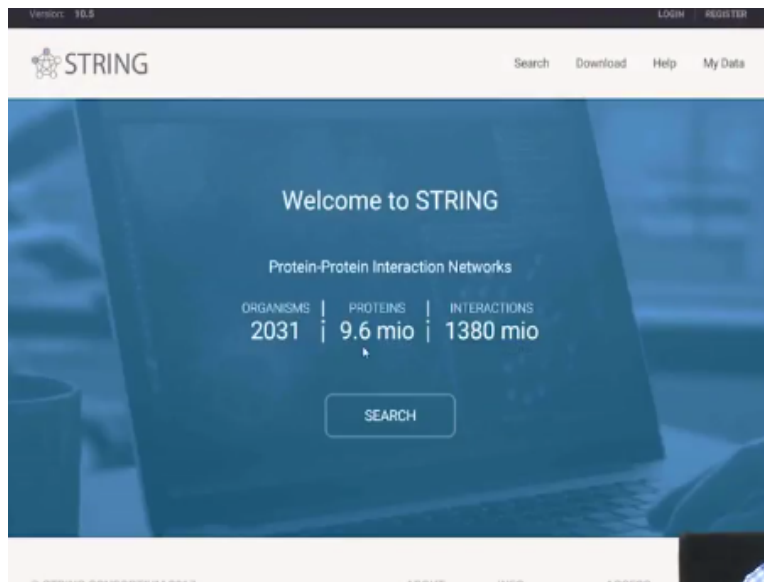
So for example there is a string app this probably the most useful because you can directly get string networks into cytoscape that is may be try installing one of those.

(Refer Slide Time: 01:59)



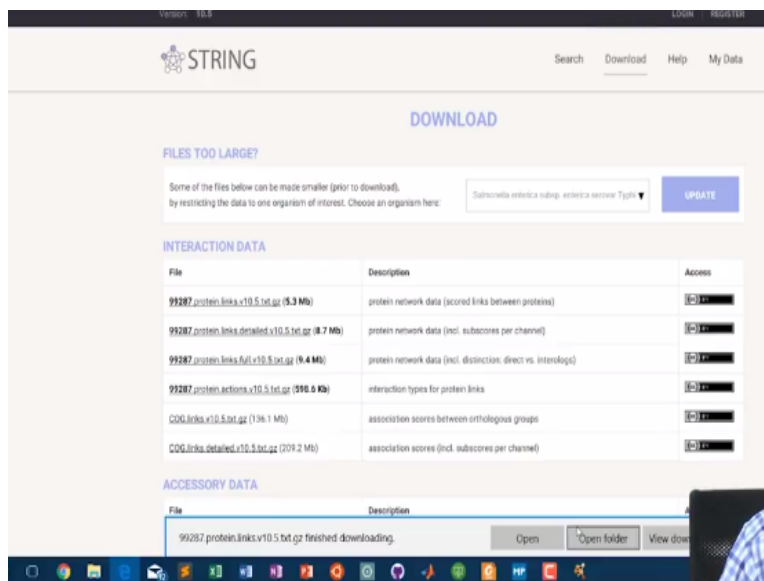
You can directly fetch proteins from the string database and start expanding the network and so on, but more than using a string app or any of those things I think the best way to go about this would be downloading network information directly from the string. So let us go to string DB.Org.

(Refer Slide Time: 02:15)



So string has information on 2031 organisms about 9.6 million proteins and about 1.3 billion interactions in all. So we can go to download. The good part of string is it allows downloads on a per organism basis.

(Refer Slide Time: 02:31)



So you just pick an organism of interest. Let us say salmonella typhimurium It2. So all organisms are basically identified by that Tax ID the taxonomy ID. This is something like 99287 for salmonella and 511145 for E. coli if you want. So you can pick any organism of interest. So let us see how this file looks like. So this will be quite a large file.

