


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)

GEXF Format : Graph Exchange XML Format

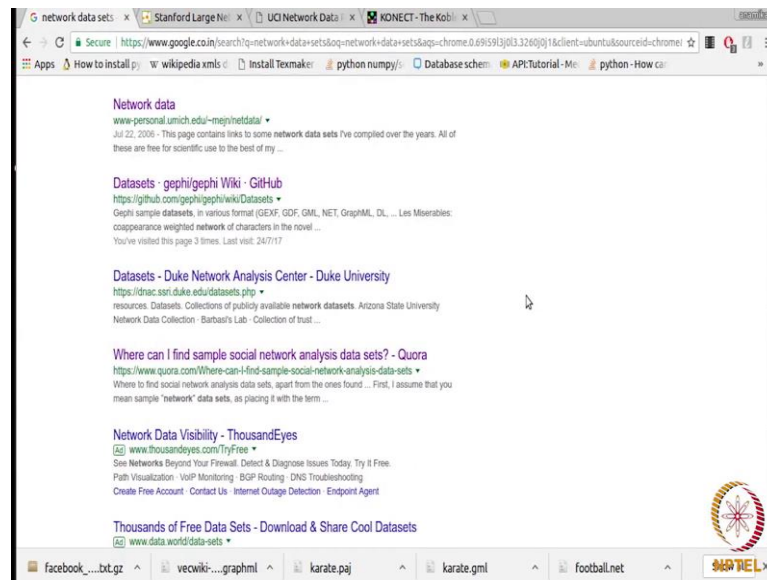
```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
  <meta lastmodifieddate="2009-03-20">
    <creator>Gexf.net</creator>
    <description>A hello world! file</description>
  </meta>
  <graph mode="static" defaultedgetype="directed">
    <nodes>
      <node id="0" label="Hello" />
      <node id="1" label="Word" />
    </nodes>
    <edges>
      <edge id="0" source="0" target="1" />
    </edges>
  </graph>
</gexf>
```


NPTEL

Real-World Network Data Sets 16 / 16

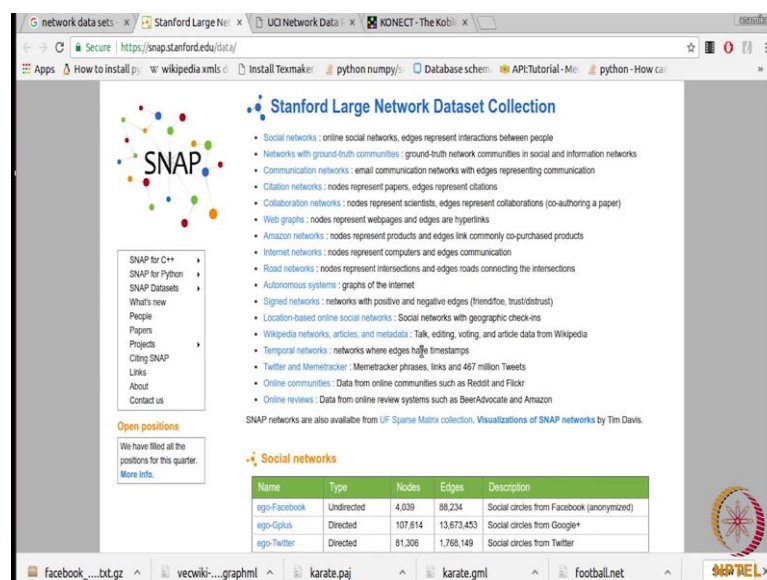
There is one more point to note here, networkx provides various functions through which we can read a network in one format and we can write the network in another format for example, we can read a network in Gmail format and we can write the network in GEXF format. So, those functions available and 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

(Refer Slide Time: 02:52)

The screenshot shows the SNAP website with the following statistics for the Facebook ego network dataset:

Nodes	Edges
4039	88234
Nodes in largest WCC	4039 (1,000)
Edges in largest WCC	88234 (1,000)
Nodes in largest SCC	4039 (1,000)
Edges in largest SCC	88234 (1,000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2847
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

Note that these statistics were compiled by combining the ego-networks, including the ego nodes themselves (along with an edge to each of their friends).

Source (citation)

- J. McKeuley and J. Leskovec, Learning to Discover Social Circles in Ego Networks, NIPS, 2012.

