Practical handout for the workshop "Introduction to the analysis of large-scale data on social connections"

Prepared for the Historical Demography Scientific Research Network workshop, Utrecth, 26 April 2012

Paul Lambert and Dave Griffiths, University of Stirling, April 2012

Version 1

Contents

Introduction

This handout accompanies the lab sessions for the workshp 'Introduction to the analysis of large-scale data on social connections' (26 April 2012, Utrecht).

For Stata and R, the step-by-step implementation instructions for each session are largely to be found within the specific 'syntax' files for the relevant sessions (.do and .R files). For Pajek, step-by-step instructions with screenshots are provided below.

Parts of this handout are extracted from a more extended handout on using data analysis packages for social science research, produced by Lambert for the DAMES Node workshop programme (see www.dames.org.uk) and for his course 'Introduction to multilevel models with applications' to the Essex Summer School in Social Science data analysis (www.staff.stir.ac.uk/paul.lambert/essex_summer_school).

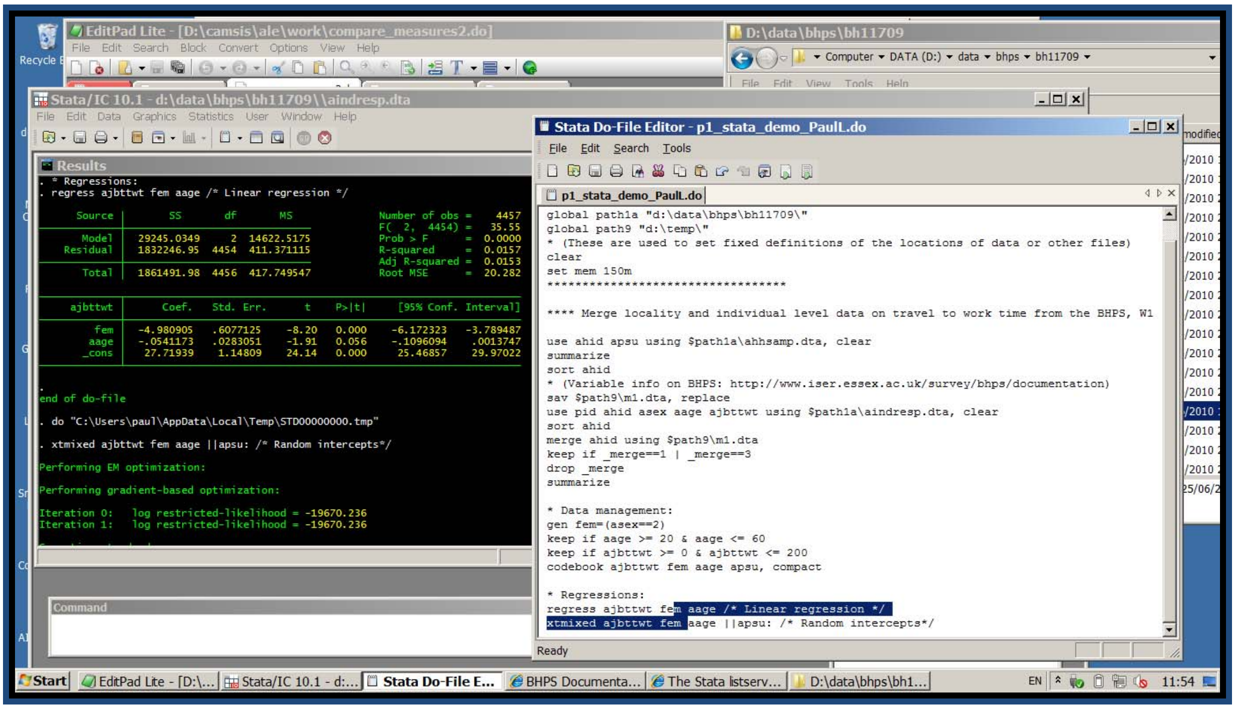
General arrangements for the practicals

Unless noted otherwise the data files used are for distribution for this lab session only and should not be transferred elsewhere. The sources of these files are ultimately avaiable online from international data providers such as NAPP (http://www.nappdata.org/napp/) or the UK's ESDS (www.esds.ac.uk).

You will need to have the relevant packages installed to undertake the relevant exercises (though it should be possible to use only some of the packages as relevant). Introductory notes on the packages are included below under 'lab 1'. Some of the packages and lab exercises have online dependencies (e.g. they need to use a data file or programme extension which is available online). We have tried to note the details below, but information may not be comprehensive.

In general terms, the task in the labs is to open up the relevant syntax files, and work your way through them, digesting the examples shown (and potentially adding your own notes, adjustments or examples). You'll ordinarily need to have the analytical software open and the relevant tool for working with a syntax file (e.g. 'syntax window' or 'do file editor'). In addition it will typically also be helpful to have open some applications to remind you of where the data is stored, and perhaps a plain text editor allowing you to conveniently open up several other syntax files for purposes of comparison.

- When working with Stata a typical view of your desktop might be something like:



EditPad Lite - [D:\camsis\ale\work\compare_measures2.do]
D:\data\bhps\bh11709
Computer ▸ DATA (D:) ▸ data ▸ bhps ▸ bh11709
Stata/IC 10.1 - d:\data\bhps\bh11709\\aindresp.dta
File Edit Data Graphics Statistics User Window Help
Results
Command
Stata Do-File Editor - p1_stata_demo_PaulL.do
File Edit Search Tools
p1_stata_demo_PaulL.do
Ready
Start
EditPad Lite - [D:\...
Stata/IC 10.1 - d:...
Stata Do-File E...
BHPS Documenta...
The Stata listserv...
D:\data\bhps\bh1...
EN
11:54

```
. * Regressions:
. regress ajbttwt fem aage /* Linear regression */

      Source |       SS       df       MS              Number of obs =    4457
-------------+------------------------------           F(  2,  4454) =   35.55
       Model |  29245.0349     2  14622.5175           Prob > F      =  0.0000
    Residual |  1832246.95  4454  411.371115           R-squared     =  0.0157
-------------+------------------------------           Adj R-squared =  0.0153
       Total |  1861491.98  4456  417.749547           Root MSE      =  20.282

------------------------------------------------------------------------------
     ajbttwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -4.980905   .6077125    -8.20   0.000    -6.172323   -3.789487
        aage |  -.0541173   .0283051    -1.91   0.056    -.1096094    .0013747
       _cons |   27.71939    1.14809    24.14   0.000     25.46857    29.97022
------------------------------------------------------------------------------

.
end of do-file

. do "C:\Users\paul\AppData\Local\Temp\STD00000000.tmp"

. xtmixed ajbttwt fem aage ||apsu: /* Random intercepts*/

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -19670.236
Iteration 1:   log restricted-likelihood = -19670.236
```

```
global path1a "d:\data\bhps\bh11709\"
global path9 "d:\temp\"
* (These are used to set fixed definitions of the locations of data or other files)
clear
set mem 150m
**********************************

**** Merge locality and individual level data on travel to work time from the BHPS, W1

use ahid apsu using $path1a\ahhsamp.dta, clear
summarize
sort ahid
* (Variable info on BHPS: http://www.iser.essex.ac.uk/survey/bhps/documentation)
sav $path9\m1.dta, replace
use pid ahid asex aage ajbttwt using $path1a\aindresp.dta, clear
sort ahid
merge ahid using $path9\m1.dta
keep if _merge==1 | _merge==3
drop _merge
summarize

* Data management:
gen fem=(asex==2)
keep if aage >= 20 & aage <= 60
keep if ajbttwt >= 0 & ajbttwt <= 200
codebook ajbttwt fem aage apsu, compact

* Regressions:
regress ajbttwt fem aage /* Linear regression */
xtmixed ajbttwt fem aage ||apsu: /* Random intercepts*/
```

Description: The first two interfaces you can see in this screenshot are respectively the Stata do file editor (where I write commands and send them to be processed, such as by highlighting the relevant lines and clicking 'ctrl-d'); and the main Stata window (here Version 10) which includes the results page. Note that the syntax file open is a modified (personalised) version of the supplied illustrative syntax file – the name has been changed so that I can save the original file plus my own version of it after edits (e.g. with my own comments). Behind the scenes I've also got open an 'Editpadlite' session which I'm using to conveniently open up and compare some other sytnax files that I don't particularly want in my do file editor itself; I've also got a file manager software open showing the data I'm drawing upon (in what Stata will call 'path1a'); and I've got some Internet Exporer (IE) sessions open (I'm looking up the online BHPS documentation, and the Stata listserv, where information on Stata commands is available).

Materials referred to in the sessions will include:

- data files (copies of survey and other data used);
- sample command files (pre-prepared materials which include programming commands in the language of the relevant software)
- supplementary 'macros' or 'sub-files' (further pre-prepared materials featuring programming commands in relevant languages, usually invoked as a sub-routine within the main sample command files)

An important point to make is that some of the command files will need to draw upon other files (e.g. data files) in order to run. To do this, they need to be able to reference the location of the required files. In most applications, we do this be defining 'macros' which point to specific 'paths' on your computer (see also software sections below). For the labs to work successfully, it will be necessary to ensure that the command file you are trying to run is pointing to the right paths at the right time. In general, this only requires one specification to be made at the start of the session, for instance whereby in Stata we define 'macros' for the relevant paths. Sometimes however it can be necessary to edit the full path reference of a particular file in order to be able to access it.

