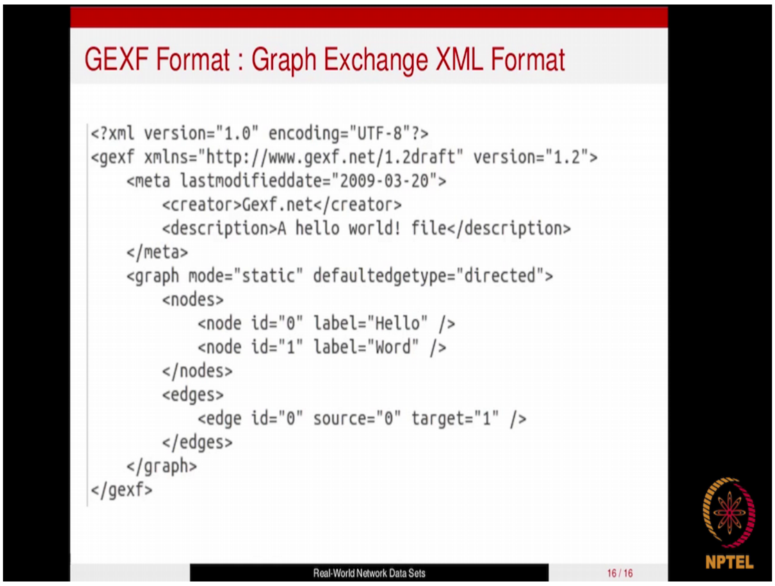
# Social Networks
## Prof. S. R. S. Iyengar
## Department of Computer Science
## Indian Institute of Technology, Ropar

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

(Refer Slide Time: 00:05)



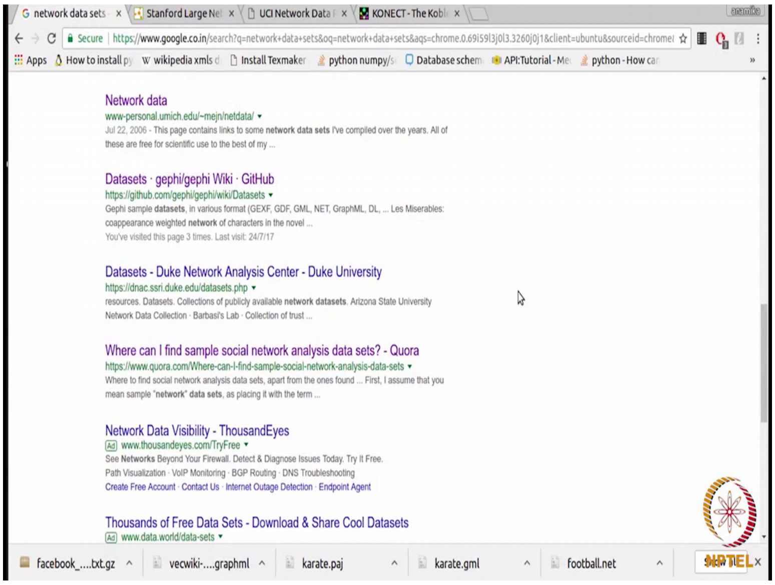
GEXF Format : Graph Exchange XML Format

```
<?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>
```

Real-World Network Data Sets 16 / 16

There is one more point to note here, network X 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)



Network data
www-personal.umich.edu/~mejn/netdata/
Jul 22, 2006 - This page contains links to some **network data sets** I've compiled over the years. All of these are free for scientific use to the best of my ...

Datasets · gephi/gephi Wiki · GitHub
https://github.com/gephi/gephi/wiki/Datasets
Gephi sample **datasets**, in various format (GEXF, GDF, GML, NET, GraphML, DL, ... Les Miserables: coappearance weighted **network** of characters in the novel ...
You've visited this page 3 times. Last visit: 24/7/17

Datasets - Duke Network Analysis Center - Duke University
https://dnac.ssri.duke.edu/datasets.php
resources. Datasets. Collections of publicly available **network datasets**. Arizona State University Network Data Collection · Barbasi's Lab · Collection of trust ...

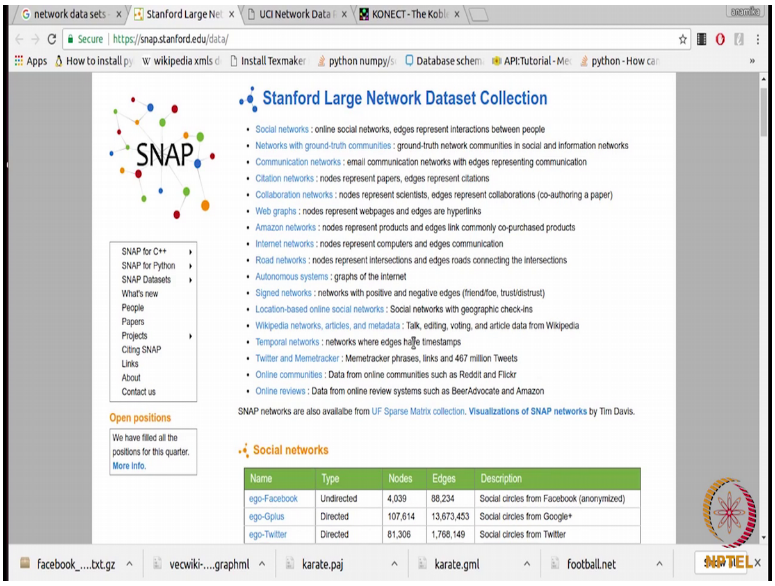
Where can I find sample social network analysis data sets? - Quora
https://www.quora.com/Where-can-I-find-sample-social-network-analysis-data-sets
Where to find social network analysis data sets, apart from the ones found ... First, I assume that you mean sample "**network**" **data sets**, as placing it with the term ...

