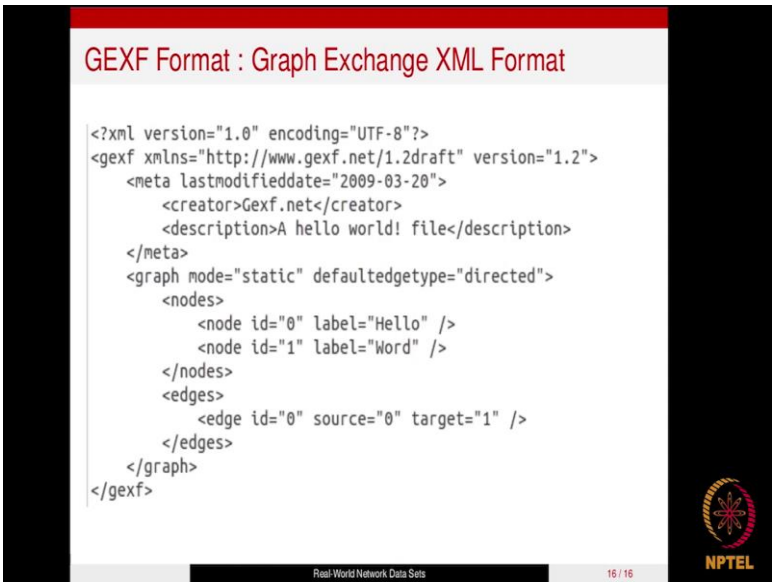
Social Networks
Prof. S. R. S. Iyengar
Department of Computer Science
Indian Institute of Technology, Ropar

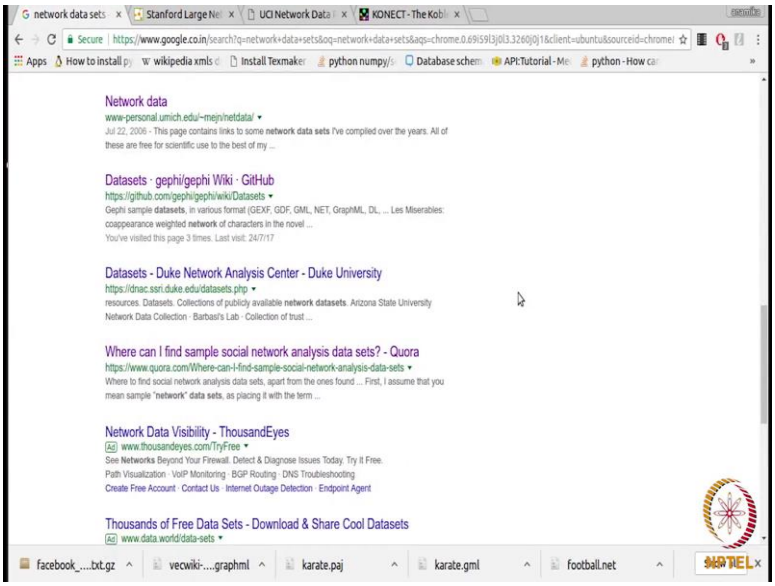
Lecture – 20
Handling Real-world Network Datasets
Datasets: How to Download?

(Refer Slide Time: 00:05)



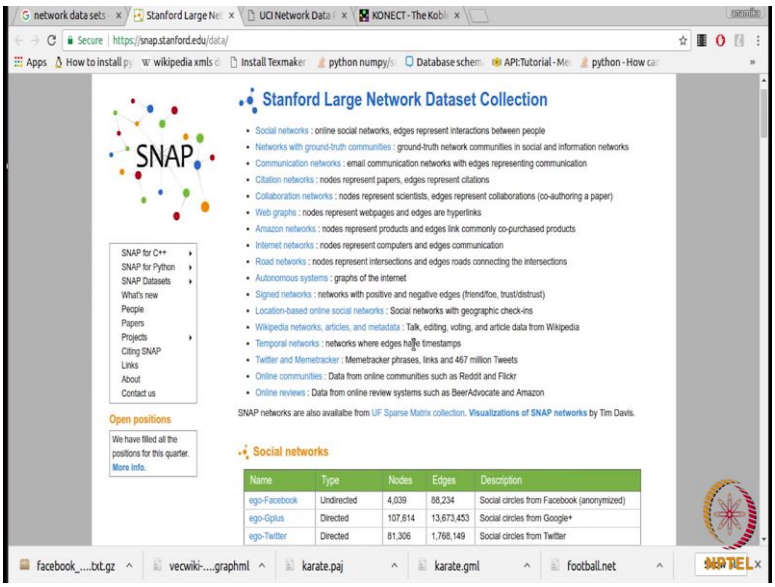
There is one more point to note here, networkx provides various functions through which we can read a network in one format and we can write the network in another format for example, we can read a network in Gmail format and we can write the network in GEXF format. So, those functions available and we can make use of them now let see how we can download these networks in different formats.