For example, in the text below, we show some Stata (and SPSS) commands which in both cases define a macro (called ‘path3a’) which gives the directory location of the data file ‘aindresp.dta’ or ‘aindresp.sav’, so that subsequent commands calling it will go directly to that path:

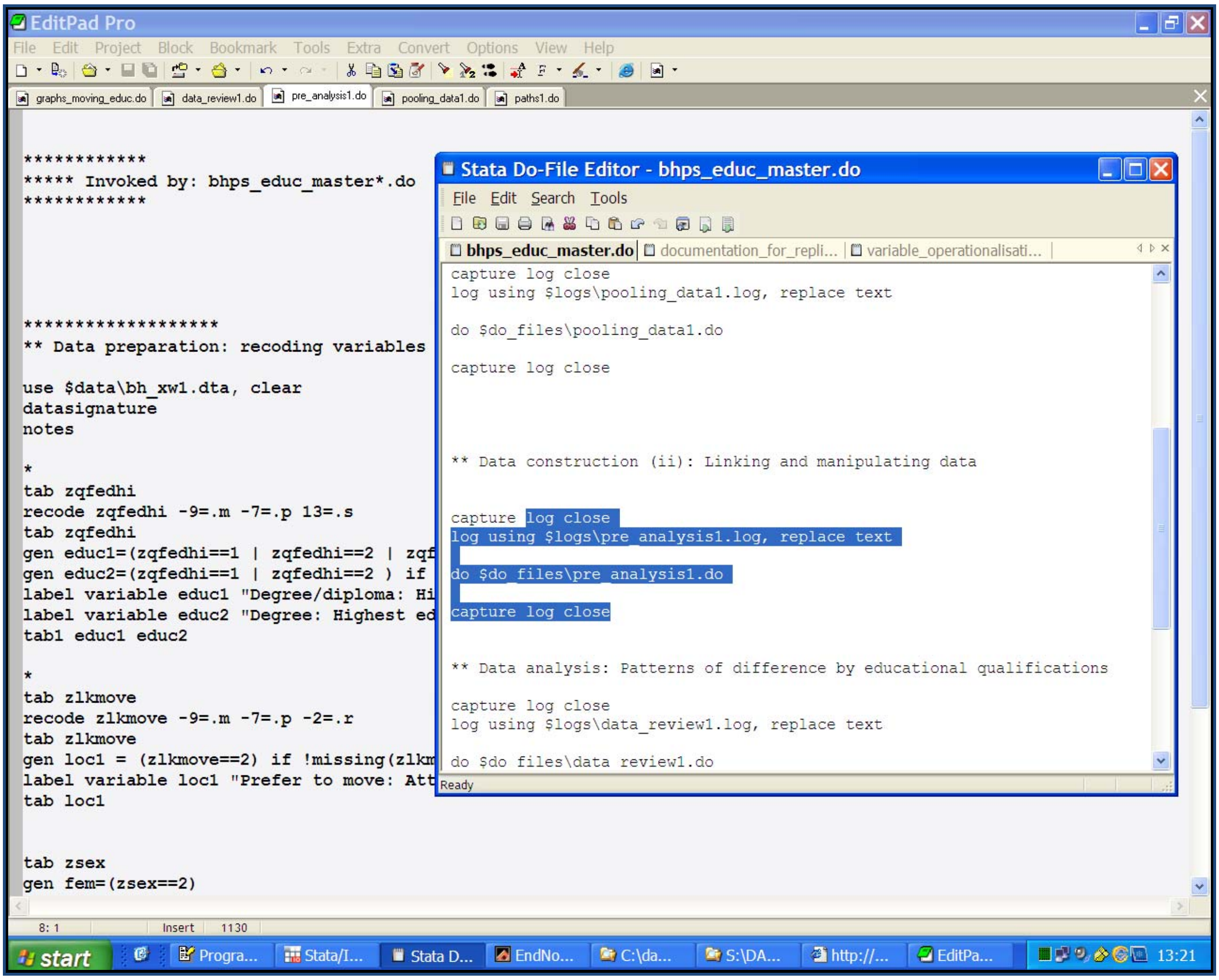
Stata example	***comparable SPSS example***
`global path3a  “d:\data\bhps\”`	`define !path3a () “d:\data\bhps\” !enddefine.`
`use pid asex using $path3a\aindresp.dta, clear`	`get file=!path3a+”aindresp.sav”.`
`tab asex`	`fre var=asex.`

Relevant background: Thinking about workflows

There are very good expositions of the idea of workflows in the social science data analysis process in, amongst others, Long (2009); Treiman (2009), and Kohler and Kreuter (2008). A workflow in itself is a representation of a series of tasks which contribute to a project or activity. It can be a useful exercise to conceptualise a research project as a workflow (with components such as data collection, data processing, data analysis, report writing). However, when dealing with large scale data, a really useful contribution is to organise your data and command files that are associated with a project in a consistent style that recognises that relevant contributions to the workflow structure.

What does that involve? The issue is that we want to construct a replicable trail of our data oriented research, which allows us to go all the way from opening the initial data file, to producing the publication quality graph or statistical results which are our end products. We need the replicable trail in order to adjust our analysis on the basis of minor changes at any possible stage of the process (or to be able to transfer a record of our work on to others). However because when dealing with large-scale and complex data (e.g. on social connections) the trail is long and complex (and not entirely linear), we can only do this, realistically, if we break down our activities into multiple separate components.

There are different ways to organise files for these purposes, but a popular and highly effective approach is to design a ‘master’ syntax command file and a series of ‘sub-files’ which it draws upon. In this model, the sub-files cover different parts of the research analysis. Personally, my preference is to construct both the master and sub-files in the principle software package being used, though Long (2009) notes that creating a documentation master file in a different software (e.g. MS Excel) is an effective way to record a wider range of activities which span across different software. Here’s an example of a series of tasks being called upon via a Stata format ‘master’ file:

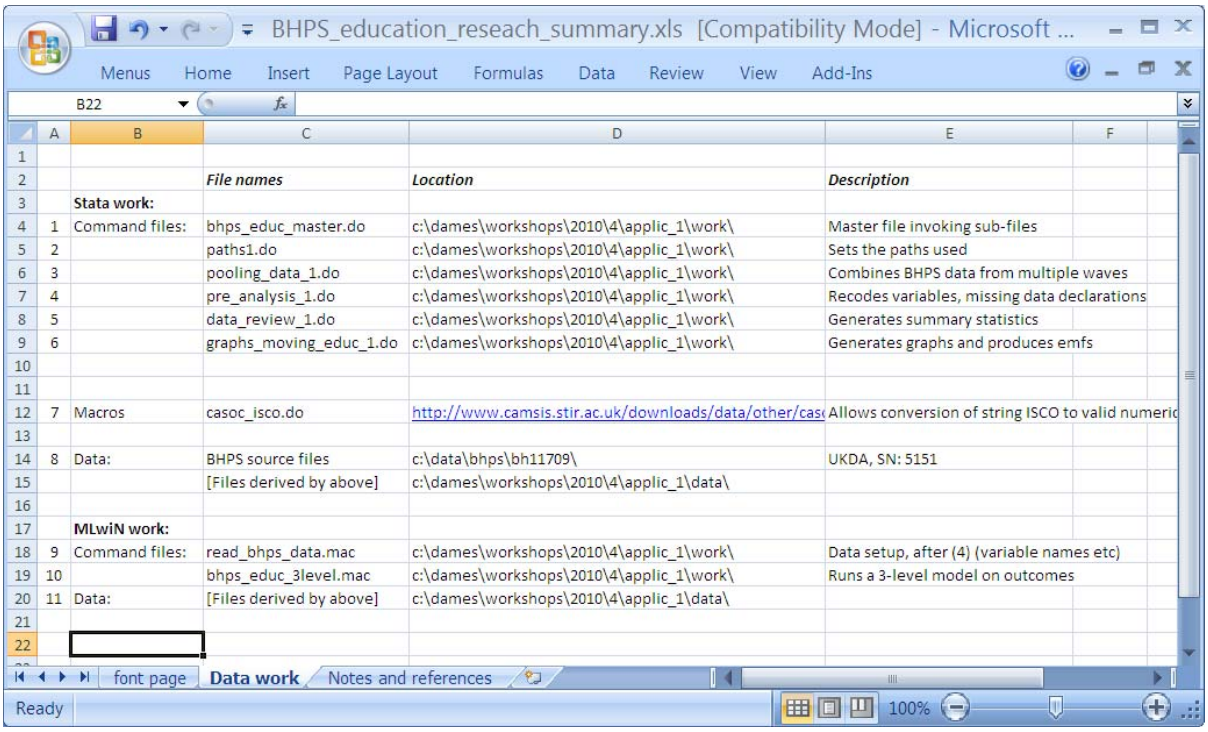


```
************
***** Invoked by: bhps_educ_master*.do
************




*******************
** Data preparation: recoding variables

use $data\bh_xw1.dta, clear
datasignature
notes

*
tab zqfedhi
recode zqfedhi -9=.m -7=.p 13=.s
tab zqfedhi
gen educ1=(zqfedhi==1 | zqfedhi==2 | zqf
gen educ2=(zqfedhi==1 | zqfedhi==2 ) if
label variable educ1 "Degree/diploma: Hi
label variable educ2 "Degree: Highest ed
tab1 educ1 educ2

*
tab zlkmove
recode zlkmove -9=.m -7=.p -2=.r
tab zlkmove
gen loc1 = (zlkmove==2) if !missing(zlkm
label variable loc1 "Prefer to move: Att
tab loc1


tab zsex
gen fem=(zsex==2)
```

```
capture log close
log using $logs\pooling_data1.log, replace text

do $do_files\pooling_data1.do

capture log close




** Data construction (ii): Linking and manipulating data


capture log close
log using $logs\pre_analysis1.log, replace text

do $do_files\pre_analysis1.do

capture log close


** Data analysis: Patterns of difference by educational qualifications

capture log close
log using $logs\data_review1.log, replace text

do $do_files\data_review1.do
```