Network Data Visibility - ThousandEyes
Ad www.thousandeyes.com/TryFree
See **Networks** Beyond Your Firewall. Detect & Diagnose Issues Today. Try It Free.
Path Visualization · VoIP Monitoring · BGP Routing · DNS Troubleshooting
Create Free Account · Contact Us · Internet Outage Detection · Endpoint Agent

Thousands of Free Data Sets - Download & Share Cool Datasets
Ad www.data.world/data-sets

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)



Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

Open positions

We have filled all the positions for this quarter. **More info.**

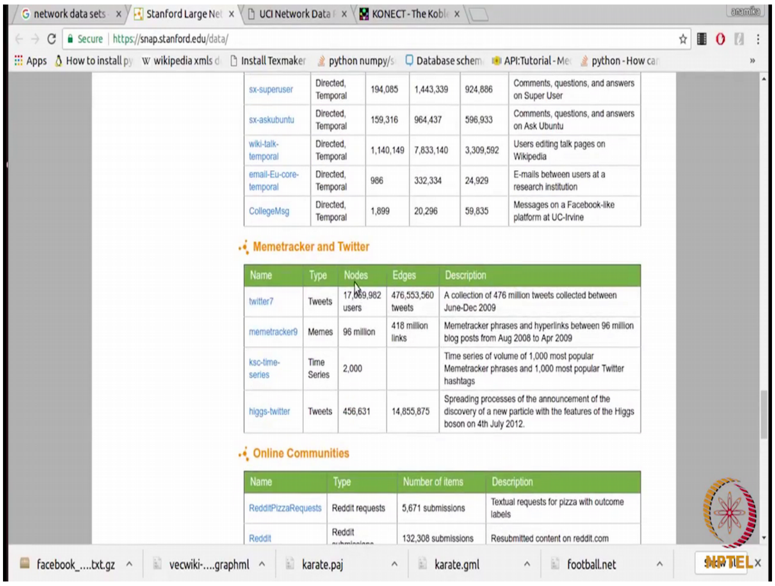
Social networks

| Name | Type | Nodes | Edges | Description |
|---|---|---|---|---|
| ego-Facebook | Undirected | 4,039 | 88,234 | Social circles from Facebook (anonymized) |
| ego-Gplus | Directed | 107,614 | 13,673,453 | Social circles from Google+ |
| ego-Twitter | Directed | 81,306 | 1,768,149 | Social circles from Twitter |

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)



| | | | | | |
|---|---|---|---|---|---|
| sx-superuser | Directed, Temporal | 194,085 | 1,443,339 | 924,886 | Comments, questions, and answers on Super User |
| sx-askubuntu | Directed, Temporal | 159,316 | 964,437 | 596,933 | Comments, questions, and answers on Ask Ubuntu |
| wiki-talk-temporal | Directed, Temporal | 1,140,149 | 7,833,140 | 3,309,592 | Users editing talk pages on Wikipedia |
| email-Eu-core-temporal | Directed, Temporal | 986 | 332,334 | 24,929 | E-mails between users at a research institution |
| CollegeMsg | Directed, Temporal | 1,899 | 20,296 | 59,835 | Messages on a Facebook-like platform at UC-Irvine |

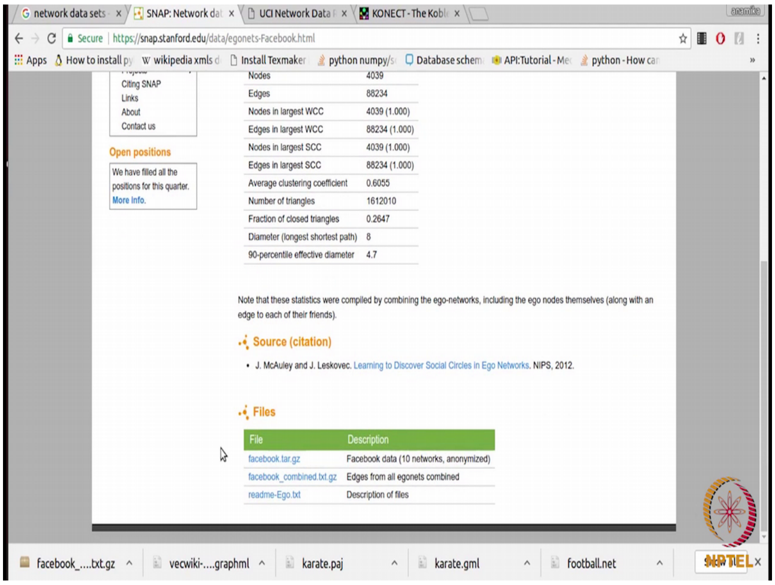
Memetracker and Twitter

| Name | Type | Nodes | Edges | Description |
|---|---|---|---|---|
| twitter7 | Tweets | 17,069,982 users | 476,553,560 tweets | A collection of 476 million tweets collected between June-Dec 2009 |
| memetracker9 | Memes | 96 million | 418 million links | Memetracker phrases and hyperlinks between 96 million blog posts from Aug 2008 to Apr 2009 |
| ksc-time-series | Time Series | 2,000 | | Time series of volume of 1,000 most popular Memetracker phrases and 1,000 most popular Twitter hashtags |
| higgs-twitter | Tweets | 456,631 | 14,855,875 | Spreading processes of the announcement of the discovery of a new particle with the features of the Higgs boson on 4th July 2012. |