(Refer Slide Time: 00:34)



So, when we google for network dataset let us see which options do we get. So, here we get repository wise Stanford university which is called snap dataset let me open this and you also get a repository by UCI that is maintained by university of California and we also get this connect repository which is maintained by Koblenz university as you go down you get a number of more resources.

(Refer Slide Time: 01:05)

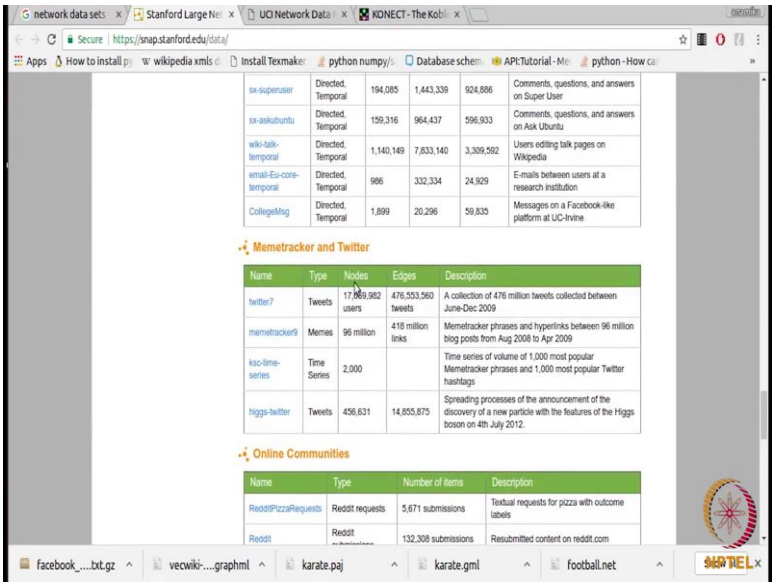


Let me show you this one first. So, snap dataset repository is the most commonly used repository for accessing the network data sets. So, here you get a number of networks of

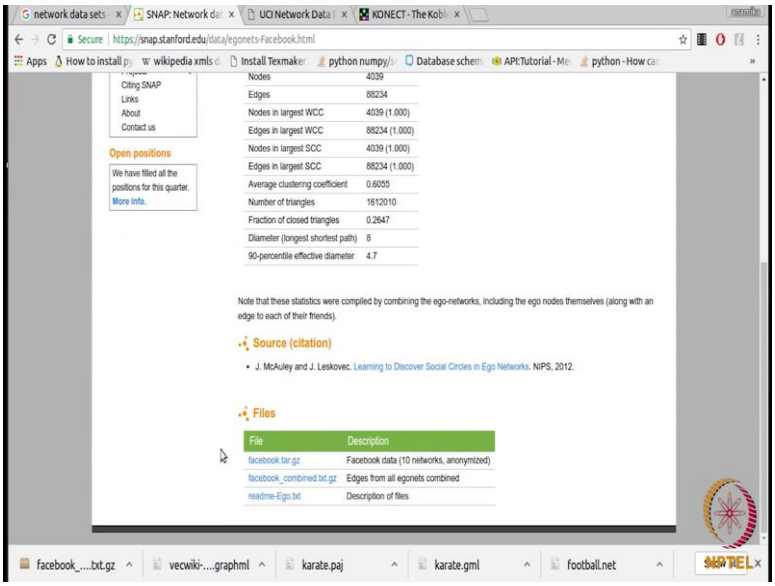
different types for example, you get social network like you like the networks on Facebook, Twitter, Slashdot, Google plus and you also get networks which have ground truth community what do you mean by that you know there are various communities of nodes in the networks. So, if there is a community detection algorithm you will come across these algorithms in the next videos. So, if you want to check whether a community detection algorithm is working fine or not, we might need to verify it.

So, these are the networks where the community details are given as ground truth values. So, they are useful in such scenarios then we have communication networks for example, networks e for example, email networks and we have citation networks and collaboration networks we already discussed about these networks and then we have web graphs we have Amazon networks, we have road networks and we have Wikipedia networks as in the networks between the articles and we have temporal networks now temporal networks are the ones which also contain the temporal information of the events. So, timing details will also be there, and we also have networks on stack over flow here. So, you see there are so many networks available over here and then the basic details you can see the number of nodes number of edges etcetera.