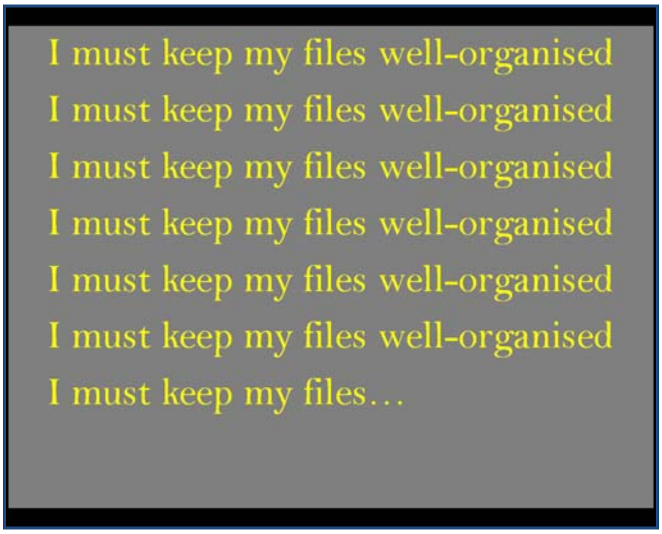
(This screenshot shows the Stata master file, and the sub-files which are mostly open within the EditPad editor - except for a few other files which I've opened in the do file editor. The Stata output file is not visible but is open behind the scenes).

Here’s an example of a project documentation file that might be constructed in Excel:



	A	B	C	D	E
1					
2			*File names*	*Location*	*Description*
3		**Stata work:**			
4	1	Command files:	bhps_educ_master.do	c:\dames\workshops\2010\4\applic_1\work\	Master file invoking sub-files
5	2		paths1.do	c:\dames\workshops\2010\4\applic_1\work\	Sets the paths used
6	3		pooling_data_1.do	c:\dames\workshops\2010\4\applic_1\work\	Combines BHPS data from multiple waves
7	4		pre_analysis_1.do	c:\dames\workshops\2010\4\applic_1\work\	Recodes variables, missing data declarations
8	5		data_review_1.do	c:\dames\workshops\2010\4\applic_1\work\	Generates summary statistics
9	6		graphs_moving_educ_1.do	c:\dames\workshops\2010\4\applic_1\work\	Generates graphs and produces emfs
10					
11					
12	7	Macros	casoc_isco.do	http://www.camsis.stir.ac.uk/downloads/data/other/casc	Allows conversion of string ISCO to valid numeric
13					
14	8	Data:	BHPS source files	c:\data\bhps\bh11709\	UKDA, SN: 5151
15			[Files derived by above]	c:\dames\workshops\2010\4\applic_1\data\	
16					
17		**MLwiN work:**			
18	9	Command files:	read_bhps_data.mac	c:\dames\workshops\2010\4\applic_1\work\	Data setup, after (4) (variable names etc)
19	10		bhps_educ_3level.mac	c:\dames\workshops\2010\4\applic_1\work\	Runs a 3-level model on outcomes
20	11	Data:	[Files derived by above]	c:\dames\workshops\2010\4\applic_1\data\	
21					
22					

Note that the other tabs in the Excel file can be used to show things like author details, context of the analysis, and last update time. The file also notes some (though not all) dependencies within the workflow – for instance step 9 requires step 4 to have been take (the macro reads in a plain text data file that was generated in Stata by do file pre_analysis1.do).


I must keep my files well-organised

I must keep my files well-organised

I must keep my files well-organised

I must keep my files well-organised

I must keep my files well-organised

I must keep my files well-organised

I must keep my files…

In summary, we can’t advise you strongly enough on the value of organising you data files around a workflow conceptualisation, such as through master and sub-files. Read the opening chapters of Long (2009), or the other references mentioned above, for more on this theme. *We’d encourage you to look at the workshop materials from the ‘DAMES’ research Node, at www.dames.org.uk, for more on this topic.*

Software alternatives

Many different software packages can be used effectively for applied research using complex data on social connections. Various packages support the estimation of a wide range of statistical models including association models, and there are numerous (mostly different) packages which feature techniques for social network analysis.

In this session we focus upon three packages which bring slightly different contributions to multilevel modelling.

- **Stata** is used because it is a popular general purpose package for data management and data analysis which also includes a substantial range of analysis options for dealing with data on social connections. Stata is attractive to applied researchers for many reasons, including its good facilities for storing and summarising estimation results; its support of a wide range of advanced analytical methods which complement a multilevel analysis (e.g. clustering estimators used in Economics); and its wide range of data management functions suited to complex data. Stata is proprietory and may be purchased from: www.stata.com.

- **R** is used because it is a popular freeware that supports many forms of statistical model estimation, social network analysis examples, and has various graphical and data construction capabilties. Many of its facilities are available via extension 'libraries' which are usually installed online. R is a difficult language for social scientists to work effectively with, however, because it brings with it very high 'overheads' in its programming requirements, especially for large and complex data. R is available to install as freeware from: http://www.r-project.org/

- **Pajek** is used because it is a freely available and popular package for social network analysis, featuring a wide range of graphical and statistical analysis possibilities. Pajek may be downloaded and installed as freeware from: http://pajek.imfm.si/doku.php

We should stress that many more packages can be used effectively for the analyses used below. In addition, an exciting software development in the area being led in the UK is the construction of a generic interface for specifying and estimating complex statistical models of 'arbitrary complexity'. These cover most forms of multilevel models, as well as many other statistical modelling devices. This project is called '**e-Stat**' and is expecting to generate it first publicly available resources over the period 2010-2012 (see http://www.cmm.bristol.ac.uk/research/NCESS-EStat/).

Lab 1: Introduction to the analysis of social connections data

This lab introduces a few examples of datasets on social connections, and provides illustrative analyses in Stata, R and Pajek.

The work of the Stata and R exercises is done by the corresponding command files, which should (hopefully) be self-explanatory. We introduce using Stata and R below.

The Pajek exercises are described through step-by-step instructions, provided below.

Background: Introducing Stata

Stata was first developed in 1984 and was originally used mainly in academic research in economics. From approximately the mid 1990's its functionalities for social survey data analysis began to filter through to other social science disciplines, and in the last decade it has displaced SPSS as the most popular intermediate-to-advanced level statistical analysis package in most academic disciplines which use social survey data (e.g. sociology, educational research, geography).

Stata is popular for many good reasons. The list of features of Stata that lead me personally to favour this package above others are:

- It supports explicit documentation of complex processes through a concise and 'human readable' syntax language
- It supports a wide range of data management functions including many routines useful in complex survey data which are not readily performed in other packages (e.g. 'egen', 'xtdes')
- It supports a very full range of statistical modelling options, including several advanced model specifications which are not widely available elsewhere
- It has excellent graphics capabilities, supporting the specification and export of publication quality graphs (in a syntactical, replicable manner)
- It features very convenient tools for storing the results from multiple models or analyses and compiling them in summary tables or files (e.g. 'est store', 'statsby')
- It can read online data files and run command files and macros from online locations
- It supports extended add-on programming capabilities, and benefits from a large, constructive community of user-contributed extensions (see e.g. http://www.stata.com/links/resources3.html)

In pragmatic terms, most users of Stata are reasonably confident programmers, and getting started with the package does need a little effort in learning about data manipulation and data analysis. This is one reason why Stata is not yet widely taught in introductory social science courses, though, in the UK for example, it is increasingly used in intermediate and advanced level teaching (e.g. MSc programmes or Undergraduate social science programmes with extended statistical components).

A common problem with working with Stata is that many institutions do not have site-level access to the software, and accordingly many individual researchers don't have access to the package - Stata is generally sold as an 'n-user' package, which means that an institution buys a specified number of copies at any one time. Recently however, access to Stata for academic researchers has probably be made easier by the Stata 'GradPlan', which allows purchase of personal copies of the package for students and faculty at fairly low price – see http://www.stata.com/order/new/edu/gradplan.html . Stata also comes in several different forms with different upper limits on the scale of data it may handle – 'Small Stata' is not normally adequate for working with advanced survey datasets; 'Intercooled' Stata (I/C) usually has more than enough capacity to support social survey research analysis (although, working with a large scale resources you may occasionally hit upper limits, such as on the number of variables or cases, it is usually possible to find an easy work-around such as by dropping unnecessary variables); Stata SE and MP offer greater capacity regarding the size of datasets and faster processing power, but they are more expensive to purchase. To my knowledge, most academic researchers use Intercooled Stata.

In summary, many users of Stata favour the package not because it offers one particular functionality which others don't, but because it offers an integrated set of advanced functionalities covering data management and data analysis which can't easily be matched by any other software. For other texts which explain the strengths and attractions of Stata, see for example Treiman (2009).