Online Communities

| Name | Type | Number of items | Description |
|---|---|---|---|
| RedditPizzaRequests | Reddit requests | 5,671 submissions | Textual requests for pizza with outcome labels |
| Reddit | Reddit submissions | 132,308 submissions | Resubmitted content on reddit.com |

(Refer Slide Time: 02:52)



| | |
|---|---|
| Nodes | 4039 |
| Edges | 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 | 0.2647 |
| 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)

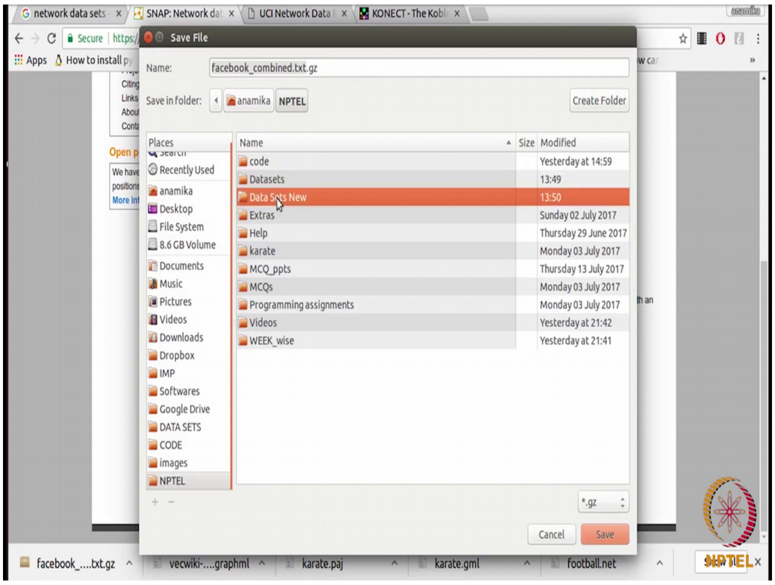
- J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

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 t he 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)



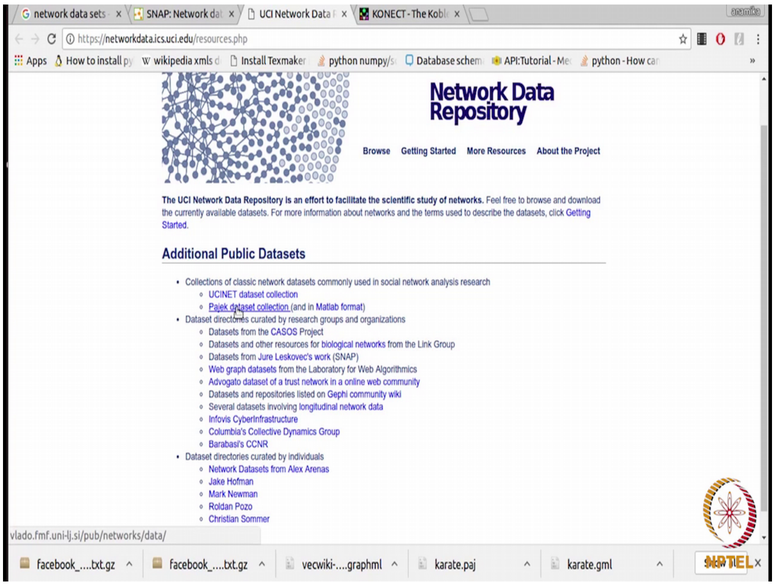
Name: facebook_combined.txt.gz

Save in folder: anamika NPTEL

| 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)



Network Data Repository

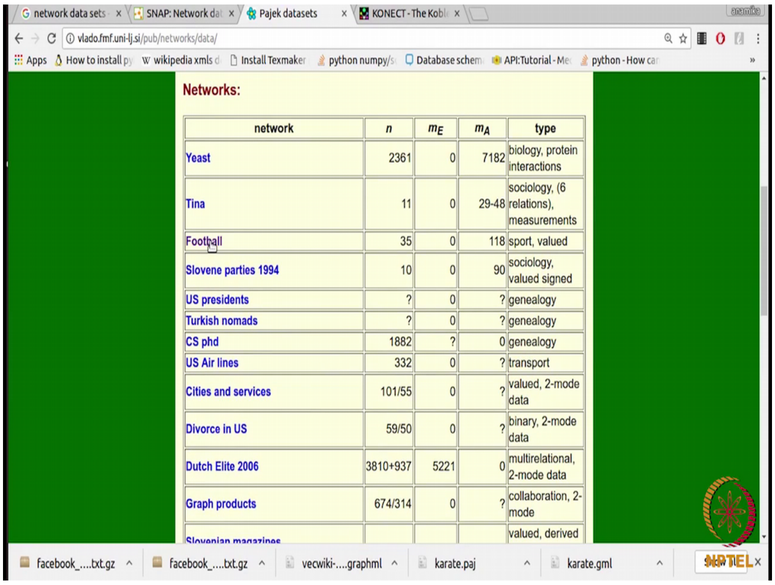
**The UCI Network Data Repository is an effort to facilitate the scientific study of networks.** Feel free to browse and download the currently available datasets. For more information about networks and the terms used to describe the datasets, click Getting Started.