(Refer Slide Time: 02:45)

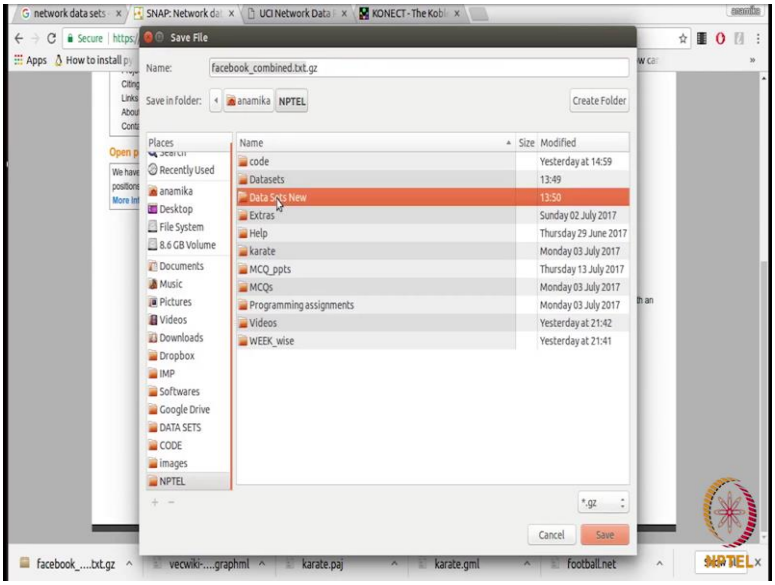


(Refer Slide Time: 02:52)



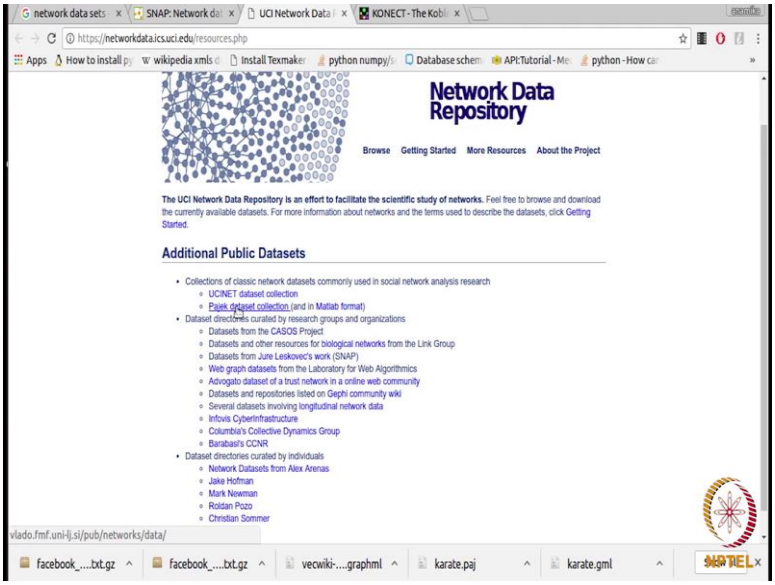
Now, let me show you how we can download one of the networks our; of these network. So, let me let me first open this and see the details. So, here you can read the details about the network and then you have basic statistics about the network and as you go down you have different files regarding the network. So, here we see 3 files one is readme file and the first file is having ten networks, the second file is having edges from all the networks you know combined. So, let me download this second file.

(Refer Slide Time: 03:31)



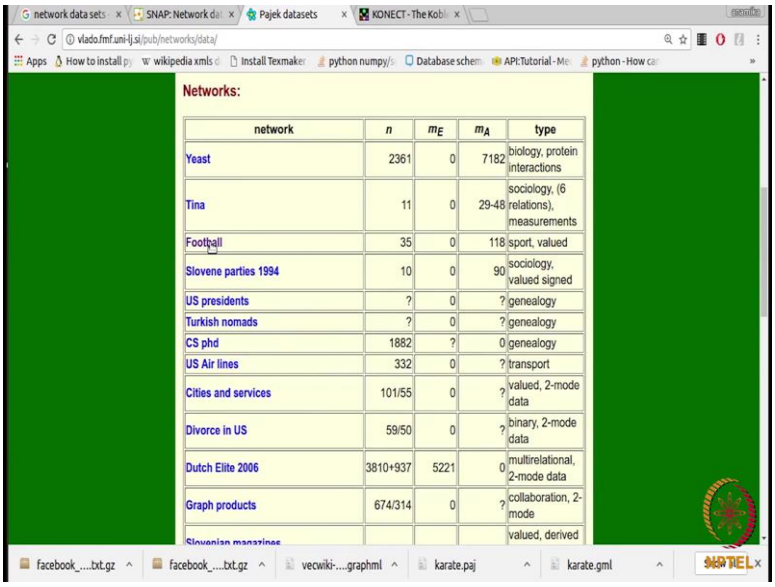
So, let me save this file and we will open it later to see its structure. So, this is how you can download the networks.