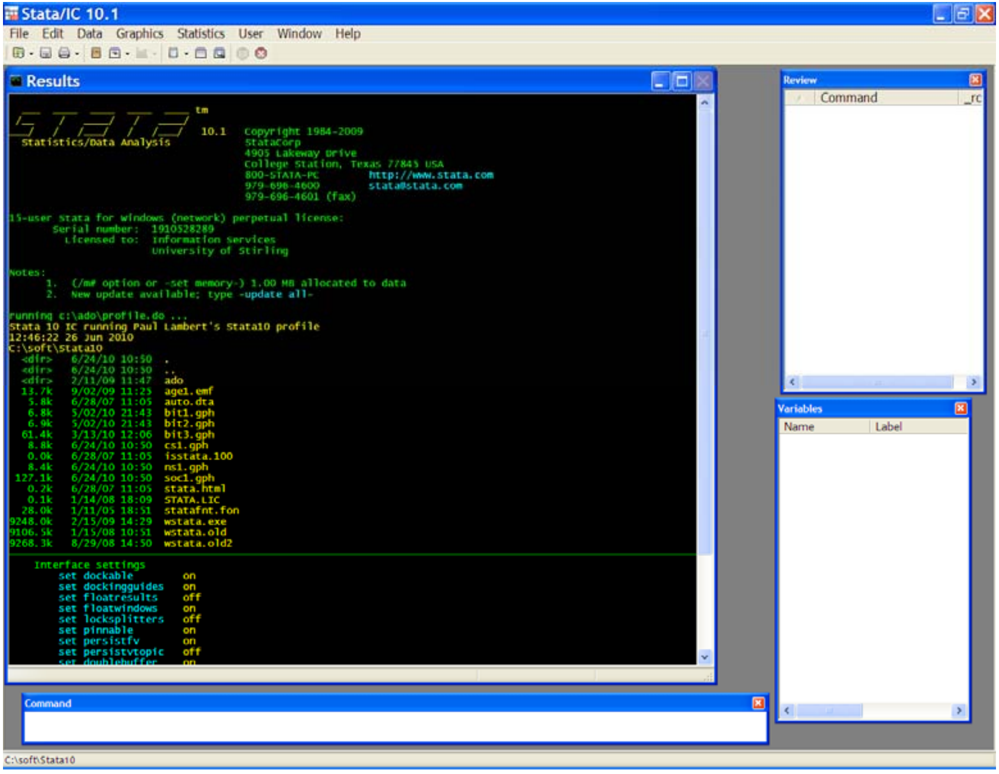
The steps below give you some relevant instructions on working with Stata for the purposes of the module (the examples are mostly from the Practical 1 Stata file). Many online resources on Stata are available, in particular we highlight:

- UCLA's ATS pages: http://www.ats.ucla.edu/stat/stata/ (Features a wide range of materials including videos of using Stata and routines across the range of the package)
- The CMM's LEMMA online course: http://www.cmm.bristol.ac.uk/learning-training/course.shtml (includes detailed descriptions of running basic regression models and of specifying random effects multilevel models in Stata)
- In the first lab session we point you to an illustrative do file which serves as an introduction to Stata, available from www.longitudinal.stir.ac.uk

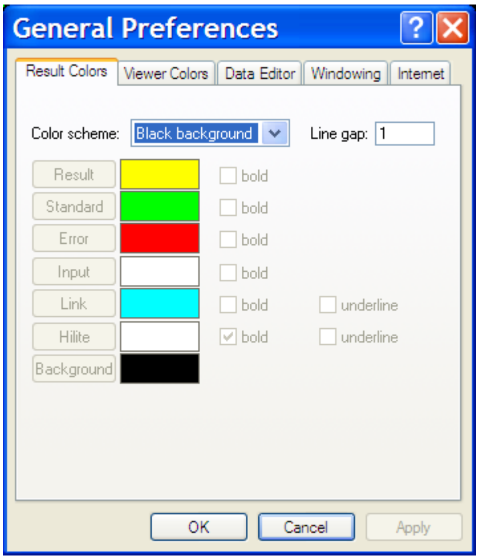
When you launch the package, you see the basic Stata window, here for version I/C 10.1.

You can customise its appearance (e.g. right click on the results window) – the image on the right is how I set up the windows on my machine and will be slightly different to what you see on first launching the package in the lab at Essex.

On opening the programme (this image shows Stata version 10):

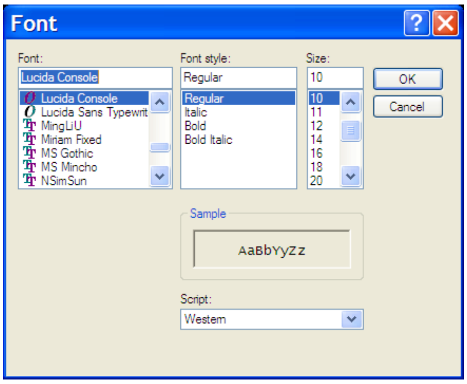


```
                                   tm
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   10.1   Copyright 1984-2009
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

15-user Stata for Windows (network) perpetual license:
       Serial number:  1910528289
         Licensed to:  Information Services
                       University of Stirling

Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data
      2.  New update available; type -update all-

running c:\ado\profile.do ...
Stata 10 IC running Paul Lambert's Stata10 profile
12:46:22 26 Jun 2010
C:\soft\Stata10
  <dir>   6/24/10 10:50  .
  <dir>   6/24/10 10:50  ..
  <dir>   2/11/09 11:47  ado
  13.7k   9/02/09 11:25  age1.emf
   5.8k   6/28/07 11:05  auto.dta
   6.8k   5/02/10 21:43  bit1.gph
   6.9k   5/02/10 21:43  bit2.gph
  61.4k   3/13/10 12:06  bit3.gph
   8.8k   6/24/10 10:50  cs1.gph
   0.0k   6/28/07 11:05  isstata.100
   8.4k   6/24/10 10:50  ns1.gph
 127.1k   6/24/10 10:50  soc1.gph
   0.2k   6/28/07 11:05  stata.html
   0.1k   1/14/08 18:09  STATA.LIC
  28.0k   1/11/05 18:51  statafnt.fon
9248.0k   2/15/09 14:29  wstata.exe
9106.5k   1/15/08 10:51  wstata.old
9268.3k   8/29/08 14:50  wstata.old2

    Interface settings
        set dockable           on
        set dockingguides      on
        set floatresults       off
        set floatwindows       on
        set locksplitters      off
        set pinnable           on
        set persistfv          on
        set persistvtopic      off
        set doublebuffer       on
```



Color scheme: Black background Line gap: 1

Result	🟨	☐ bold	
Standard	🟩	☐ bold	
Error	🟥	☐ bold	
Input	⬜	☐ bold	
Link	🟦	☐ bold	☐ underline
Hilite	⬜	☑ bold	☐ underline
Background	⬛		



Font: Lucida Console

Font style: Regular

Size: 10

Font list: Lucida Console, Lucida Sans Typewrit, MingLiU, Miriam Fixed, MS Gothic, MS Mincho, NSimSun

Font style list: Regular, Italic, Bold, Bold Italic

Size list: 10, 11, 12, 14, 16, 18, 20

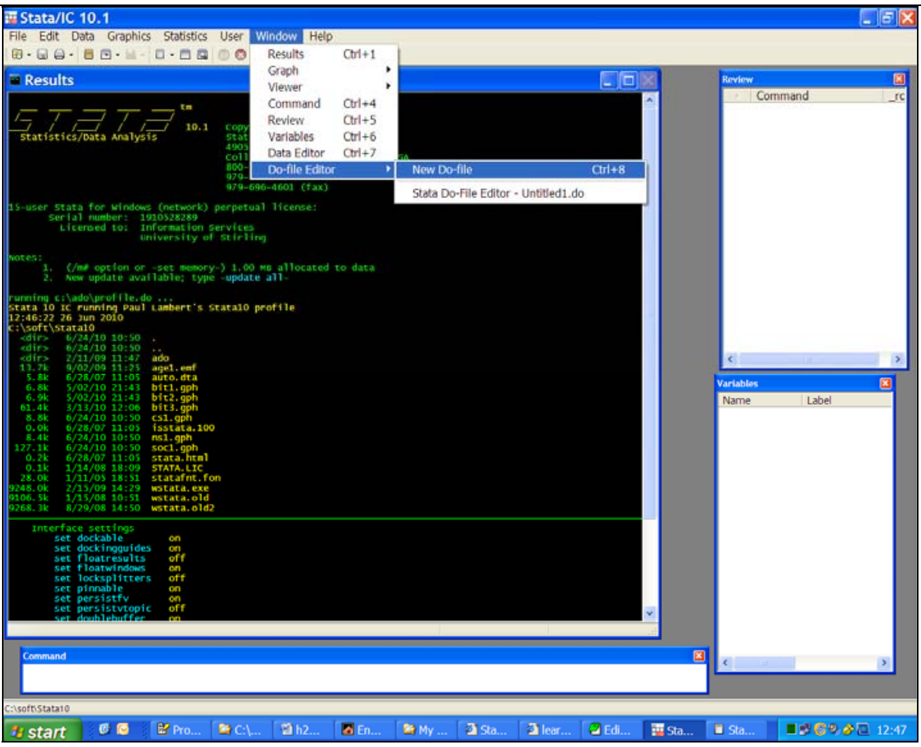
Sample: AaBbYyZz

Script: Western

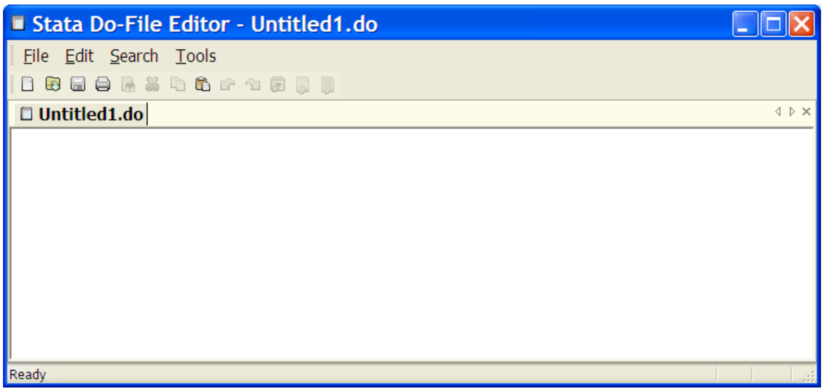
The very first thing you should do at the start of every session it so ask explicitly to open the 'do file' editor with '**ctrl+8**' or via the GUI.