Files

File	Description
facebook.tar.gz	Facebook data (10 networks, anonymized)
facebook_combined.txt.gz	Edges from all egonets combined
readme-Ego.txt	Description of files

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)

The screenshot shows a 'Save File' dialog box with the filename 'facebook_combined.txt.gz' and the save location set to 'anamika NPTEL'. The file explorer on the left shows the following structure:

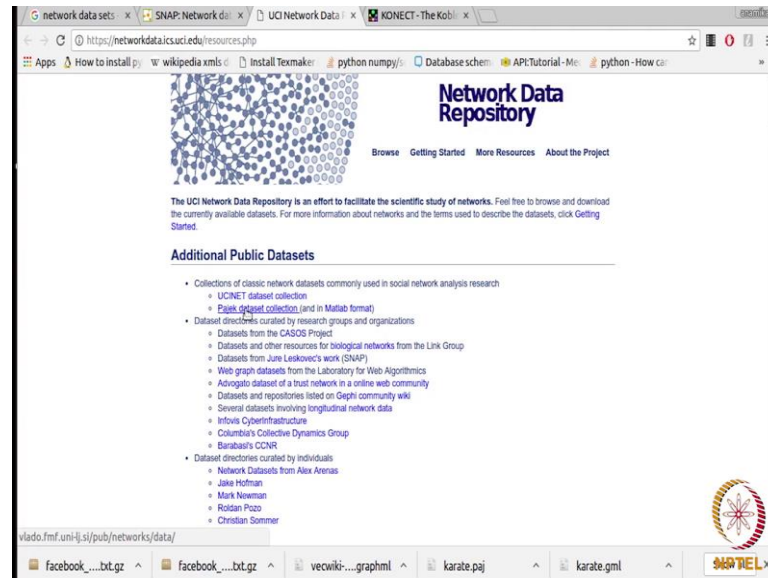
- Places
 - anamika
 - Desktop
 - File System
 - 8.6 GB Volume
 - Documents
 - Music
 - Pictures
 - Videos
 - Downloads
 - Dropbox
 - IMP
 - Softwares
 - Google Drive
 - DATA SETS
 - CODE
 - images
 - NPTEL

The file explorer on the right shows the following files:

Name	Size	Modified
code		Yesterday at 14:59
Datasets		13:49
Data Sets New		13:50
Extras		Sunday 02 July 2017
Help		Thursday 29 June 2017
karate		Monday 03 July 2017
MCQ_ppts		Thursday 13 July 2017
MCQs		Monday 03 July 2017
Programming assignments		Monday 03 July 2017
Videos		Yesterday at 21:42
WEEK_wise		Yesterday at 21:41

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.

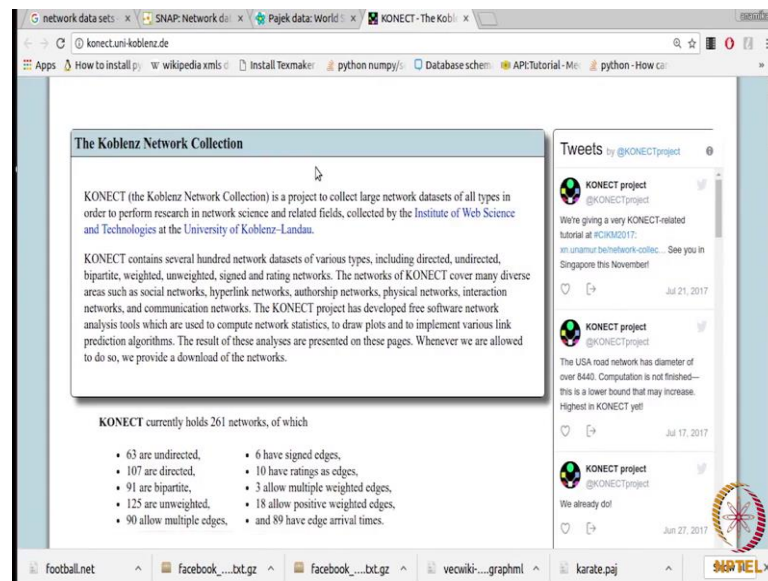
(Refer Slide Time: 04:02)