(Refer Slide Time: 03:43)



Let me show another repository this is maintained by university of California and they are also you see their in a given additional public datasets you can you can explore all these lines and get different kinds of network since we also have to download Pajek network and click this link. So, here we have all the Pajek datasets.

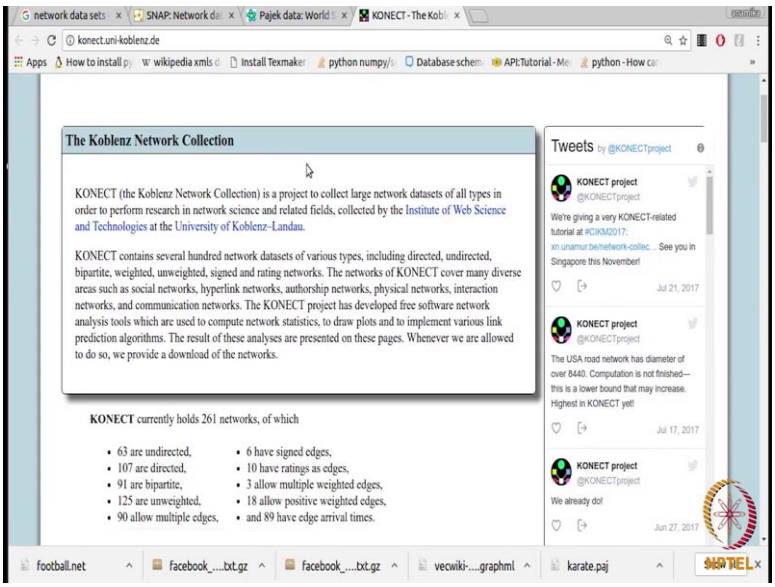
(Refer Slide Time: 04:02)



network	n	m_E	m_A	type
Yeast	2361	0	7182	biology, protein interactions
Tina	11	0	29-48	sociology, (6 relations), measurements
Football	35	0	118	sport, valued
Slovene parties 1994	10	0	90	sociology, valued signed
US presidents	?	0	?	genealogy
Turkish nomads	?	0	?	genealogy
CS phd	1882	?	0	genealogy
US Air lines	332	0	?	transport
Cities and services	101/55	0	?	valued, 2-mode data
Divorce in US	59/50	0	?	binary, 2-mode data
Dutch Elite 2006	3810+937	5221	0	multirelational, 2-mode data
Graph products	674/314	0	?	collaboration, 2-mode
Slovenian magazines				valued, derived

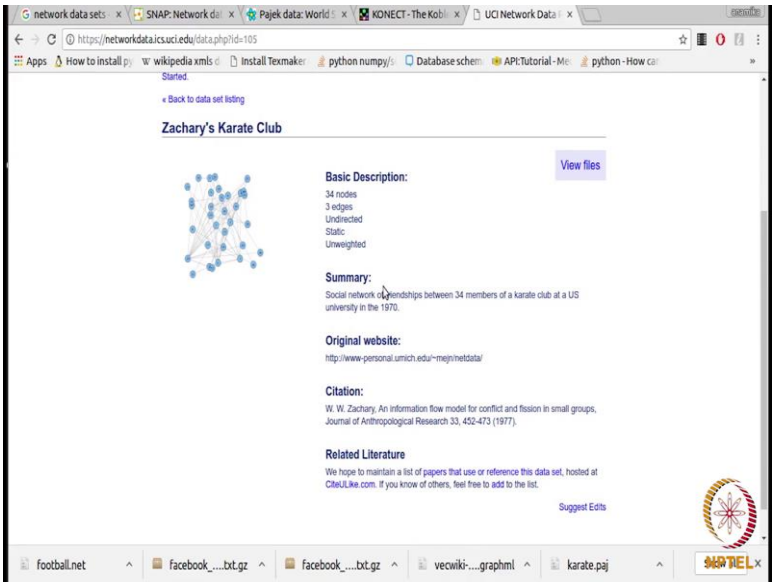
Let me download the small network just to show you the kind of structure that Pajek file contains. So, I am saving this I am sorry let me first see this. So, we have to save this file. So, this is a dot net file I told you previously that the Pajek files will be available in one of the 2 formats either it will be dot net, or it will be dot paj. So, this one is dot net a very small network just to show you the structure.

(Refer Slide Time: 04:43)



Let us go to the next one. So, here you see this repository again is having a lot of networks on different topics you see Wikipedia site you like and there are air traffic control there are different topics and you can just choose based on your requirement there is twitter dataset. So, you can download the data set based on the resource that you want to access or you can download dataset based on the format that you want for example, I want the data from say UCI.