## Additional Public Datasets

- Collections of classic network datasets commonly used in social network analysis research
  - UCINET dataset collection
  - Pajek dataset collection (and in Matlab format)
- Dataset directories curated by research groups and organizations
  - Datasets from the CASOS Project
  - Datasets and other resources for biological networks from the Link Group
  - Datasets from Jure Leskovec's work (SNAP)
  - Web graph datasets from the Laboratory for Web Algorithmics
  - Advogato dataset of a trust network in a online web community
  - Datasets and repositories listed on Gephi community wiki
  - Several datasets involving longitudinal network data
  - Infovis CyberInfrastructure
  - Columbia's Collective Dynamics Group
  - Barabasi's CCNR
- Dataset directories curated by individuals
  - Network Datasets from Alex Arenas
  - Jake Hofman
  - Mark Newman
  - Roldan Pozo
  - Christian Sommer

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 Pajik network and click this link. So, here we have all the Pajik datasets.

(Refer Slide Time: 04:02)

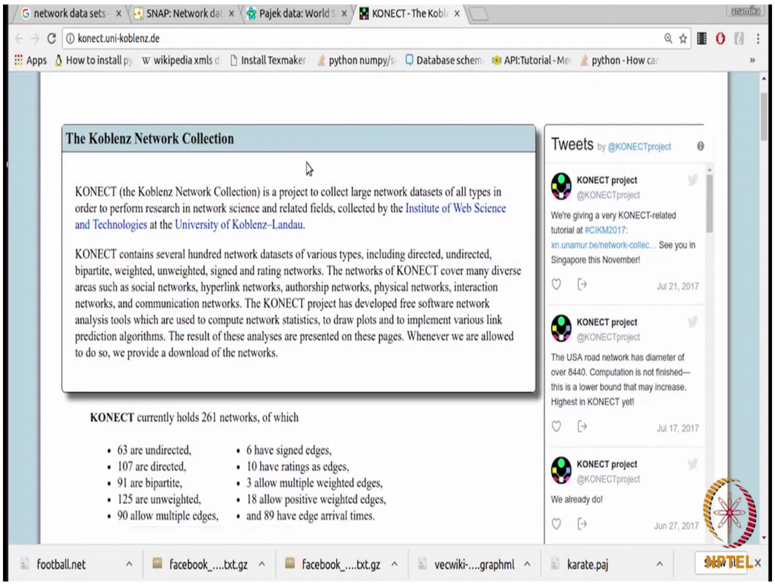


Networks:

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

(Refer Slide Time: 04:43)



### The Koblenz Network Collection

KONECT (the Koblenz Network Collection) is a project to collect large network datasets of all types in order to perform research in network science and related fields, collected by the Institute of Web Science and Technologies at the University of Koblenz–Landau.

KONECT contains several hundred network datasets of various types, including directed, undirected, bipartite, weighted, unweighted, signed and rating networks. The networks of KONECT cover many diverse areas such as social networks, hyperlink networks, authorship networks, physical networks, interaction networks, and communication networks. The KONECT project has developed free software network analysis tools which are used to compute network statistics, to draw plots and to implement various link prediction algorithms. The result of these analyses are presented on these pages. Whenever we are allowed to do so, we provide a download of the networks.

**KONECT** currently holds 261 networks, of which

- 63 are undirected,
- 107 are directed,
- 91 are bipartite,
- 125 are unweighted,
- 90 allow multiple edges,
- 6 have signed edges,
- 10 have ratings as edges,
- 3 allow multiple weighted edges,
- 18 allow positive weighted edges,
- and 89 have edge arrival times.

Tweets by @KONECTproject

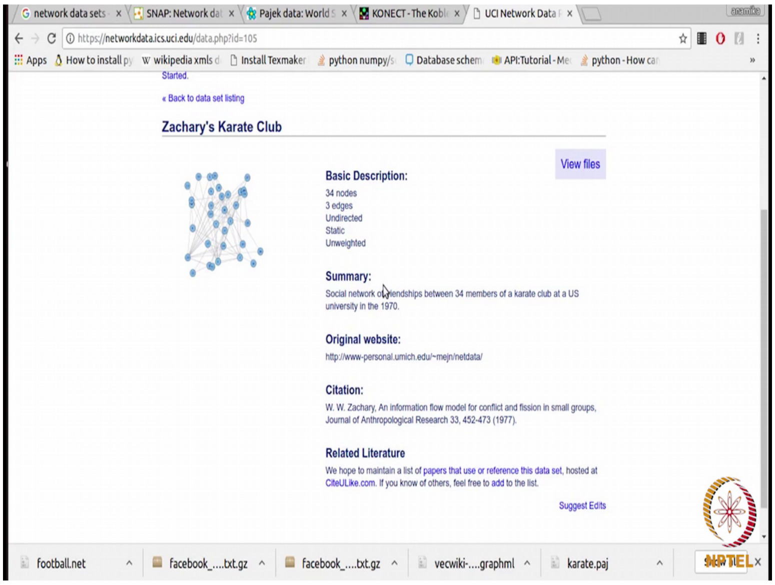
KONECT project @KONECTproject
We're giving a very KONECT-related tutorial at #CIKM2017: xn.unamur.be/network-collec... See you in Singapore this November!
Jul 21, 2017

KONECT project @KONECTproject
The USA road network has diameter of over 8440. Computation is not finished—this is a lower bound that may increase. Highest in KONECT yet!
Jul 17, 2017

KONECT project @KONECTproject
We already do!
Jun 27, 2017

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)



Started.

« Back to data set listing

Zachary's Karate Club

View files

Basic Description:

34 nodes
3 edges
Undirected
Static
Unweighted

Summary:

Social network of friendships between 34 members of a karate club at a US university in the 1970.

Original website:

http://www-personal.umich.edu/~mejn/netdata/

Citation:

W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).