The screenshot shows the Pajek datasets website. The table lists various network datasets with columns for network name, n, mE, mA, and type. The datasets include Yeast, Tina, Football, Slovene parties 1994, US presidents, Turkish nomads, CS phd, US Air lines, Cities and services, Divorce in US, Dutch Elite 2006, Graph products, and Simian mansonae.

network	n	mE	mA	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
Simian mansonae				valued, derived

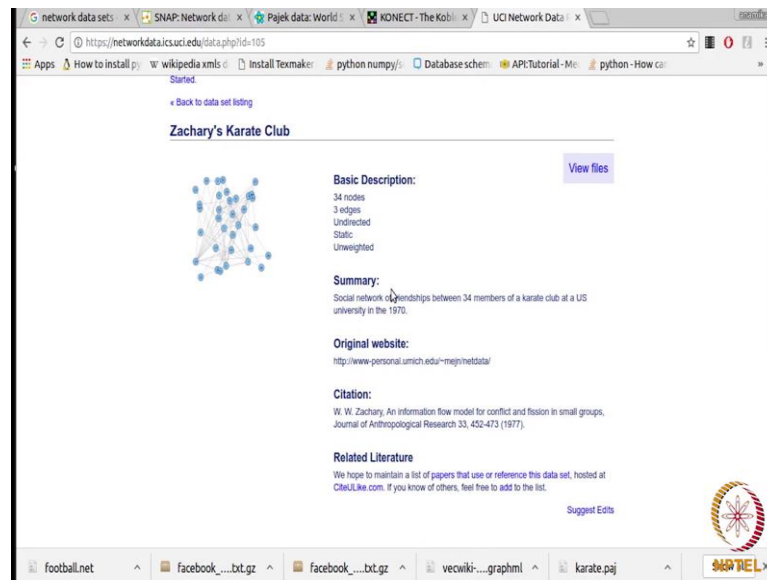
Let me download the small network just to show you the kind of structure that Pajek file contains. So, I am saving this I am sorry let me first see this. So, we have to save this file. So, this is a dot net file I told you previously that the Pajek files will be available in one of the 2 formats either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

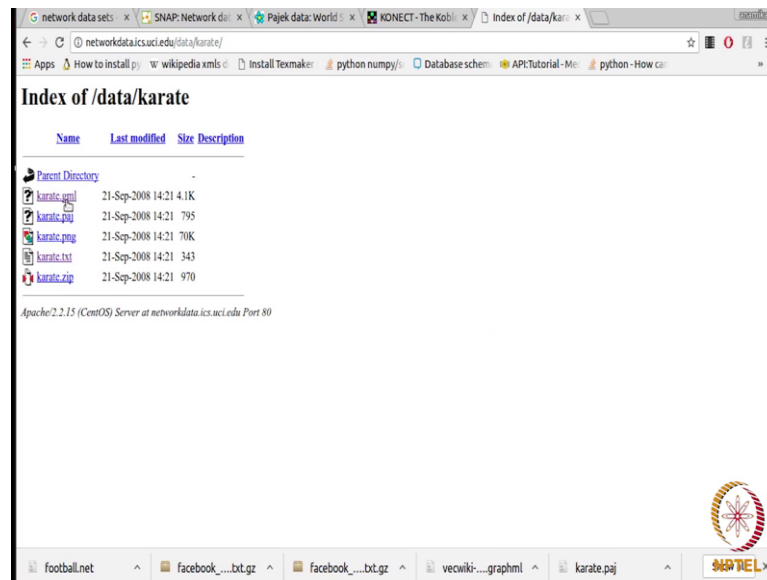
(Refer Slide Time: 05:24)



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.

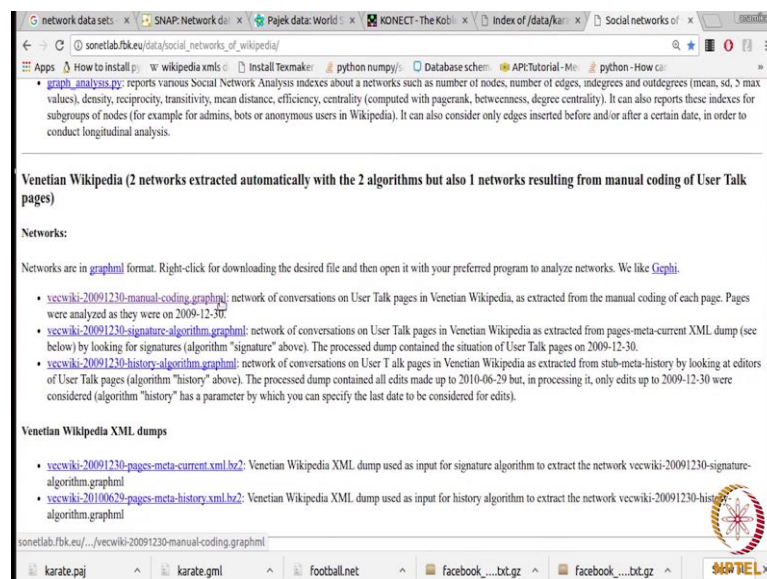
(Refer Slide Time: 06:35)