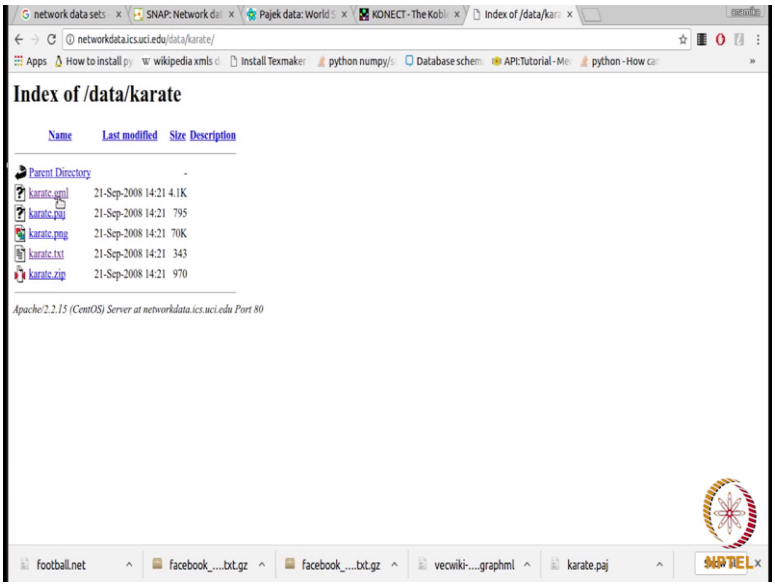
(Refer Slide Time: 05:24)



So, I click this and here you have all the network from this repository there is an interesting network here which is the Zachary Karate, let me download this. So, there is an interesting history attached to this network there was a karate club in us university and it had 34 participants, there was a fight that happened between 2 important people of that club and they were the instructor and the administrator.

So, after a period of 3 year the due to this fight the network got divided into 2 communities Zachary was a person who analyze this network over a over this period and he Indian basically predicted the communities that are going to form. So, this network is well known and it is called the Zachary's karate club network and it is very small as well its use for the community detection algorithms as well and it is also a nice starting step for you to start your analysis on the networks. So, we are going to download this network.

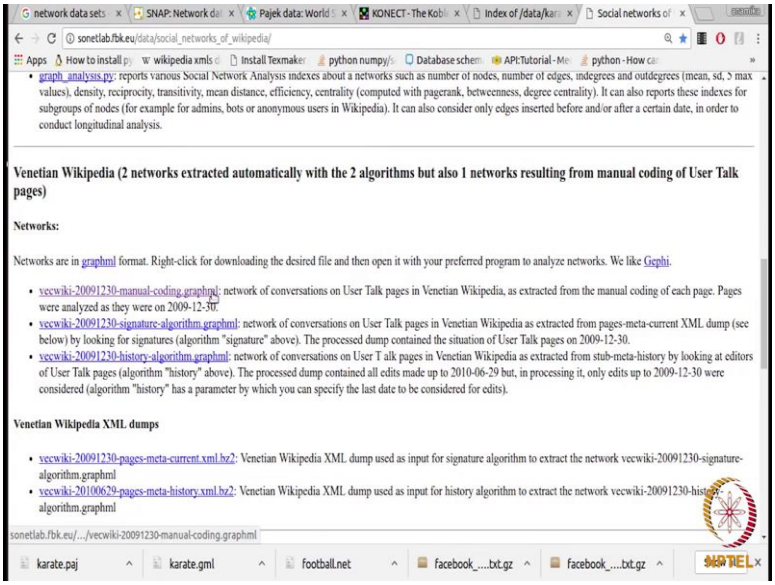
(Refer Slide Time: 06:35)



This network is available in GML format as well as Pajek format. So, let me download both of them and then we have this Pajek. So, we have done with Ajelius format and dot net format and dot gml format and dot Pajek format.

So, we are left with GraphML. So, let me download one network in that format I know one resource a sonnet is a repository that contains data in graph ml format.

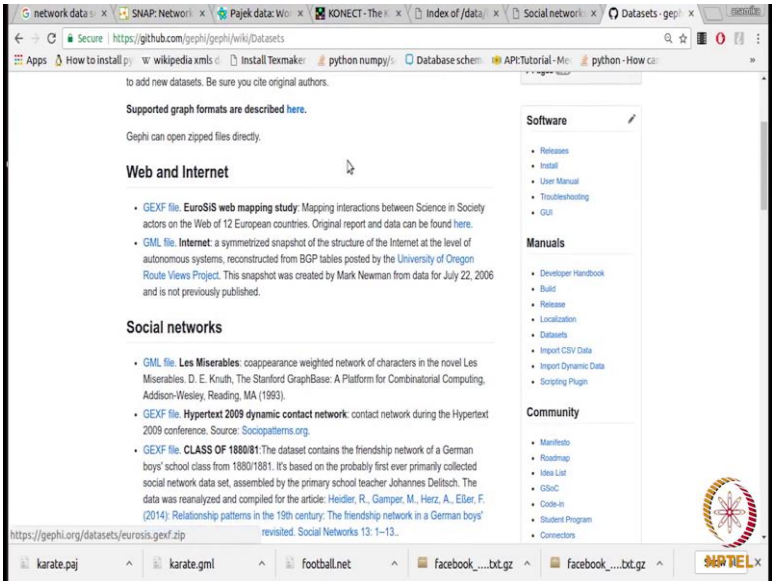
(Refer Slide Time: 07:12)