Related Literature

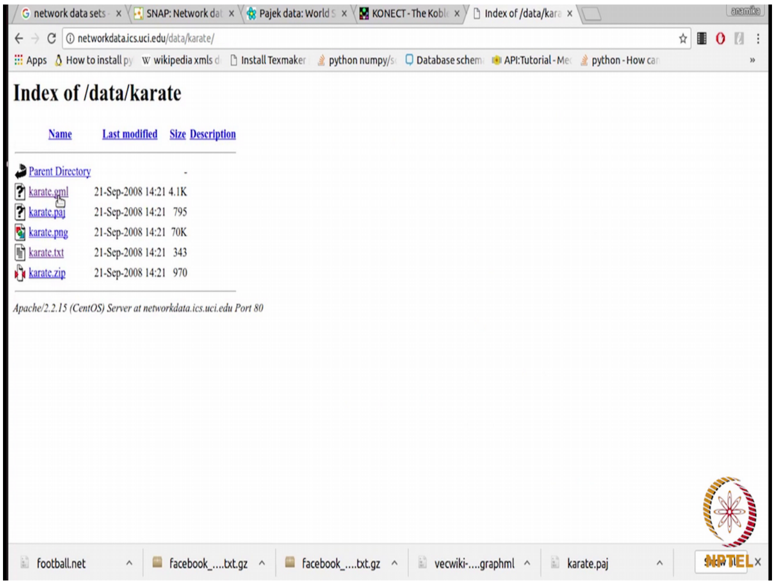
We hope to maintain a list of papers that use or reference this data set, hosted at CiteULike.com. If you know of others, feel free to add to the list.

Suggest Edits

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)



## Index of /data/karate

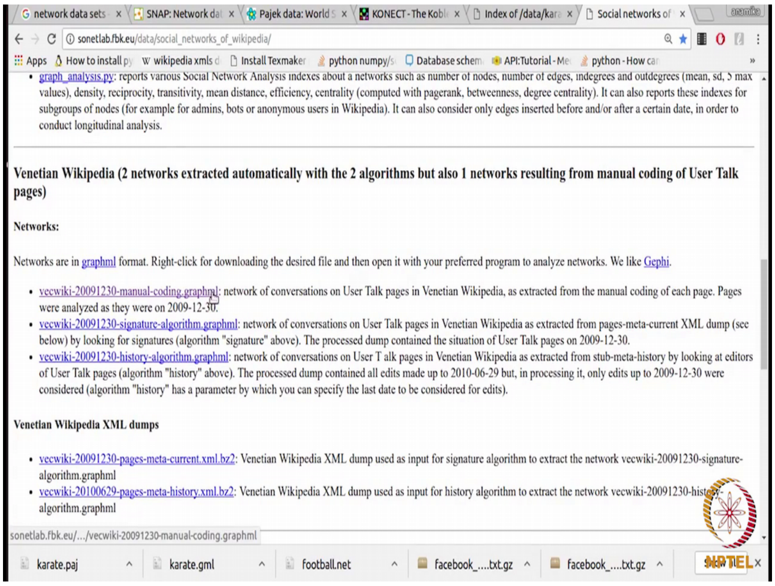
| Name | Last modified | Size | Description |
|---|---|---|---|
| Parent Directory |  | - |  |
| karate.gml | 21-Sep-2008 14:21 | 4.1K |  |
| karate.paj | 21-Sep-2008 14:21 | 795 |  |
| karate.png | 21-Sep-2008 14:21 | 70K |  |
| karate.txt | 21-Sep-2008 14:21 | 343 |  |
| karate.zip | 21-Sep-2008 14:21 | 970 |  |

*Apache/2.2.15 (CentOS) Server at networkdata.ics.uci.edu Port 80*

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



- graph_analysis.py: reports various Social Network Analysis indexes about a networks such as number of nodes, number of edges, indegrees and outdegrees (mean, sd, 5 max values), density, reciprocity, transitivity, mean distance, efficiency, centrality (computed with pagerank, betweenness, degree centrality). It can also reports these indexes for subgroups of nodes (for example for admins, bots or anonymous users in Wikipedia). It can also consider only edges inserted before and/or after a certain date, in order to conduct longitudinal analysis.

**Venetian Wikipedia (2 networks extracted automatically with the 2 algorithms but also 1 networks resulting from manual coding of User Talk pages)**

**Networks:**

Networks are in graphml format. Right-click for downloading the desired file and then open it with your preferred program to analyze networks. We like Gephi.