(Refer Slide Time: 03:04)

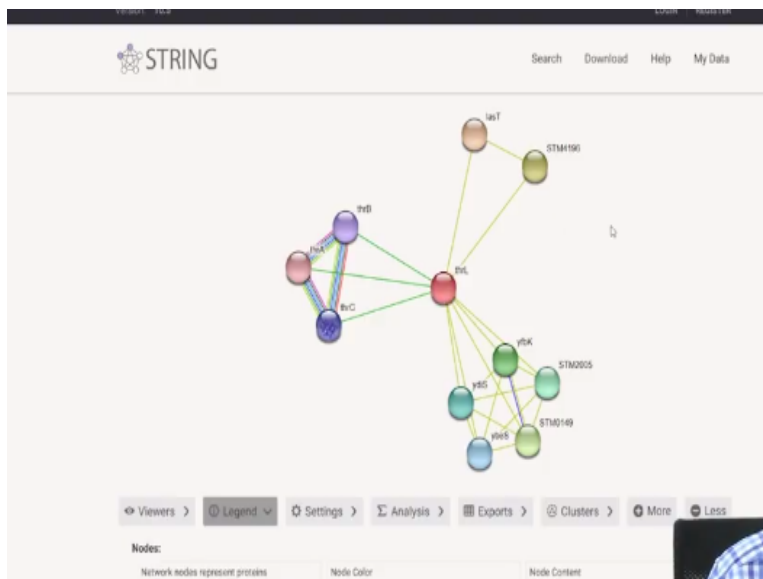
```

1 protein1 protein2 combined_score
2 99287.STM0001 99287.STM4476.S 260
3 99287.STM0001 99287.STM1240 205
4 99287.STM0001 99287.STM2772 208
5 99287.STM0001 99287.STM2805 710
6 99287.STM0001 99287.STM0004 622
7 99287.STM0001 99287.STM2515 268
8 99287.STM0001 99287.STM4488 189
9 99287.STM0001 99287.STM0419 502
10 99287.STM0001 99287.STM1722 266
11 99287.STM0001 99287.STM2704 510
12 99287.STM0001 99287.STM1802 209
13 99287.STM0001 99287.STM0906 597
14 99287.STM0001 99287.STM3034 499
15 99287.STM0001 99287.STM3028 189
16 99287.STM0001 99287.STM2816 418
17 99287.STM0001 99287.STM0197 311
18 99287.STM0001 99287.STM0003 621
19 99287.STM0001 99287.STM4246 206
20 99287.STM0001 99287.STM0301 390
21 99287.STM0001 99287.STM1350 522
22 99287.STM0001 99287.STM0529 204
23 99287.STM0001 99287.STM4571 507
24 99287.STM0001 99287.STM4219.S 506
25 99287.STM0001 99287.STM2016 186
26 99287.STM0001 99287.STM0658 311
27 99287.STM0001 99287.STM4218 317
28 99287.STM0001 99287.STM2616 420
29 99287.STM0001 99287.STM4257 263
30 99287.STM0001 99287.STM4065 311
31 99287.STM0001 99287.STM4261 208
32 99287.STM0001 99287.STM1829 597
33 99287.STM0001 99287.STM2063 208
34 99287.STM0001 99287.STM4436 268
35 99287.STM0001 99287.STM2138 499
36 99287.STM0001 99287.STM4196 715