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.

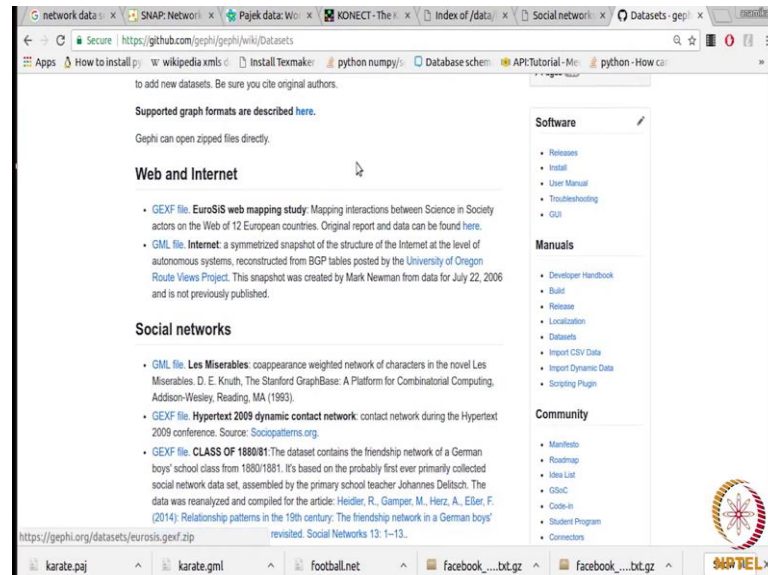
(Refer Slide Time: 07:12)



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

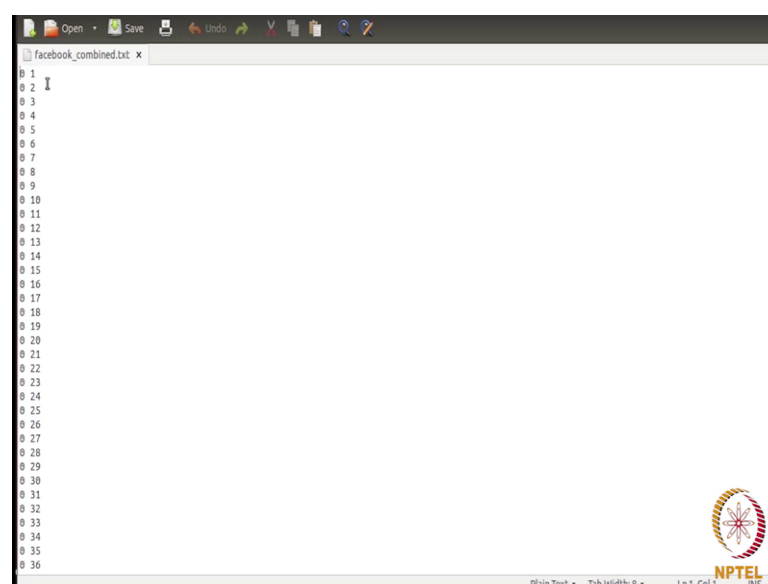
download all any of these networks and you can basically read the details about the network and then accordingly download if it chooses your purpose.

(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.