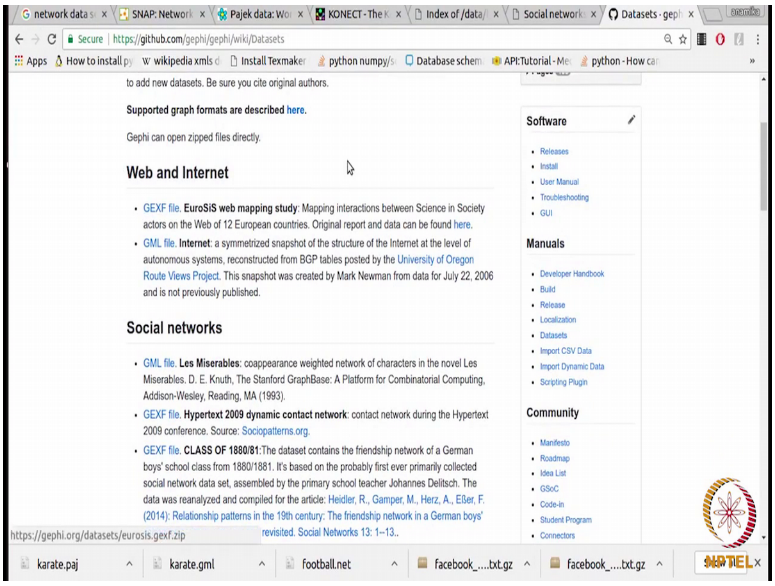
- vecwiki-20091230-manual-coding.graphml: network of conversations on User Talk pages in Venetian Wikipedia, as extracted from the manual coding of each page. Pages were analyzed as they were on 2009-12-30.
- vecwiki-20091230-signature-algorithm.graphml: network of conversations on User Talk pages in Venetian Wikipedia as extracted from pages-meta-current XML dump (see below) by looking for signatures (algorithm "signature" above). The processed dump contained the situation of User Talk pages on 2009-12-30.
- vecwiki-20091230-history-algorithm.graphml: network of conversations on User T alk pages in Venetian Wikipedia as extracted from stub-meta-history by looking at editors of User Talk pages (algorithm "history" above). The processed dump contained all edits made up to 2010-06-29 but, in processing it, only edits up to 2009-12-30 were considered (algorithm "history" has a parameter by which you can specify the last date to be considered for edits).

**Venetian Wikipedia XML dumps**

- vecwiki-20091230-pages-meta-current.xml.bz2: Venetian Wikipedia XML dump used as input for signature algorithm to extract the network vecwiki-20091230-signature-algorithm.graphml
- vecwiki-20100629-pages-meta-history.xml.bz2: Venetian Wikipedia XML dump used as input for history algorithm to extract the network vecwiki-20091230-history-algorithm.graphml

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)



to add new datasets. Be sure you cite original authors.

**Supported graph formats are described here.**

Gephi can open zipped files directly.

Web and Internet

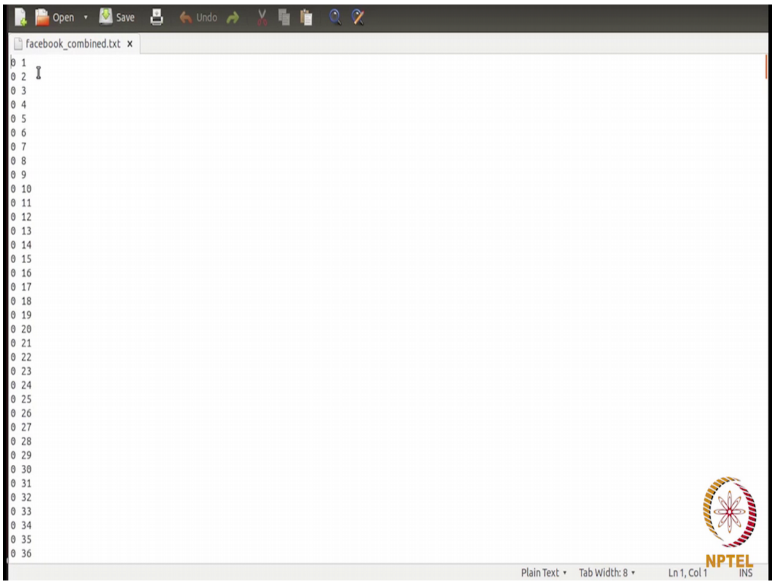
- GEXF file. **EuroSiS web mapping study**: Mapping interactions between Science in Society actors on the Web of 12 European countries. Original report and data can be found here.
- GML file. **Internet**: a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project. This snapshot was created by Mark Newman from data for July 22, 2006 and is not previously published.

Social networks

- GML file. **Les Miserables**: coappearance weighted network of characters in the novel Les Miserables. D. E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA (1993).
- GEXF file. **Hypertext 2009 dynamic contact network**: contact network during the Hypertext 2009 conference. Source: Sociopatterns.org.
- GEXF file. **CLASS OF 1880/81**:The dataset contains the friendship network of a German boys' school class from 1880/1881. It's based on the probably first ever primarily collected social network data set, assembled by the primary school teacher Johannes Delitsch. The data was reanalyzed and compiled for the article: Heidler, R., Gamper, M., Herz, A., Eßer, F. (2014): Relationship patterns in the 19th century: The friendship network in a German boys' revisited. Social Networks 13: 1--13..

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)

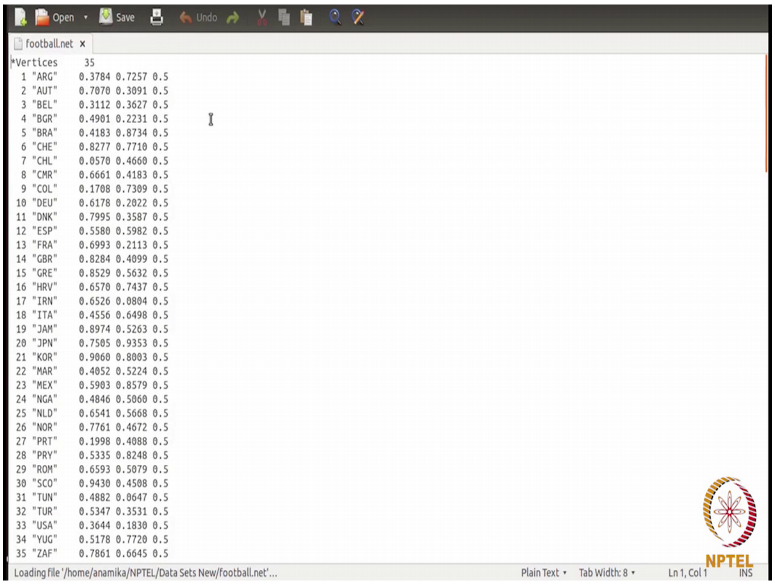


```
0 1
0 2
0 3
0 4
0 5
0 6
0 7
0 8
0 9
0 10
0 11
0 12
0 13
0 14
0 15
0 16
0 17
0 18
0 19
0 20
0 21
0 22
0 23
0 24
0 25
0 26
0 27
0 28
0 29
0 30
0 31
0 32
0 33
0 34
0 35
0 36
```