```

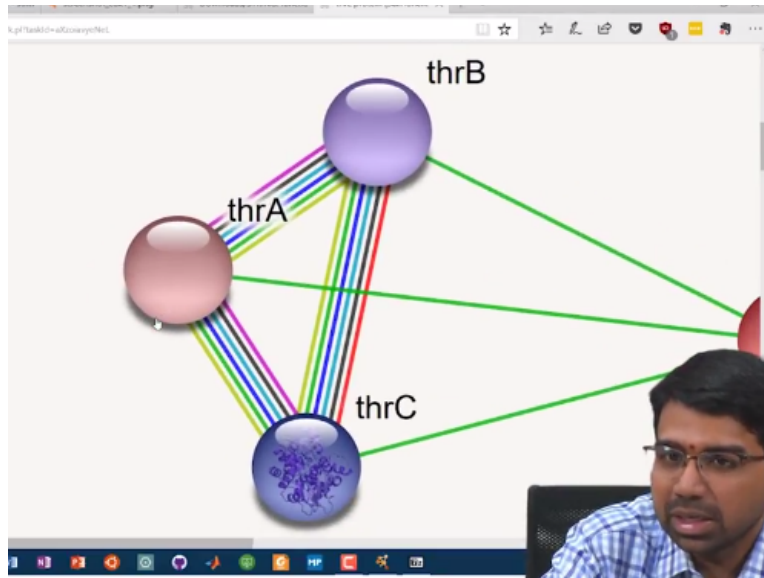
So it has 3 columns; organism ID. protein ID, space organism ID. protein ID space score of interaction. Do you remember how the string's score is calculated? So string has information coming from several channels. Let us just go back to string.

(Refer Slide Time: 03:33)



So this is the first protein from salmonella which is TCRL or whatever it is. So you can do a lot of interesting analysis right off of the string database itself would I think you know we would like to do it computationally offline. So this even tells you if there is a structure available and so on. Can you see that there is a structure that is there? Now what is interesting for us is this.

(Refer Slide Time: 04:09)



Can you see the multicolored lines? Each of those lines represent a different channel of interaction coming from curative databases, experimentally determined, gene neighborhood, gene fusions, gene co-occurrence, text mining, co-expression, homology and so on. So you have multiple evidences that are used for that are available for every interaction and these are finally integrated so how are these integrated.

(Refer Slide Time: 04:44)

STRING-db

P_1-P_2 0.35

P_1-P_2 0.45

P_1-P_2 0.86

$$1 - \frac{(1-0.35)(1-0.45)}{(1-0.86)}$$

$$\text{score} = 1 - \prod_i (1 - p_i)$$

So string has a nice way of integrating this information. Let us say score are. Let us say you have a protein P1 and a protein P2 and interaction score is 0.35 and for the same P1, P2 you have another interaction score of 0.45. This comes from filogeny, this comes from say text mining

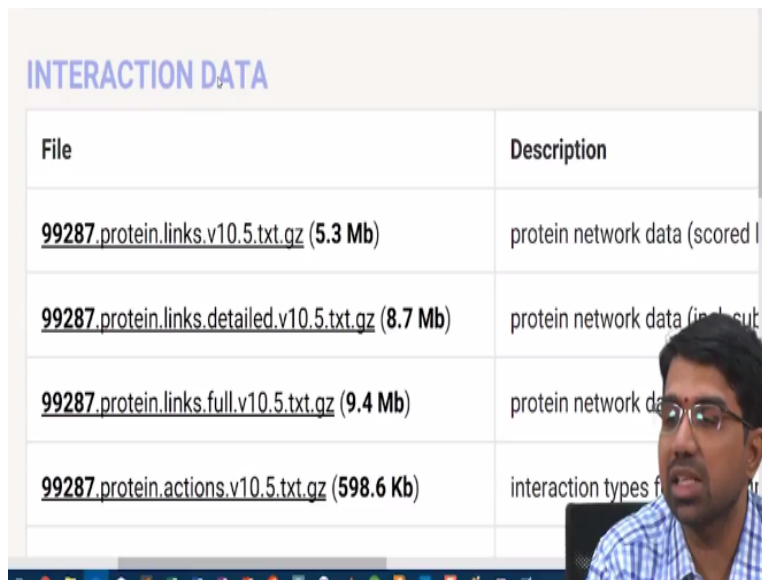
and something of this sort. How do you integrate these course? You could take average. The idea string is a little better.

What it does is? If this is you can consider this as a probability of interaction how do you combine these probabilities so it depends upon what you want to ask so if this is the probability that p1 and p2 interact based on channel 1. This is the probability that p1 and p2 interact based on channel 2 and this is the probability that p1 and p2 interact based on channel 3. What is the probability that p1 and p2 interact?