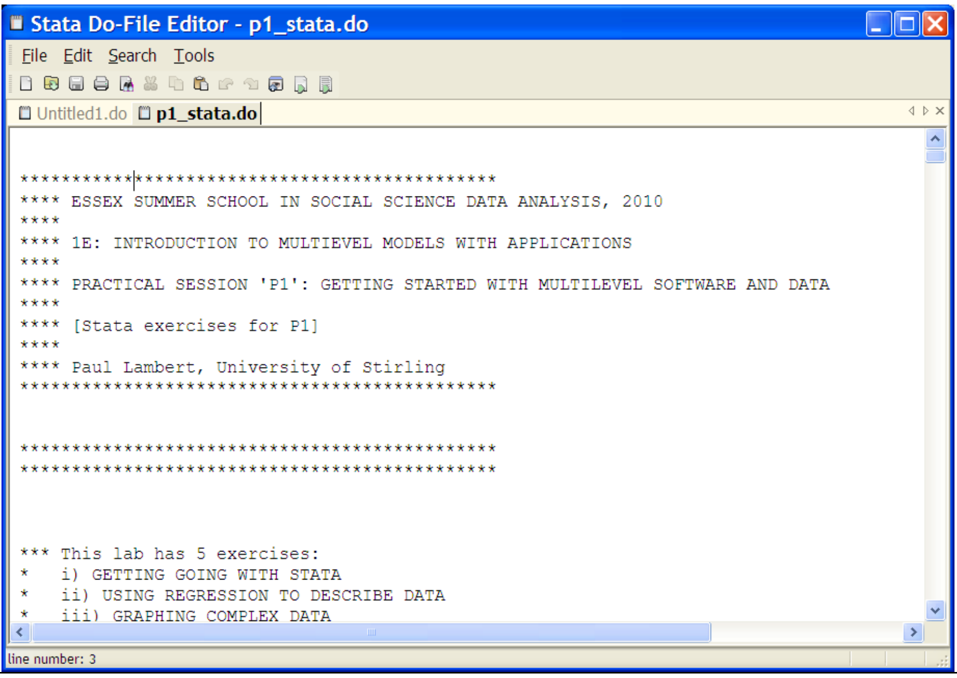
Note below that we can have several do files open at once.

Not shown below, but from Stata 11 onwards, it is possible to permit various formatting options in the do file editor (e.g. colour coding). It is also possible to set up Stata to run directly from a plain text editor if you wish to (search online for how to do this).



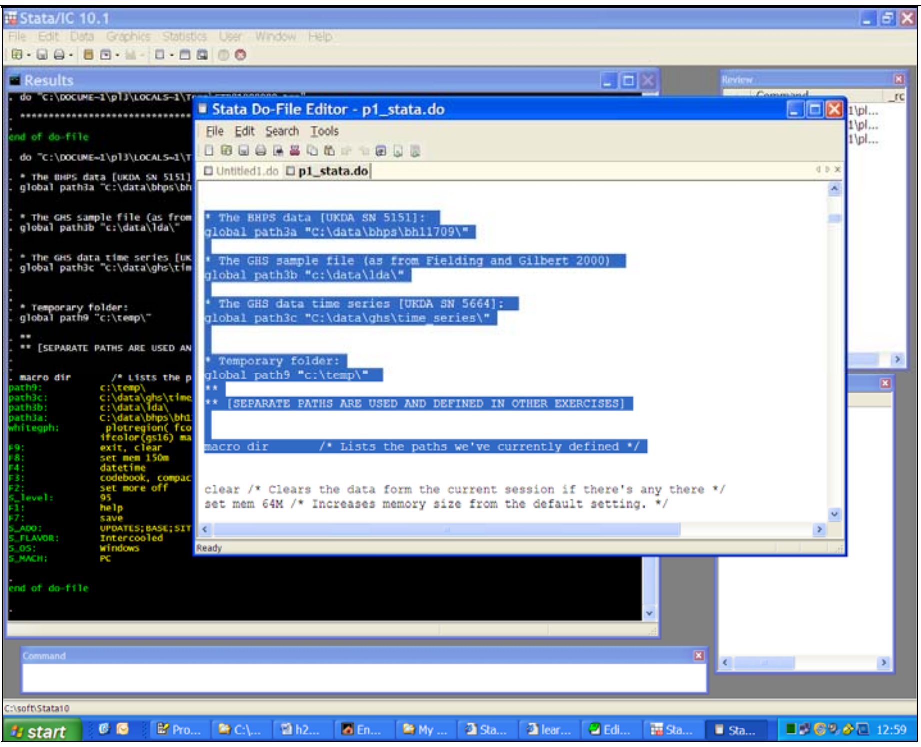
```
                                      tm
  STATA                        10.1   copy
  Statistics/Data Analysis            Stat
                                      4905
                                      coll
                                      800-
                                      979-
                                      979-696-4601 (fax)

15-user Stata for Windows (network) perpetual license:
       Serial number:  1910528289
         Licensed to:  Information Services
                       University of Stirling

Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data
      2.  New update available; type -update all-

running c:\ado\profile.do ...
Stata 10 IC running Paul Lambert's Stata10 profile
12:46:22 26 Jun 2010
c:\soft\Stata10
  <dir>   6/24/10 10:50  .
  <dir>   6/24/10 10:50  ..
  <dir>   2/11/09 11:47  ado
  13.7k   9/02/09 11:25  age1.emf
   5.8k   6/28/07 11:05  auto.dta
   6.8k   5/02/10 21:43  bit1.gph
   6.9k   5/02/10 21:43  bit2.gph
  61.4k   3/13/10 12:06  bit3.gph
   8.8k   6/24/10 10:50  cs1.gph
   0.0k   6/28/07 11:05  isstata.100
   8.4k   6/24/10 10:50  ns1.gph
 127.1k   6/24/10 10:50  soc1.gph
   0.2k   6/28/07 11:05  stata.html
   0.1k   1/14/08 18:09  STATA.LIC
  28.0k   1/11/05 18:51  statafnt.fon
9248.0k   2/15/09 14:29  wstata.exe
9106.5k   1/15/08 10:51  wstata.old
9268.3k   8/29/08 14:50  wstata.old2

    Interface settings
        set dockable          on
        set dockingguides     on
        set floatresults      off
        set floatwindows      on
        set locksplitters     off
        set pinnable          on
        set persistfv         on
        set persistvtopic     off
        set doublebuffer      on
```





```
************************************************
**** ESSEX SUMMER SCHOOL IN SOCIAL SCIENCE DATA ANALYSIS, 2010
****
**** 1E: INTRODUCTION TO MULTIEVEL MODELS WITH APPLICATIONS
****
**** PRACTICAL SESSION 'P1': GETTING STARTED WITH MULTILEVEL SOFTWARE AND DATA
****
**** [Stata exercises for P1]
****
**** Paul Lambert, University of Stirling
************************************************


************************************************
************************************************



*** This lab has 5 exercises:
*   i) GETTING GOING WITH STATA
*   ii) USING REGRESSION TO DESCRIBE DATA
*   iii) GRAPHING COMPLEX DATA
```

Once you've opened a 'do file' you can begin running commands by clicking on the segments of the relevant command lines and clicking 'ctrl+R'.