So, let us look at what we have downloaded we have all these network let me extract this first. So, here we have six network GEXF format txt format dot net format dot gml format dot pajik 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)

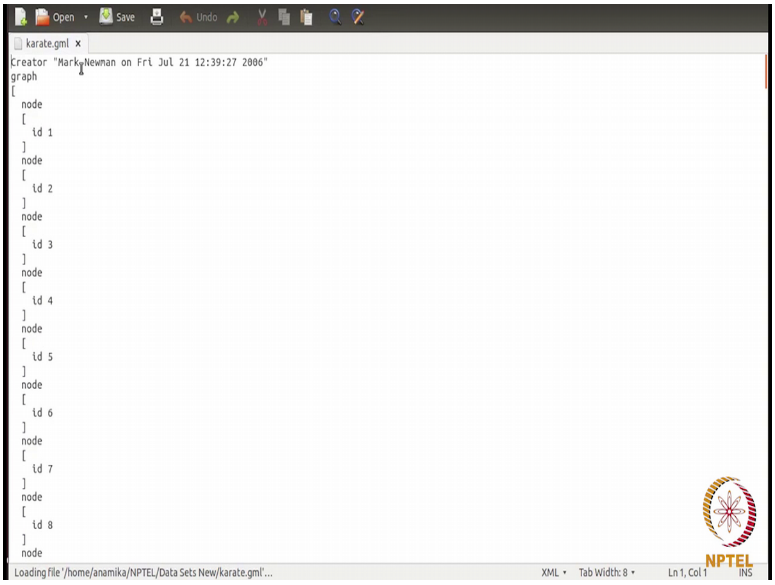


```
*Vertices     35
  1 "ARG"     0.3784 0.7257 0.5
  2 "AUT"     0.7070 0.3091 0.5
  3 "BEL"     0.3112 0.3627 0.5
  4 "BGR"     0.4901 0.2231 0.5
  5 "BRA"     0.4183 0.8734 0.5
  6 "CHE"     0.8277 0.7710 0.5
  7 "CHL"     0.0570 0.4660 0.5
  8 "CMR"     0.6661 0.4183 0.5
  9 "COL"     0.1708 0.7309 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 0.0804 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"     0.5903 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 "PRT"     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"     0.4882 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
```

Loading file '/home/anamika/NPTEL/Data Sets New/football.net'...

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)

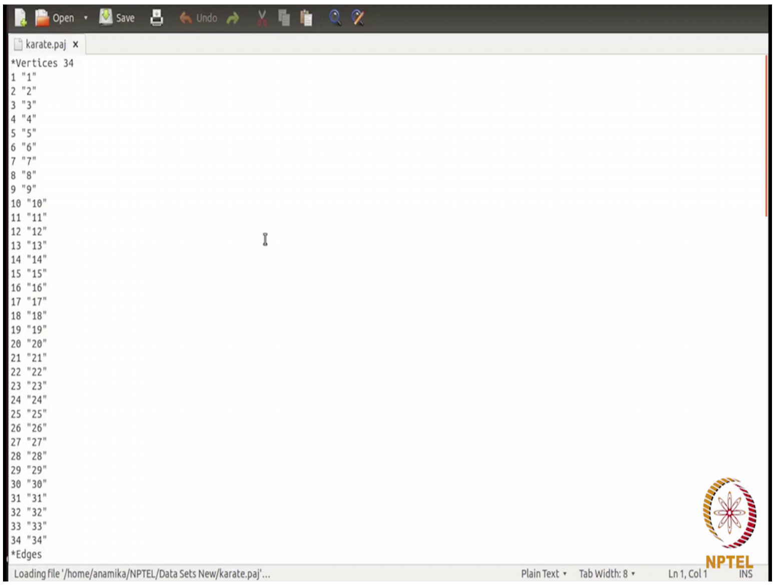


```
Creator "Mark Newman on Fri Jul 21 12:39:27 2006"
graph
[
  node
  [
    id 1
  ]
  node
  [
    id 2
  ]
  node
  [
    id 3
  ]
  node
  [
    id 4
  ]
  node
  [
    id 5
  ]
  node
  [
    id 6
  ]
  node
  [
    id 7
  ]
  node
  [
    id 8
  ]
  node
```

```
Loading file '/home/anamika/NPTEL/Data Sets New/karate.gml'...    XML ▾   Tab Width: 8 ▾   Ln 1, Col 1   INS
```

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)