So it is going to be $1 - (1 - 0.35)(1 - 0.45) * (1 - 0.80)$. So, basically the final interaction score is going to be 1 - does it make sense? So what is the probability that they do not interact at all. The probability that they do not interact according to this according to this, according to this, will be this number product of these probabilities 1 - of that is the probability that they interact. So this is how you combine string combines these scores up. So it is quite useful.

So it is very useful if you choose to take the string there is 1 more file here actually. So there is a file with links detail.

(Refer Slide Time: 07:05)



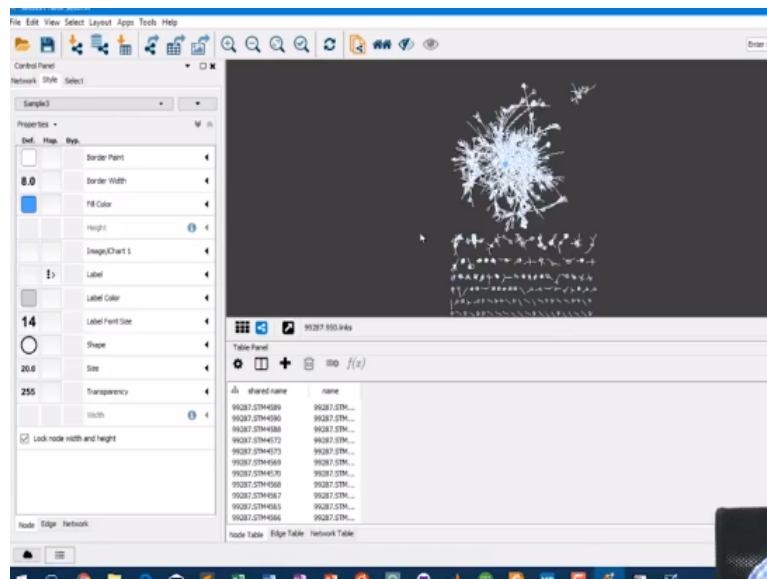
File	Description
99287.protein.links.v10.5.txt.gz (5.3 Mb)	protein network data (scored)
99287.protein.links.detailed.v10.5.txt.gz (8.7 Mb)	protein network data (incl. cut
99287.protein.links.full.v10.5.txt.gz (9.4 Mb)	protein network da
99287.protein.actions.v10.5.txt.gz (598.6 Kb)	interaction types f

So this will have all the channel specific course. So if you take all the channel specific course and you want to describe a couple of channels and rewrite this course now you know how to do

it and in string there is like a small 1 on 1 is there. All scores are given in 3 digits. So it will be written as 0.356 or actually just 356 and there are a few cut off for confidence so 400 is you know medium confidence, 700 is high confidence.

You can also use 900 if you want. I liked to use 900 to start off with because it will give you a very sparse network that can be more easily analyzed, but if you are going to do computationally you can do obviously any confidence that you want. So for using cytoscape I prefer something like 900. So let us see how we go about doing that. So let us import a network into cytosphere. So already have a filter network which has only the links about 990 maybe that is too much. Let us start with 950 and I can assign something to each column.

(Refer Slide Time: 08:29)