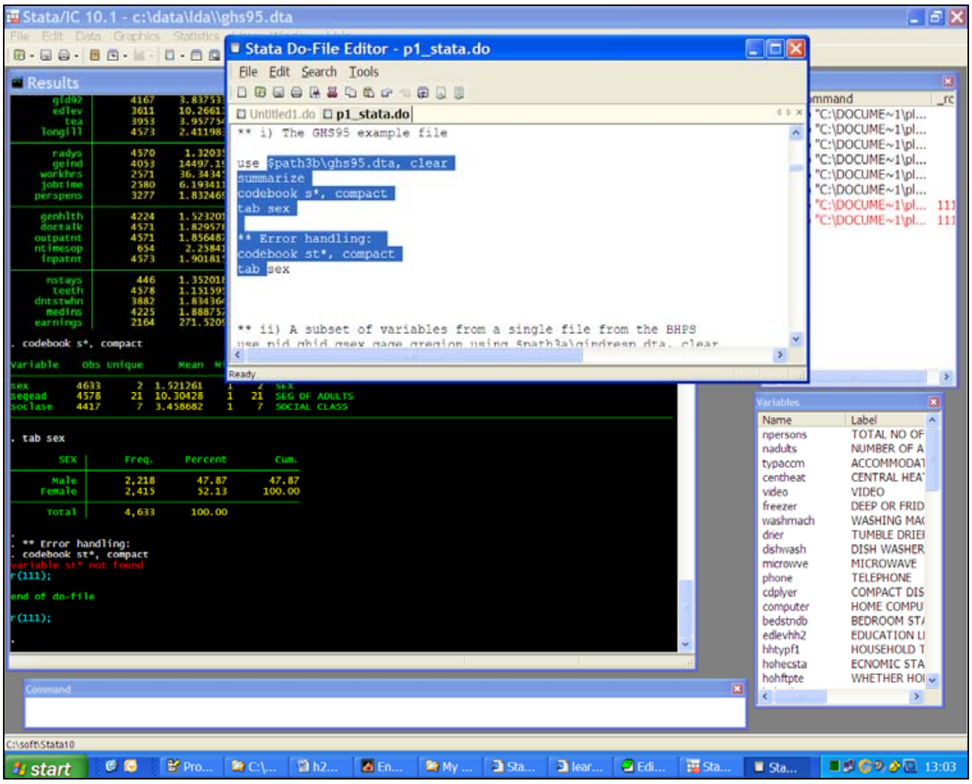
```
* The BHPS data [UKDA SN 5151]:
global path3a "C:\data\bhps\bh11709\"

* The GHS sample file (as from Fielding and Gilbert 2000)
global path3b "c:\data\lda\"

* The GHS data time series [UKDA SN 5664]:
global path3c "C:\data\ghs\time_series\"


* Temporary folder:
global path9 "c:\temp\"
**
** [SEPARATE PATHS ARE USED AND DEFINED IN OTHER EXERCISES]


macro dir        /* Lists the paths we've currently defined */


clear /* Clears the data form the current session if there's any there */
set mem 64M /* Increases memory size from the default setting. */
```

```
. macro dir      /* Lists the p
path9:          c:\temp\
path3c:         c:\data\ghs\time
path3b:         c:\data\lda\
path3a:         C:\data\bhps\bh1
whitegph:        plotregion( fco
                ifcolor(gs16) ma
F9:             exit, clear
F8:             set mem 150m
F4:             datetime
F3:             codebook, compac
F2:             set more off
S_level:        95
F1:             help
F7:             save
S_ADO:          UPDATES;BASE;SIT
S_FLAVOR:       Intercooled
S_OS:           Windows
S_MACH:         PC

end of do-file
```

Important: Defining macros for paths. This particular image shows an important component of the start of every session in the module lab exercises. The lines beginning with 'global' are ways of defining permanent 'macros' for the session. The macros serve to define locations on my machine where my files (e.g. data files) are actually stored. Doing this means that in later commands (e.g. the image below) I can call on files via their macro folder locations rather than their full path – this aids transferability of work between machines/locations.

The results are shown in the results window. Error messages are by default shown in red text and lead to the termination of the sequence of commands at that point

(unlike in SPSS, which carries on, disregarding the error).


Stata/IC 10.1 - c:\data\lda\\ghs95.dta
File Edit Data Graphics Statistics
Results
Stata Do-File Editor - p1_stata.do
File Edit Search Tools
Untitled1.do
p1_stata.do
Ready
Command
Variables
C:\soft\Stata10
start
13:03

```
** i) The GMS95 example file

use $path3b\ghs95.dta, clear
summarize
codebook s*, compact
tab sex

** Error handling:
codebook st*, compact
tab sex



** ii) A subset of variables from a single file from the BHPS
use pid ghid qsex qage qregion using $path3a\qindresp.dta, clear
```

```
. codebook s*, compact

Variable     Obs Unique      Mean  Min  Max  Label

sex         4633      2  1.521261    1    2  SEX
segead      4578     21  10.30428    1   21  SEG OF ADULTS
soclase     4417      7  3.458682    1    7  SOCIAL CLASS


. tab sex

        SEX |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      2,218       47.87       47.87
     Female |      2,415       52.13      100.00
------------+-----------------------------------
      Total |      4,633      100.00

.
. ** Error handling:
. codebook st*, compact
variable st* not found
r(111);

end of do-file

r(111);
```

```
Name       Label
npersons   TOTAL NO OF
nadults    NUMBER OF A
typaccm    ACCOMMODAT
centheat   CENTRAL HEA
video      VIDEO
freezer    DEEP OR FRID
washmach   WASHING MA
drier      TUMBLE DRIE
dishwash   DISH WASHER
microwve   MICROWAVE
phone      TELEPHONE
cdplyer    COMPACT DIS
computer   HOME COMPU
bedstndb   BEDROOM ST
edlevhh2   EDUCATION L
hhtypf1    HOUSEHOLD T
hohecsta   ECNOMIC STA
hohftpte   WHETHER HO
```

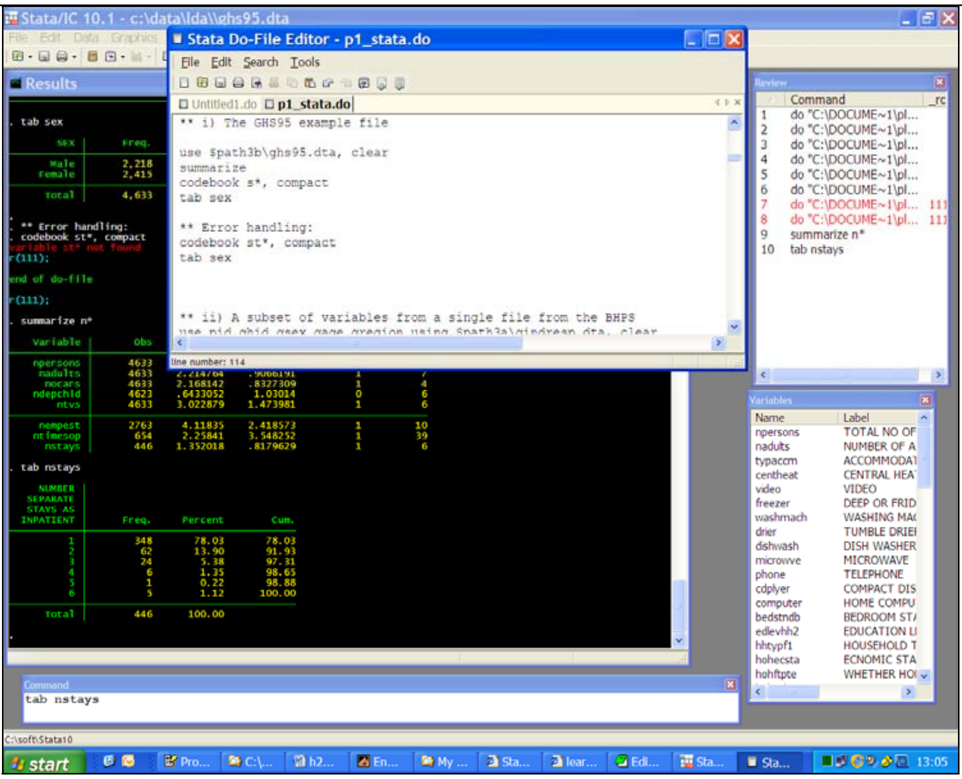
In the above, the macro which I have called 'path3b' means that when Stata reads the line:

    use $path3b\ghs95.dta, clear

what it reads behind the scenes is

    use c:\data\lda\ghs95.dta, clear

You can also submit commands line by line through the command interface (e.g. if you don't want to log them in the do file).



Stata/IC 10.1 - c:\data\lda\\ghs95.dta

Results

```
. tab sex

        SEX |      Freq.
------------+-----------
       Male |      2,218
     Female |      2,415
------------+-----------
      Total |      4,633

.
. ** Error handling:
. codebook st*, compact
variable st* not found
r(111);

end of do-file

r(111);

. summarize n*

    Variable |       Obs
-------------+-----------
    npersons |      4633
     nadults |      4633   2.214764    .9066191          1          7
      nocars |      4633   2.168142    .8327309          1          4
    ndepchld |      4623   .6433052     1.03014          0          6
        ntvs |      4633   3.022879    1.473981          1          6
-------------+------------------------------------------------------
     nempest |      2763    4.11835    2.418573          1         10
   ntimesop |       654    2.25841    3.548252          1         39
     nstays |       446   1.352018    .8179629          1          6

. tab nstays

    NUMBER |
  SEPARATE |
  STAYS AS |
 INPATIENT |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |        348       78.03       78.03
         2 |         62       13.90       91.93
         3 |         24        5.38       97.31
         4 |          6        1.35       98.65
         5 |          1        0.22       98.88
         6 |          5        1.12      100.00
-----------+-----------------------------------
     Total |        446      100.00

.
```

Stata Do-File Editor - p1_stata.do

```
** i) The GHS95 example file

use $path3b\ghs95.dta, clear
summarize
codebook s*, compact
tab sex

** Error handling:
codebook st*, compact
tab sex



** ii) A subset of variables from a single file from the BHPS
use pid ghid gsex gage gregion using $path3a\gindresp.dta, clear
```

line number: 114

Review

	Command	_rc
1	do "C:\DOCUME~1\pl...	
2	do "C:\DOCUME~1\pl...	
3	do "C:\DOCUME~1\pl...	
4	do "C:\DOCUME~1\pl...	
5	do "C:\DOCUME~1\pl...	
6	do "C:\DOCUME~1\pl...	
7	do "C:\DOCUME~1\pl...	111
8	do "C:\DOCUME~1\pl...	111
9	summarize n*	
10	tab nstays	

Variables

Name	Label
npersons	TOTAL NO OF
nadults	NUMBER OF A
typaccm	ACCOMMODAT
centheat	CENTRAL HEA
video	VIDEO
freezer	DEEP OR FRID
washmach	WASHING MA(
drier	TUMBLE DRIEI
dishwash	DISH WASHER
microwve	MICROWAVE
phone	TELEPHONE
cdplyer	COMPACT DIS
computer	HOME COMPU
bedstndb	BEDROOM ST/
edlevhh2	EDUCATION L
hhtypf1	HOUSEHOLD T
hohecsta	ECNOMIC STA
hohftpte	WHETHER HO(

Command

```
tab nstays
```

C:\soft\Stata10

Note how the 'review' window shows lines that were entered through the command window, but it just shows some programming code for commands run through the do file editor.