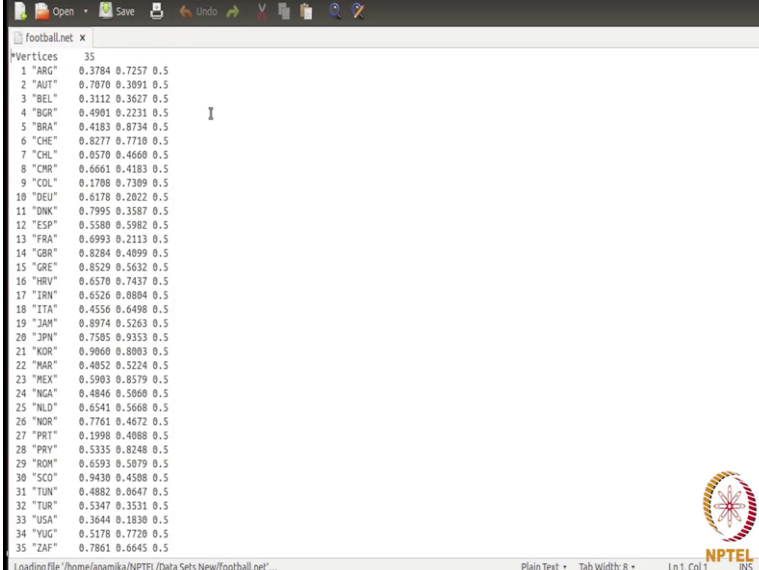
(Refer Slide Time: 08:05)



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.

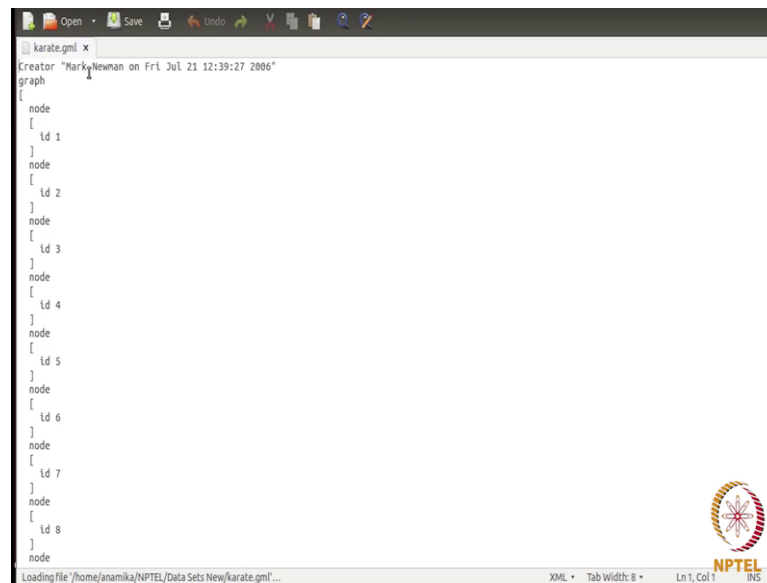
(Refer Slide Time: 08:59)



```
football.net x
Vertices 35
1 "ARG" 0.3784 0.7257 0.5
2 "AUT" 0.7070 0.3091 0.5
3 "BEL" 0.3112 0.3627 0.5
4 "BGR" 0.4901 0.2231 0.5
5 "BRA" 0.4183 0.8734 0.5
6 "CHE" 0.8277 0.7710 0.5
7 "CHL" 0.8570 0.4660 0.5
8 "CMR" 0.6661 0.4183 0.5
9 "COL" 0.1708 0.7309 0.5
10 "DEU" 0.6178 0.2022 0.5
11 "DNK" 0.7995 0.3587 0.5
12 "ESP" 0.5580 0.5982 0.5
13 "FRA" 0.6993 0.2113 0.5
14 "GBR" 0.8284 0.4099 0.5
15 "GRE" 0.8529 0.5632 0.5
16 "HRV" 0.6570 0.7437 0.5
17 "IRN" 0.6526 0.8004 0.5
18 "ITA" 0.4556 0.6498 0.5
19 "JAM" 0.8974 0.5263 0.5
20 "JPN" 0.7505 0.9353 0.5
21 "KOR" 0.9060 0.8003 0.5
22 "MAR" 0.4052 0.5224 0.5
23 "MEX" 0.5993 0.8579 0.5
24 "NGA" 0.4846 0.5060 0.5
25 "NLD" 0.6541 0.5668 0.5
26 "NOR" 0.7761 0.4672 0.5
27 "PRY" 0.1998 0.4088 0.5
28 "PRY" 0.5335 0.8248 0.5
29 "ROM" 0.6593 0.5079 0.5
30 "SCO" 0.9430 0.4508 0.5
31 "TUN" 0.4082 0.0647 0.5
32 "TUR" 0.5347 0.3531 0.5
33 "USA" 0.3644 0.1830 0.5
34 "YUG" 0.5178 0.7720 0.5
35 "ZAF" 0.7861 0.6645 0.5
```

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.