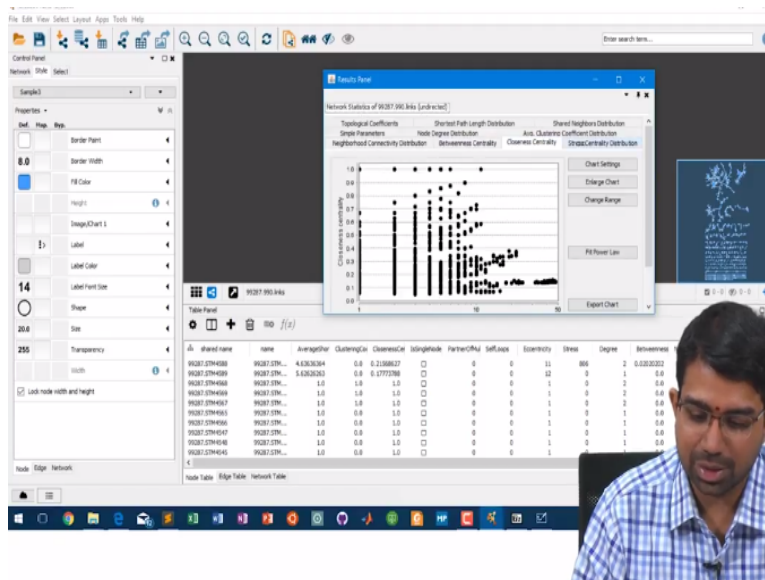
So this I can say this is a source node the first column, the second node is the second column is the target node and the third column is basically an age attributes. You may have it to be disconnected. This is you how it looks. You can also lay it out for you and it usually has nice layouts. So you see there are so many components clearly. So let look at this style, may be a there are different layout styles you can have.

I think you cannot even see it is already not it is too many nodes so but if you start zooming in this is what the network looks like. So you can see like there are so many separate components and so on and let us may be look at 990. **“Professor - student conversation starts”** Even if we

look at 990 the number of nodes are going to be the same. No the orphan nodes will be removed.
“Professor - student conversation ends”.

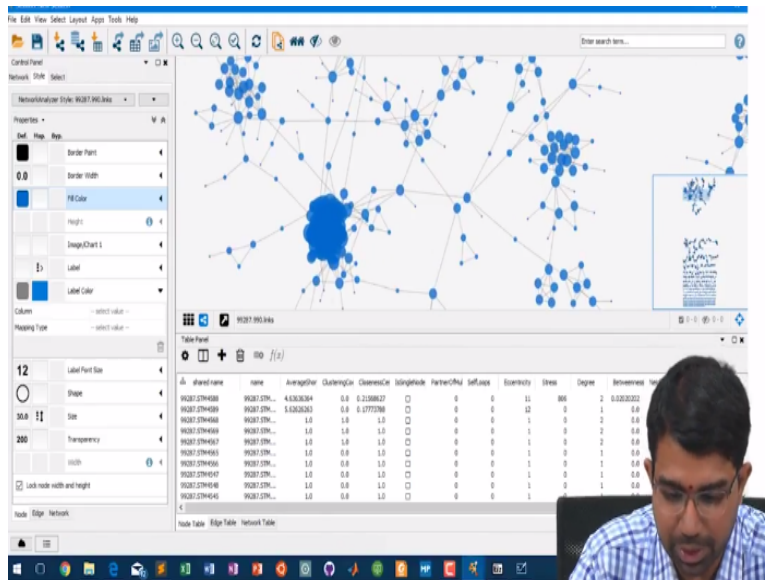
See these are orphaned in clusters and the small clusters with if you zoom it will show also the different features that you have and so on like the any node features so you can show node labels and whatever else that you want. The most useful aspect is there is a network analyzer just say analyze network treated as under directed in this case and just computes a dozen properties.

(Refer Slide Time: 11:35)



You can look at the node degree distribution looks reasonably power law, clustering coefficient distribution, number of shared neighbors, shortest path lengths, topological coefficients, neighborhood connectivity between a centrality, closeness centrality, other measures like stress centrality whatever and so on. The other useful thing is you can just go and say directly visualize parameters and map node size to degree and to say apply zoom can sometimes play up a little you can see some nodes that are larger here.

(Refer Slide Time: 14:14)



Helix has a tight cluster here. You have lot of biggish nodes here. Again when we just move out a single node or you want to just delete this node you can just delete it. If you select the first neighbors and so this is how you go about studying perturbations. I just removed a node in all its edges and now you can recalculate. You can recalculate properties and so on, but obviously you can image it is going to be a lot harder doing it with GUI.