So, let me just open that only here these are the Wikipedia network which are in graph ml format I had already downloaded this once. So, I will access that only. So, you can

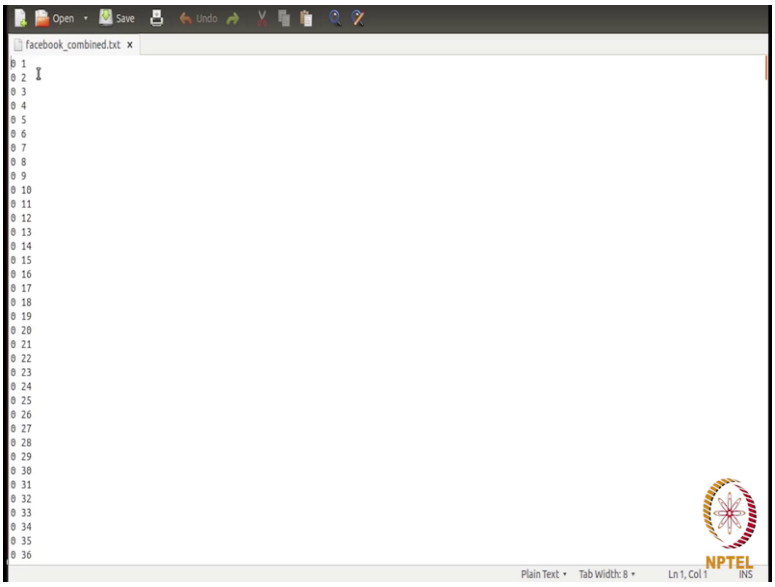
download all any of these networks and you can basically read the details about the network and then accordingly download if it chooses your purpose.

(Refer Slide Time: 07:47)



Now, let see GEXF format let me show you how we can download that let me open this link. So, here you can see various networks in GML as well as GEXF format. So, you can download any of these after reading the details I have already downloaded one in GEXF files. So, I will show you that only.

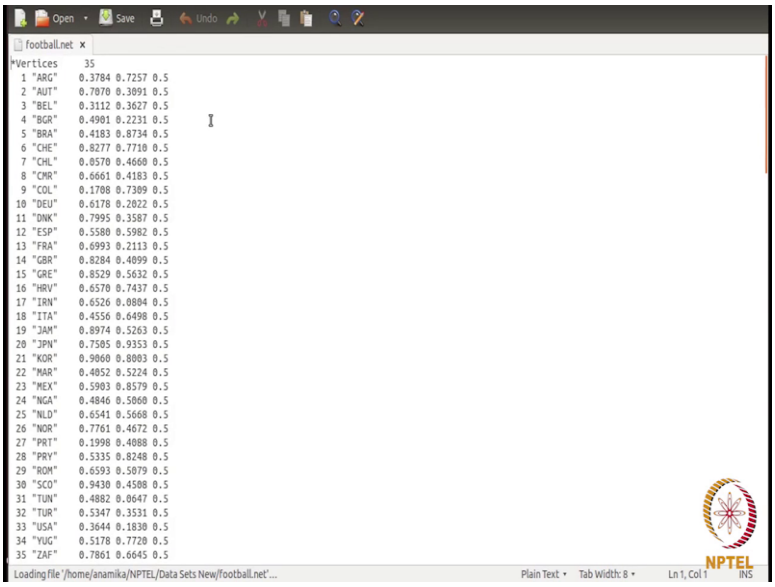
(Refer Slide Time: 08:05)



So, let us look at what we have downloaded we have all these networks let me extract this first. So, here we have six network GEXF format txt format dot net format dot gml format dot Pajek format and dot graph ml format. So, I am quickly going to show you the structure of these files although I had already introduce the formats to you let me first open the txt format which is in edge list format.