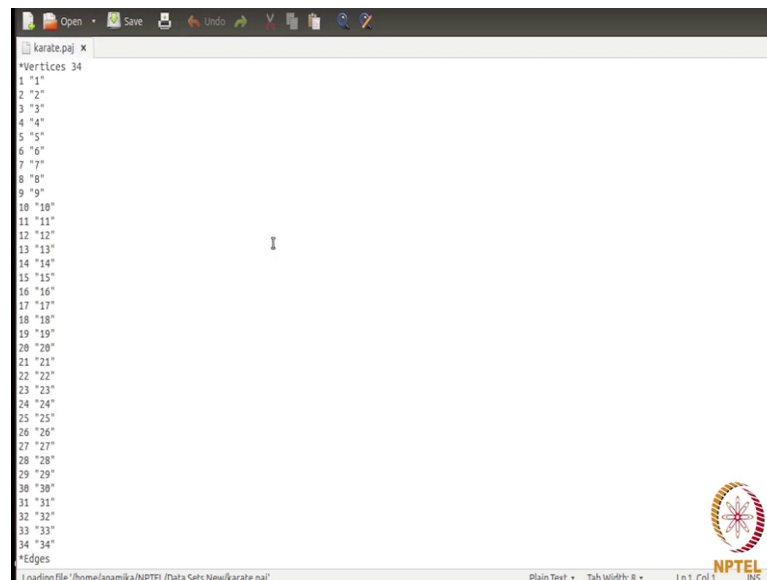
(Refer Slide Time: 09:40)



```
graph
[
  node
  [
    id 1
  ]
  node
  [
    id 2
  ]
  node
  [
    id 3
  ]
  node
  [
    id 4
  ]
  node
  [
    id 5
  ]
  node
  [
    id 6
  ]
  node
  [
    id 7
  ]
  node
  [
    id 8
  ]
]
```

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)

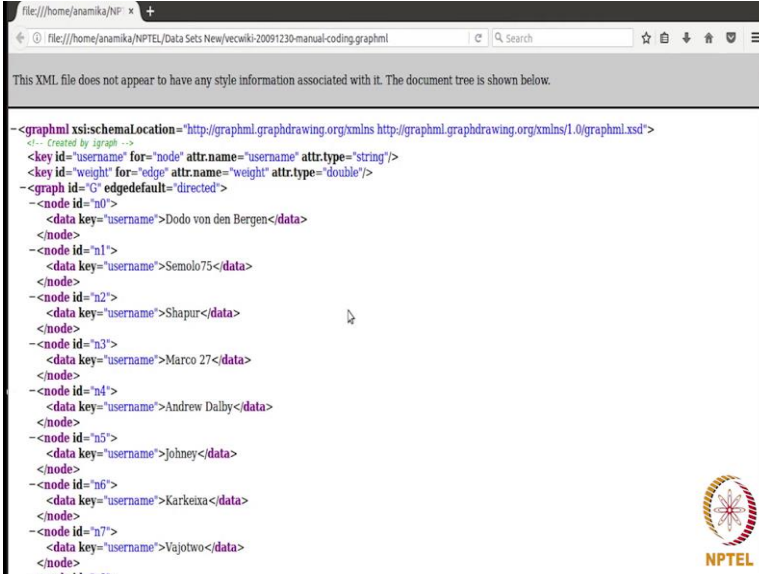


```
*Vertices 34
1 "1"
2 "2"
3 "3"
4 "4"
5 "5"
6 "6"
7 "7"
8 "8"
9 "9"
10 "10"
11 "11"
12 "12"
13 "13"
14 "14"
15 "15"
16 "16"
17 "17"
18 "18"
19 "19"
20 "20"
21 "21"
22 "22"
23 "23"
24 "24"
25 "25"
26 "26"
27 "27"
28 "28"
29 "29"
30 "30"
31 "31"
32 "32"
33 "33"
34 "34"
*Edges
```

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.