Note some of the features of the Stata syntax language:

You need to 'clear' the dataspace to read in a new file, e.g.

```
use $path3b\ghs95.dta, clear
```

You can't create a new variable with the same name as an existing one – if it's there already you need to delete it first, e.g.

```
drop fem
gen fem=(sex==2)
```

The 'capture' command suppresses potential error messages so is a useful way to make commands generic

```
capture drop fem
gen fem=(sex==2)
```

Using 'by:' or 'if..' within a command can usefully restrict the specifications:

```
tab nstays if sex==1
bysort sex: tab nstays if age >= 20 & age <= 30
```

The 'numlabel' command is a useful way to show both numeric values and categorical labels compared to the default (labels only), e.g.;

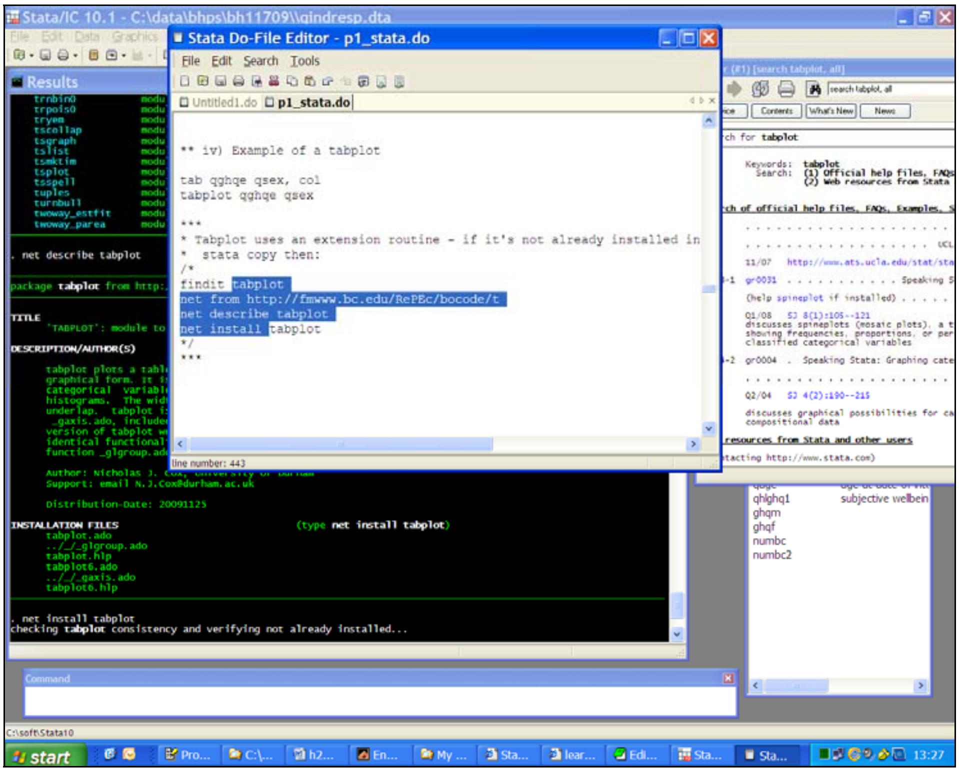
```
tab ecstaa
numlabel _all, add
tab ecstaa
```

There's no requirement for full stops at the end of lines, but a carriage return serves as the default delimiter, and so we usually use '///' to extend a command over more than one line.

```
bysort sex: tab nstays ///
   if age >= 20 & age <= 60 & ecstaa==1
```

Extension routines are often written by users and made available to the wider community. To exploit them, you need to run either 'net install' or 'ssc install'

Example of finding and installing the 'tabplot' extension routine:

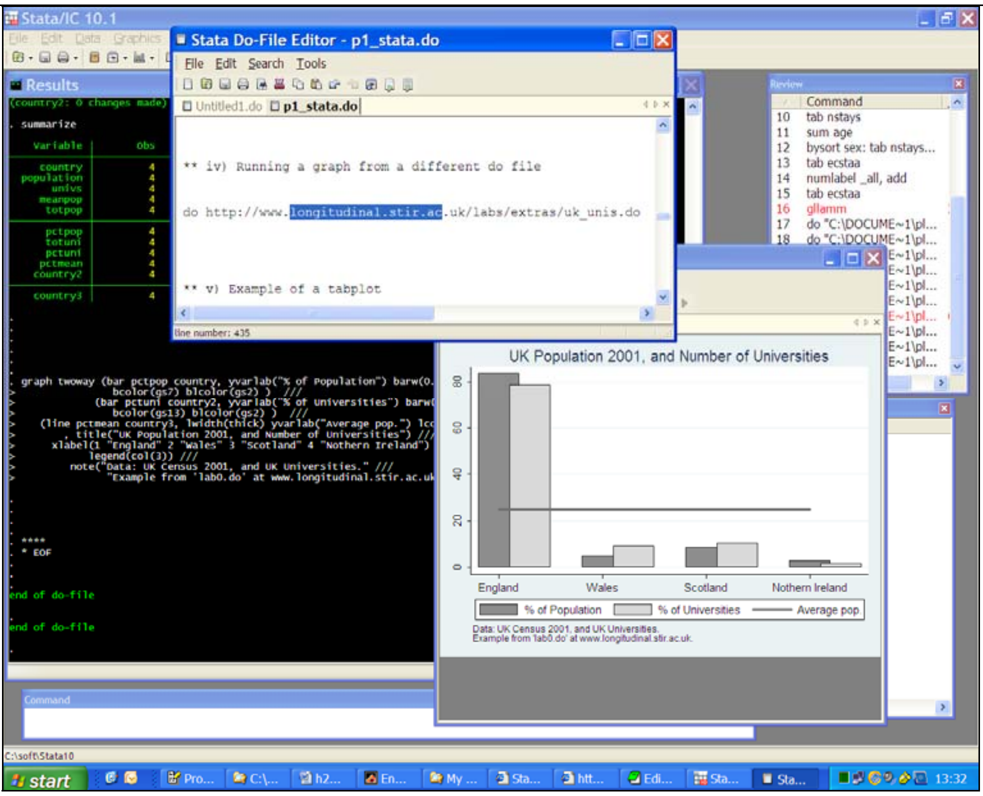


```
** iv) Example of a tabplot

tab qghqe qsex, col
tabplot qghqe qsex

***
* Tabplot uses an extension routine - if it's not already installed in
*  stata copy then:
/*
findit tabplot
net from http://fmwww.bc.edu/RePEc/bocode/t
net describe tabplot
net install tabplot
*/
***
```

```
. net describe tabplot

package tabplot from http:

TITLE
      'TABPLOT': module to

DESCRIPTION/AUTHOR(S)

      tabplot plots a tabl
      graphical form. It i
      categorical  variabl
      histograms.  The wid
      underlap.  tabplot i
       _gaxis.ado, include
      version of tabplot w
      identical functional
      function _glgroup.ad

      Author: Nicholas J. Cox, University of Durham
      Support: email N.J.Cox@durham.ac.uk

      Distribution-Date: 20091125

INSTALLATION FILES                              (type net install tabplot)
      tabplot.ado
      ../_/_glgroup.ado
      tabplot.hlp
      tabplot6.ado
      ../_/_gaxis.ado
      tabplot6.hlp

. net install tabplot
checking tabplot consistency and verifying not already installed...
```

(the exact code needed may depend on which machine you are working on – you may have to define a folder for the installation that you have permission to write to)

We often run subfiles, or define macros or programmes, via calling upon other do files with the 'do' command