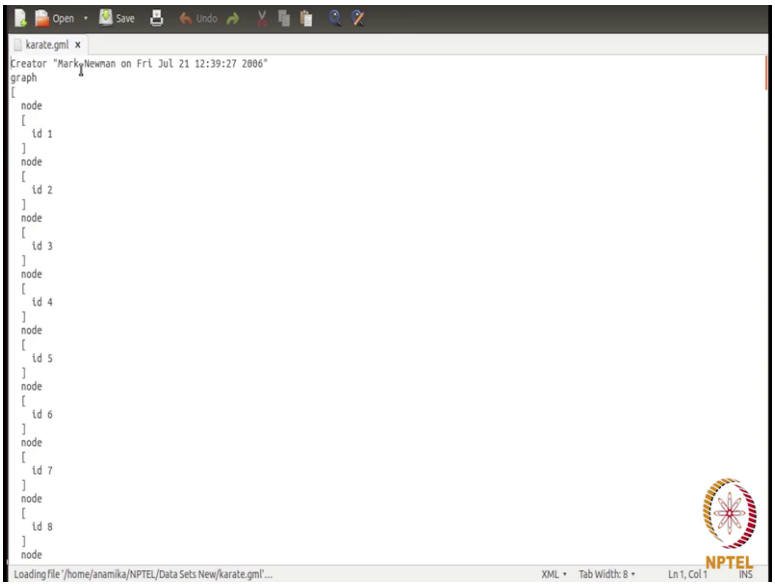
So, you can see this simplicity of the format again you just have 2 things in every row and these 2 things are the source and the target of the edges. So, there will be a link from 0 to 1, 0 to 2, basically this is undirected. So, there will be an edge between 0 and 3, 0 and 4 and so on. So, this is the edge list format.

(Refer Slide Time: 08:59)



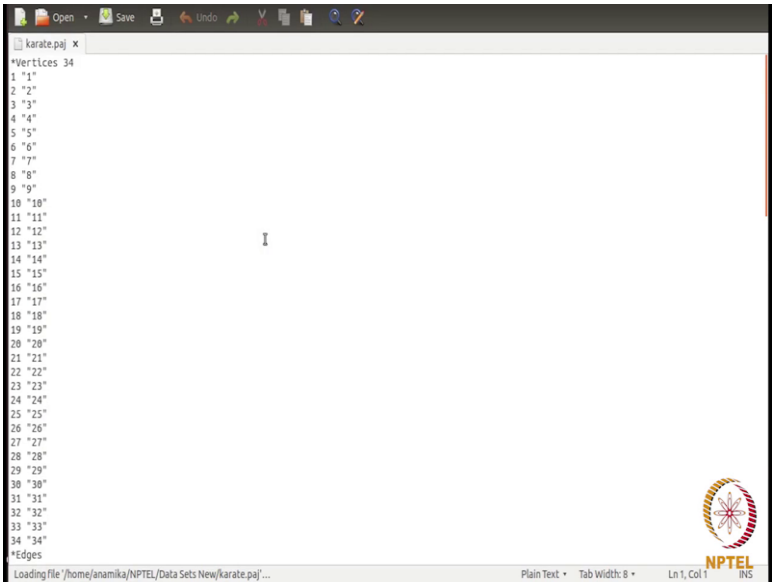
Next let us check the dot net format. So, as I told you this starts with the key words star vertices and then you have the number of vertices in that network. So, we these are the ids and these are the labels of the vertices and then we have 3 details attached to every vertex which are the attributes and you can basically see the documentation to see what these attributes mean then after the nodes are done you have these arcs which are basically the edges and for every edge you have this attributes attached.

(Refer Slide Time: 09:40)



So, this is an example of dot net file and then we have GML. So, you see there is a graph keyboard and then you have square brackets and then you have all the nodes. So, first all the nodes will be there and then once the nodes are done you have these edge details. So, this is a Gmail format. So, this is Zachary karate network which had thirty-four nodes.

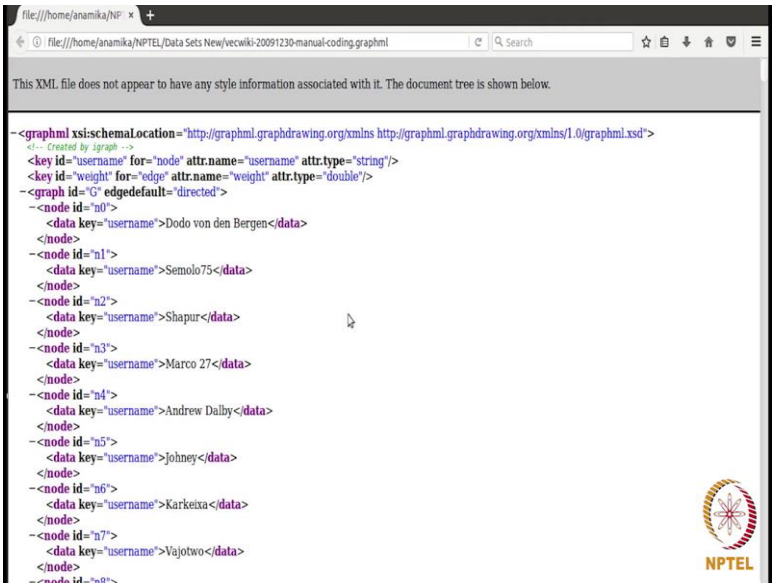
(Refer Slide Time: 10:05)