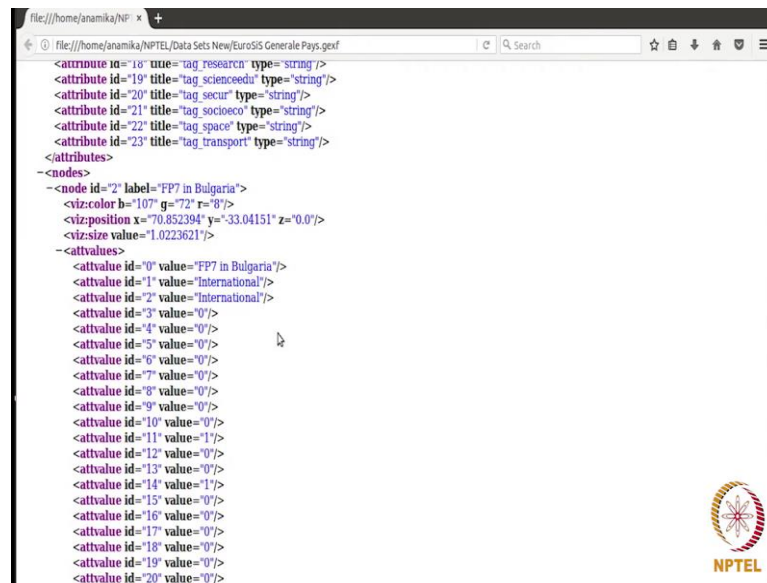
(Refer Slide Time: 10:26)

A screenshot of a web browser window. The address bar shows a file path: file:///home/ananika/NPTEL/Data Sets New/vecwiki-20091230-manual-coding.graphml. Below the address bar, a message states: "This XML file does not appear to have any style information associated with it. The document tree is shown below." The main content area displays an XML document. The root element is <graphml>, which includes an xsi:schemaLocation attribute pointing to a URL. Inside the <graphml> tag, there are two <key> tags: one for nodes with attribute 'username' and type 'string', and one for edges with attribute 'weight' and type 'double'. Following these keys is a <graph id='G' edgedefault='directed'> tag. Inside this graph tag, there are eight <node> elements, each with an id attribute (n0 through n7) and a <data key='username'> attribute containing a name and a weight. The names are: Dodo von den Bergen, Semolo75, Shapur, Marco, Andrew Dalby, Johney, Karkkna, and Vajotwo. The XML code is color-coded, with tags in green and attributes/values in black. An NPTEL logo is visible in the bottom right corner of the browser window.

```
<graphml xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <!-- Created by igraph -->
  <key id="username" for="node" attr.name="username" attr.type="string"/>
  <key id="weight" for="edge" attr.name="weight" attr.type="double"/>
  <graph id="G" edgedefault="directed">
    <node id="n0">
      <data key="username">Dodo von den Bergen</data>
    </node>
    <node id="n1">
      <data key="username">Semolo75</data>
    </node>
    <node id="n2">
      <data key="username">Shapur</data>
    </node>
    <node id="n3">
      <data key="username">Marco 27</data>
    </node>
    <node id="n4">
      <data key="username">Andrew Dalby</data>
    </node>
    <node id="n5">
      <data key="username">Johney</data>
    </node>
    <node id="n6">
      <data key="username">Karkkna</data>
    </node>
    <node id="n7">
      <data key="username">Vajotwo</data>
    </node>
  </graph>
</graphml>
```

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.