And so on which is why I try to focus on mat lab BGL earlier, but this is another nice way. If you have a smallish network and if want to visually try to understand a few things is a great way to go about it. You can directly visualize between a centrality or closeness centrality or any of those measures and see how it affects. **“Professor - student conversation starts”** (()) (15:39) Built in app for perturbation or something that might be something, I am not aware of it.

There are so many different things. **“Professor - student conversation ends”** Something like motive discovery. So you have nice things actually. So you can even decide how you are going to visualize that. What I showed you here was I did some auto visualization using the network analyzer tool box itself plug in. So what you can actually do is you can pick what you want to visualize so you can just say let us look at film colour. So this is the default value.

This is the mapping and if it bypasses to something let us see what that means. So here I can say column fill colour, I can say clustering coefficient and obviously I cannot have a pass through

mapping here. If you have an integer if color were an integer if size I can have a pass through mapping to degree or something like that, higher the degree, higher the size. So here I basically make some sort of either a discrete or I can make a continuous mapping.

So see what happens to the figures. **“Professor - student conversation starts”** Why cannot size be associated to clustering coefficient? No, size can be associated to clustering. See I am trying to colour so here node colour I cannot map to some integer. So it has to be map through in this continuous fashion. So I cannot do a pass through mapping. There are 3 types of mapping you have in cytoscape.

There is discrete mapping, a continuous mapping, and a pass through mapping. Pass through mapping is useful when you are trying to say node title I mean what is the label of a node you pass through map in the label of the node to the node ID or something like that. Degree again you can pass through map it. So the size is very small. Clustering coefficient is a float.

So you do not want to have say non numbers are going to be between 0 and 1 why do you want to make a pass through mapping of it. You make a discrete mapping. You will say 0.3. I will just show you how that works, but you can now see this you can maybe you should have a reverse mapping. **“Professor - student conversation ends”**

(Refer Slide Time: 18:56)

The screenshot displays the Cytoscape software interface. The main window shows a network graph with nodes and edges. The left sidebar contains various toolbars and panels, including 'Properties', 'Columns', 'Mapping Type', and 'Label Part Size'. The bottom panel shows a 'Table Panel' with a table of node data.

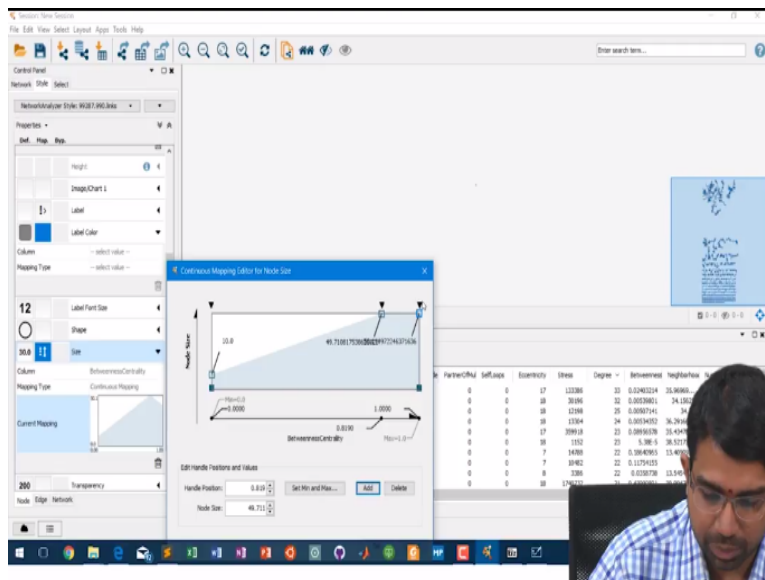
#	strand name	name	AveragePath	ClusteringC	ClosenessC	IsingNode	FactorOfN	SelfLoops	Essentiality	Stress	Degree	Betweenness
1	PK387.STRM1152	PK387.STRM...	6.8372947	0.7210891	0.1404837	<input type="checkbox"/>	0	0	17	13386	13	0.2462214
2	PK387.STRM1157	PK387.STRM...	3.3882884	0.6849461	0.2302842	<input type="checkbox"/>	0	0	18	26166	12	0.2022869
3	PK387.STRM1213	PK387.STRM...	3.2637842	0.7533133	0.1373385	<input type="checkbox"/>	0	0	18	2188	25	0.0087767
4	PK387.STRM1136	PK387.STRM...	3.2637842	0.7627128	0.1373385	<input type="checkbox"/>	0	0	18	1334	24	0.0034133
5	PK387.STRM1412	PK387.STRM...	6.1481851	0.6381284	0.2882919	<input type="checkbox"/>	0	0	17	23918	13	0.2884539
6	PK387.STRM1204	PK387.STRM...	3.2637842	0.5288137	0.1373385	<input type="checkbox"/>	0	0	18	1152	23	5.38E-5
7	PK387.STRM1217	PK387.STRM...	2.8167638	0.4353842	0.3750861	<input type="checkbox"/>	0	0	7	1400	22	0.3849163
8	PK387.STRM1216	PK387.STRM...	2.7401925	0.4391242	0.3637359	<input type="checkbox"/>	0	0	7	1842	22	0.1172184
9	PK387.STRM1268	PK387.STRM...	2.3008051	0.4545416	0.34375	<input type="checkbox"/>	0	0	8	336	22	0.0328738
10	PK387.STRM1242	PK387.STRM...	6.5758228	0.4652261	0.1521947	<input type="checkbox"/>	0	0	18	17942	11	0.2884539