Now, this is a karate network in Pajek format, let me show you that here there is no attribute for the vertices and there is no attribute for the edges as well. So, this is again a very simple network in Pajek. Now let me show you graph ml format. So, this is an

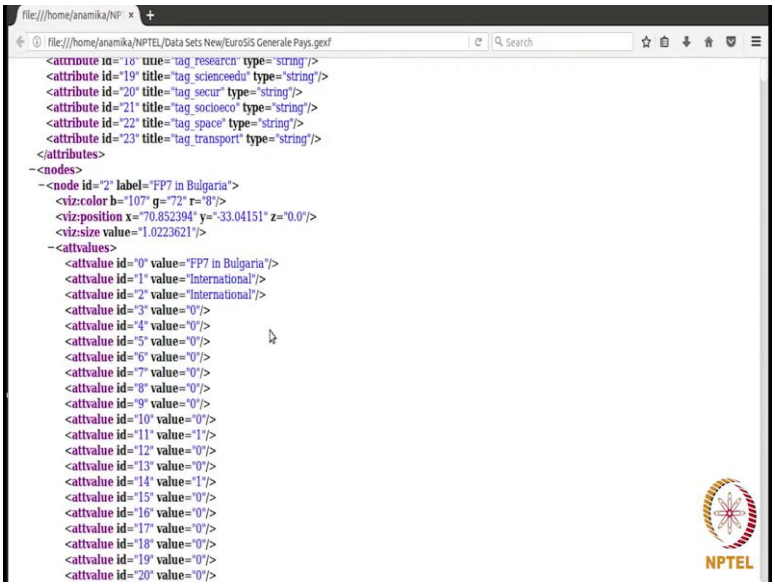
example of a GraphML network. So, it as the number of tags the first tag is GraphML xl and after that we have 2 key tags. I told you that key tags for adding the attributes to nodes and edges.

(Refer Slide Time: 10:26)



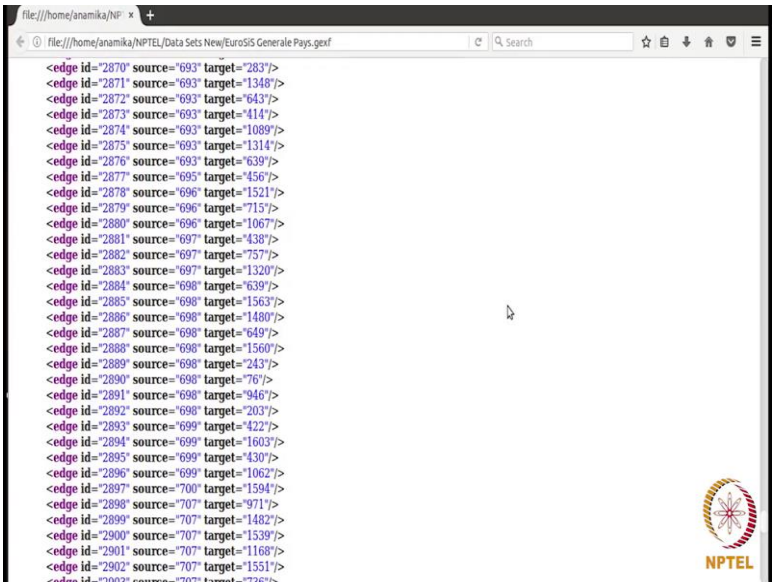
So, first one is for nodes and the second one is for edges and then we start the graph tag and inside graph we have a number of node tags and the nodes are given attributes the key using the you know data tag and we are making use of this key and as you go down after the nodes are done you can go down gap after the nodes are done you have the edge stacks. So, again the edges are given attributes using the data tag.

(Refer Slide Time: 11:21)



Now, let us check the GEXF format. So, here you see there are again because this is also based on xml there is GEXF tag inside that we have graph tag and there are. So, many attributes for this graph and after that we have nodes tag inside these nodes there will be all the nodes. So, the first node is here and then there are. So, many attributes for this node and then second node comes with all these attributes. So, basically here in this network they are assigning a lot of attributes for every node as you can see. So, as you go down once the nodes are done the edges will be there it should be the yeah. So, here you see once the nodes are done, we have edge tags they have not assigned any attributes for the edges as you can see. So, this is an example of GEXF format.

(Refer Slide Time: 12:01)



So, you saw there are so many resources available, you can just explore and download the once which suit your purpose for an analysis. So, this is the basic introduction to how we can download datasets. Next, we will see how we can analysis these network datasets that we have downloaded.