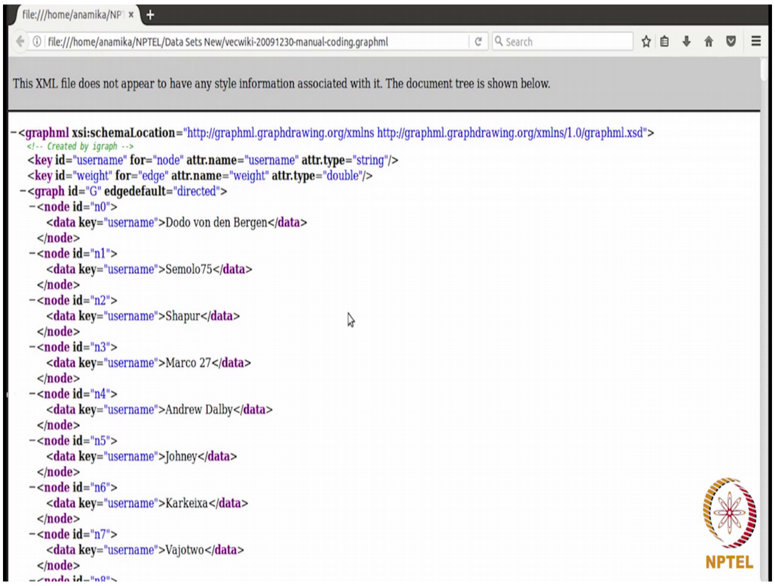


```
*Vertices 34
1 "1"
2 "2"
3 "3"
4 "4"
5 "5"
6 "6"
7 "7"
8 "8"
9 "9"
10 "10"
11 "11"
12 "12"
13 "13"
14 "14"
15 "15"
16 "16"
17 "17"
18 "18"
19 "19"
20 "20"
21 "21"
22 "22"
23 "23"
24 "24"
25 "25"
26 "26"
27 "27"
28 "28"
29 "29"
30 "30"
31 "31"
32 "32"
33 "33"
34 "34"
*Edges
```

```
Loading file '/home/anamika/NPTEL/Data Sets New/karate.paj'...    Plain Text ▾   Tab Width: 8 ▾   Ln 1, Col 1   INS
```

Now, this is a karate network in Pajik 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 Pajik. 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)



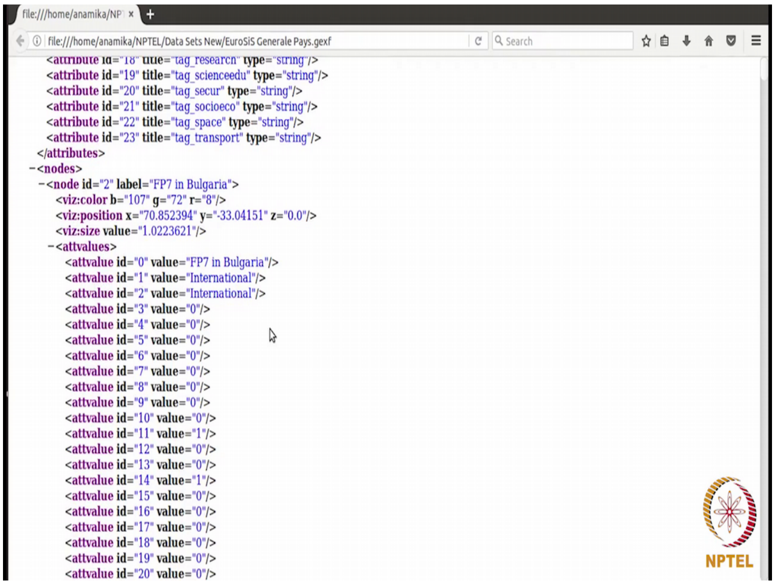
file:///home/anamika/NPTEL/Data Sets New/vecwiki-20091230-manual-coding.graphml

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
-<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">Karkeixa</data>
   </node>
  -<node id="n7">
     <data key="username">Vajotwo</data>
   </node>
  -<node id="n8">
```

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)

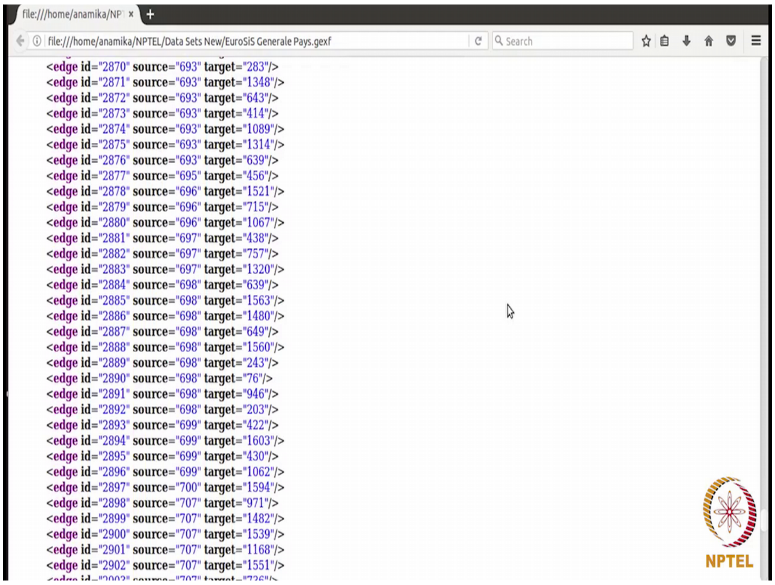


file:///home/anamika/NP
file:///home/anamika/NPTEL/Data Sets New/EuroSiS Generale Pays.gexf
Search
NPTEL

```
    <attribute id="18" title="tag_research" type="string"/>
    <attribute id="19" title="tag_scienceedu" 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 this 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 nodes 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)


file:///home/anamika/NP ×
file:///home/anamika/NPTEL/Data Sets New/EuroSiS Generale Pays.gexf
Search
NPTEL

```
<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="715"/>
<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="698" 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"/>
<edge id="2903" source="707" target="736"/>
```

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.