So lighter the node higher the clustering coefficients that is the mapping that is going out currently. So you see all these dark nodes are basically 0 clustering coefficient. They do not have any triangles incident on them. So let us look at node size. So node size a continuous mapping with degree. You can also have a discrete mapping if you want which is actually painful. You can say degree value of 1 I will map it all size. Degree value of 2. I will map it to some size.

If you have like this 5 suppose you have lethal nodes, non lethal nodes, sick nodes some 3 or 4 things like that access via node attribute you can map them to some 3 or 4 colors of your liking. So, all these things are so trivially easily done with cytoscape. **“Professor - student conversation starts”** Pass through is just pass through. So I can just pass through is just identity. So now I can do pass through mapping for so the size is 1 if the degree is 1.

The size is 30 if the degree is 30 pass through. So 3 types of mapping, pass through, continuous, and discrete. You will rarely use pass through mapping almost except for something like default it will be there if you look at node label or something as a pass through mapping for name you can see that here. So it says that there is a pass through mapping for column name. **“Professor - student conversation ends”**

(Refer Slide Time: 20:42)



You can also have other very interesting ways to do things here so if your values are not if your values are distributed in a particular way let us see so let us say we are going to look at between

a centrality and I am going to make a discrete or continuous mapping and this I can scale it the way I want. So minimum between a centrality is 0 where in I am going to have a size of 30 but after this I just say after a point I do not care I am just going to leave the same size.

I want to really show the difference in size from this point to the next. So anything beyond a particular cut off I am not going to make the size much different. Are you able to understand this? I can have more points if I want. I did not want to do something like this, but something like this. This is almost like you know bring it to discrete mappings. So how does this look?

(Refer Slide Time: 22:39)

Recap

Topics covered

- ▶ Cytoscape Introduction
- ▶ STRING Database
- ▶ Loading and Visualising Networks
- ▶ NetworkAnalyzer

In the next video ...

- ▶ Simple Network Walkthrough
- ▶ Other Tools

I hope you got a quick introduction to cytoscape in this video and also the very interesting database known as string and I hope you now know how to load and visualize networks using cytoscape and how you can visualize many of the network parameters which are computed using network analyzer such as so you want to size a node based on its degree and so on all those things are easily doable in cytoscape.

In the next video take a simple network and have a walkthrough of the different kinds of analysis one can do and I have also mentioned to you briefly about other tools that exist.