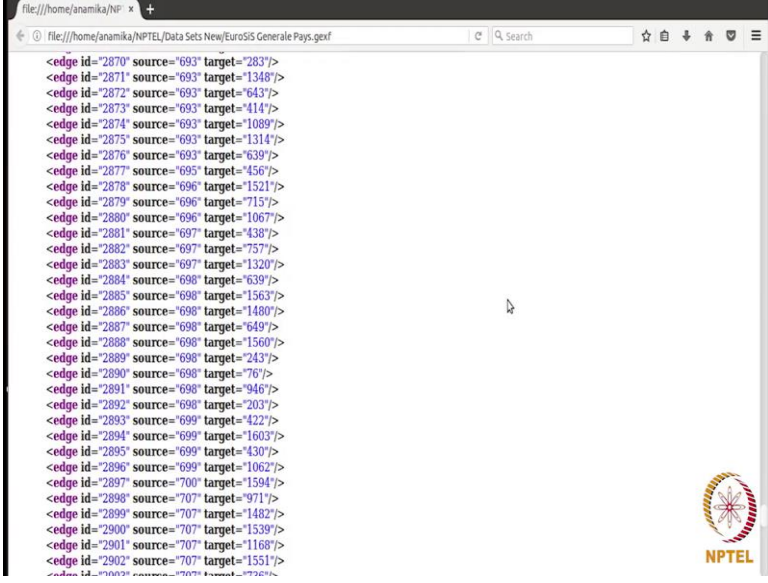
(Refer Slide Time: 11:21)



```
<!-- attribute id="18" title="tag_research" type="string"/>
<!-- attribute id="19" title="tag_sciencedu" type="string"/>
<!-- attribute id="20" title="tag_secur" type="string"/>
<!-- attribute id="21" title="tag_socioeco" type="string"/>
<!-- attribute id="22" title="tag_space" type="string"/>
<!-- attribute id="23" title="tag_transport" type="string"/>
</attributes>
<!-- nodes -->
<!-- node id="2" label="FP7 in Bulgaria" -->
<!-- viz:color b="107" g="72" r="8"/>
<!-- viz:position x="70.852394" y="-33.04151" z="0.0"/>
<!-- viz:size value="1.0223621"/>
<!-- attvalues -->
<!-- attvalue id="0" value="FP7 in Bulgaria"/>
<!-- attvalue id="1" value="International"/>
<!-- attvalue id="2" value="International"/>
<!-- attvalue id="3" value="0"/>
<!-- attvalue id="4" value="0"/>
<!-- attvalue id="5" value="0"/>
<!-- attvalue id="6" value="0"/>
<!-- attvalue id="7" value="0"/>
<!-- attvalue id="8" value="0"/>
<!-- attvalue id="9" value="0"/>
<!-- attvalue id="10" value="0"/>
<!-- attvalue id="11" value="1"/>
<!-- attvalue id="12" value="0"/>
<!-- attvalue id="13" value="0"/>
<!-- attvalue id="14" value="1"/>
<!-- attvalue id="15" value="0"/>
<!-- attvalue id="16" value="0"/>
<!-- attvalue id="17" value="0"/>
<!-- attvalue id="18" value="0"/>
<!-- attvalue id="19" value="0"/>
<!-- attvalue id="20" value="0"/>
```

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)



The screenshot shows a text editor window with a file path in the title bar: `file:///home/anamika/NPTEL/Data Sets New/EuroSIS Generale Pays.gexf`. The editor displays a list of network edges in GEXF format. Each line represents an edge with its ID, source node, and target node. The edges are listed sequentially from ID 2870 to 2902. The NPTEL logo is visible in the bottom right corner of the editor window.

```
<edge id="2870" source="693" target="283"/>
<edge id="2871" source="693" target="1348"/>
<edge id="2872" source="693" target="643"/>
<edge id="2873" source="693" target="414"/>
<edge id="2874" source="693" target="1089"/>
<edge id="2875" source="693" target="1314"/>
<edge id="2876" source="693" target="639"/>
<edge id="2877" source="695" target="456"/>
<edge id="2878" source="696" target="1521"/>
<edge id="2879" source="696" target="713"/>
<edge id="2880" source="696" target="1067"/>
<edge id="2881" source="697" target="438"/>
<edge id="2882" source="697" target="757"/>
<edge id="2883" source="697" target="1320"/>
<edge id="2884" source="698" target="639"/>
<edge id="2885" source="698" target="1563"/>
<edge id="2886" source="698" target="1480"/>
<edge id="2887" source="698" target="649"/>
<edge id="2888" source="698" target="1560"/>
<edge id="2889" source="698" target="243"/>
<edge id="2890" source="698" target="76"/>
<edge id="2891" source="698" target="946"/>
<edge id="2892" source="699" target="203"/>
<edge id="2893" source="699" target="422"/>
<edge id="2894" source="699" target="1603"/>
<edge id="2895" source="699" target="430"/>
<edge id="2896" source="699" target="1062"/>
<edge id="2897" source="700" target="1594"/>
<edge id="2898" source="707" target="971"/>
<edge id="2899" source="707" target="1482"/>
<edge id="2900" source="707" target="1539"/>
<edge id="2901" source="707" target="1168"/>
<edge id="2902" source="707" target="1551"/>
```

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