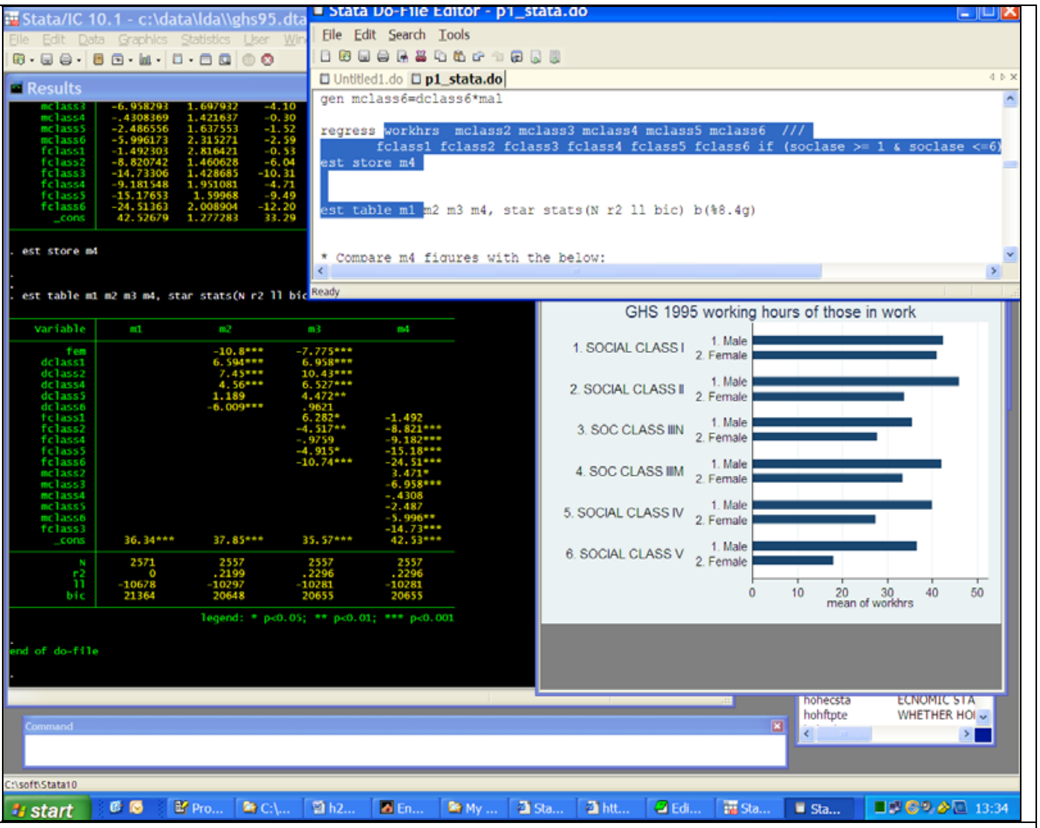
```
** iv) Running a graph from a different do file

do http://www.longitudinal.stir.ac.uk/labs/extras/uk_unis.do


** v) Example of a tabplot
```

```
(country2: 0 changes made)

. summarize

    Variable |       Obs
-------------+----------
     country |         4
  population |         4
       univs |         4
     meanpop |         4
      totpop |         4
-------------+----------
      pctpop |         4
      totuni |         4
      pctuni |         4
     pctmean |         4
    country2 |         4
-------------+----------
    country3 |         4

. graph twoway (bar pctpop country, yvarlab("% of Population") barw(0.
>              bcolor(gs7) blcolor(gs2) )  ///
>           (bar pctuni country2, yvarlab("% of Universities") barw(
>              bcolor(gs13) blcolor(gs2) )  ///
>    (line pctmean country3, lwidth(thick) yvarlab("Average pop.") lc
>         , title("UK Population 2001, and Number of Universities") ///
>      xlabel(1 "England" 2 "Wales" 3 "Scotland" 4 "Nothern Ireland")
>          legend(col(3)) ///
>        note("Data: UK Census 2001, and UK Universities." ///
>             "Example from 'lab0.do' at www.longitudinal.stir.ac.uk

. ****
. * EOF

end of do-file

end of do-file
```

UK Population 2001, and Number of Universities

Data: UK Census 2001, and UK Universities.
Example from 'lab0.do' at www.longitudinal.stir.ac.uk

‘est store’ is a great way to collate and review multiple model results

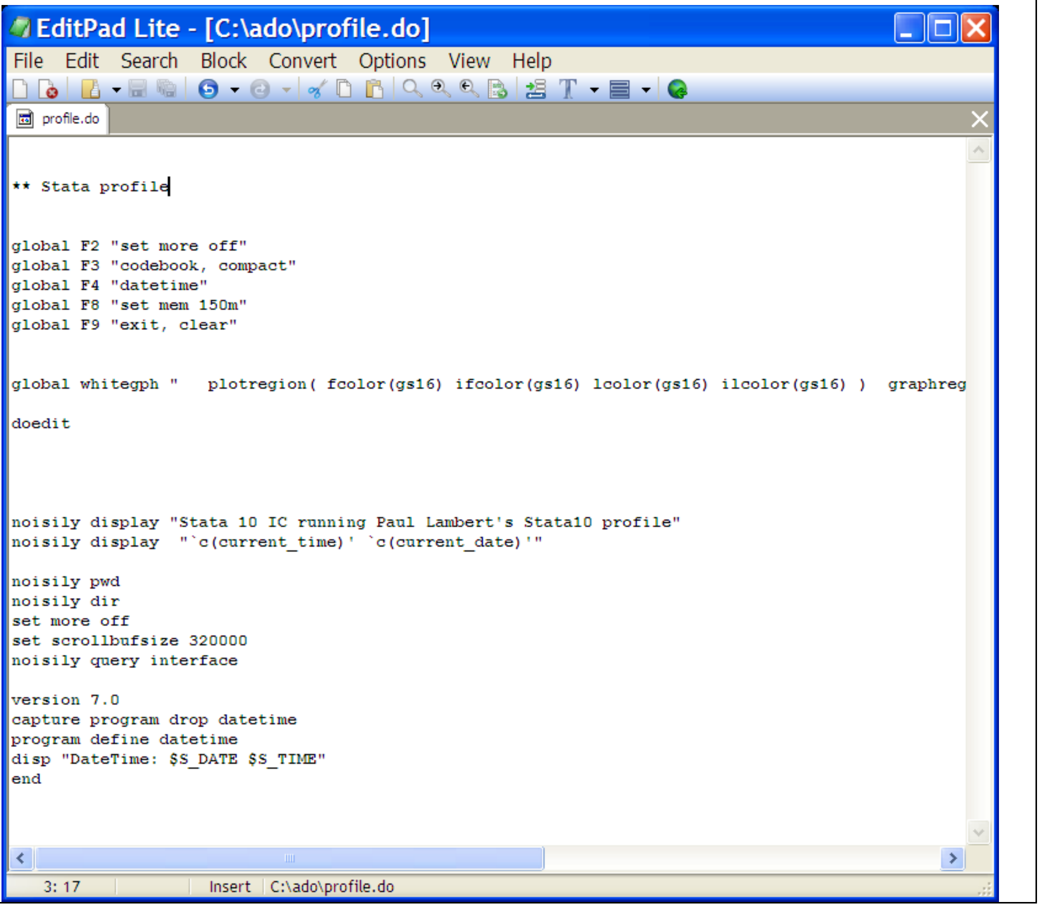


```
     mclass3 |  -6.958293   1.697932    -4.10
     mclass4 |  -.4308369   1.421637    -0.30
     mclass5 |  -2.486556   1.637553    -1.52
     mclass6 |  -5.996173   2.315271    -2.59
     fclass1 |  -1.492303   2.816421    -0.53
     fclass2 |  -8.820742   1.460628    -6.04
     fclass3 |  -14.73306   1.428685   -10.31
     fclass4 |  -9.181548   1.951081    -4.71
     fclass5 |  -15.17653    1.59968    -9.49
     fclass6 |  -24.51363   2.008904   -12.20
       _cons |   42.52679   1.277283    33.29
-------------------------------------------

. est store m4

.
.
. est table m1 m2 m3 m4, star stats(N r2 ll bic

--------------------------------------------------------------------
    Variable |     m1           m2           m3           m4
-------------+------------------------------------------------------
         fem |               -10.8***    -7.775***
     dclass1 |               6.594***     6.958***
     dclass2 |                7.45***     10.43***
     dclass4 |                4.56***     6.527***
     dclass5 |               1.189        4.472**
     dclass6 |              -6.009***      .9621
     fclass1 |                            6.282*      -1.492
     fclass2 |                           -4.517**     -8.821***
     fclass4 |                           -.9759       -9.182***
     fclass5 |                           -4.915*      -15.18***
     fclass6 |                           -10.74***    -24.51***
     mclass2 |                                         3.471*
     mclass3 |                                        -6.958***
     mclass4 |                                        -.4308
     mclass5 |                                        -2.487
     mclass6 |                                        -5.996**
     fclass3 |                                        -14.73***
       _cons |  36.34***     37.85***     35.57***     42.53***
-------------+------------------------------------------------------
           N |     2571         2557         2557         2557
          r2 |        0        .2199        .2296        .2296
          ll |   -10678       -10297       -10281       -10281
         bic |    21364        20648        20655        20655
--------------------------------------------------------------------
                          legend: * p<0.05; ** p<0.01; *** p<0.001

.
end of do-file

.
```

```
gen mclass6=dclass6*mal

regress workhrs  mclass2 mclass3 mclass4 mclass5 mclass6  ///
         fclass1 fclass2 fclass3 fclass4 fclass5 fclass6 if (soclase >= 1 & soclase <=6)
est store m4


est table m1 m2 m3 m4, star stats(N r2 ll bic) b(%8.4g)


* Compare m4 figures with the below:
```

Extension: You can write a ‘.profile’ file to load some starting specifications into your session

For lots more on Stata, see the links and references given above, or the DAMES Node workshops at dames.org.uk



EditPad Lite - [C:\ado\profile.do]
File Edit Search Block Convert Options View Help
profile.do
3: 17
Insert
C:\ado\profile.do

```
** Stata profile


global F2 "set more off"
global F3 "codebook, compact"
global F4 "datetime"
global F8 "set mem 150m"
global F9 "exit, clear"



global whitegph "   plotregion( fcolor(gs16) ifcolor(gs16) lcolor(gs16) ilcolor(gs16) )  graphreg

doedit




noisily display "Stata 10 IC running Paul Lambert's Stata10 profile"
noisily display  "`c(current_time)' `c(current_date)'"

noisily pwd
noisily dir
set more off
set scrollbufsize 320000
noisily query interface

version 7.0
capture program drop datetime
program define datetime
disp "DateTime: $S_DATE $S_TIME"
end
```

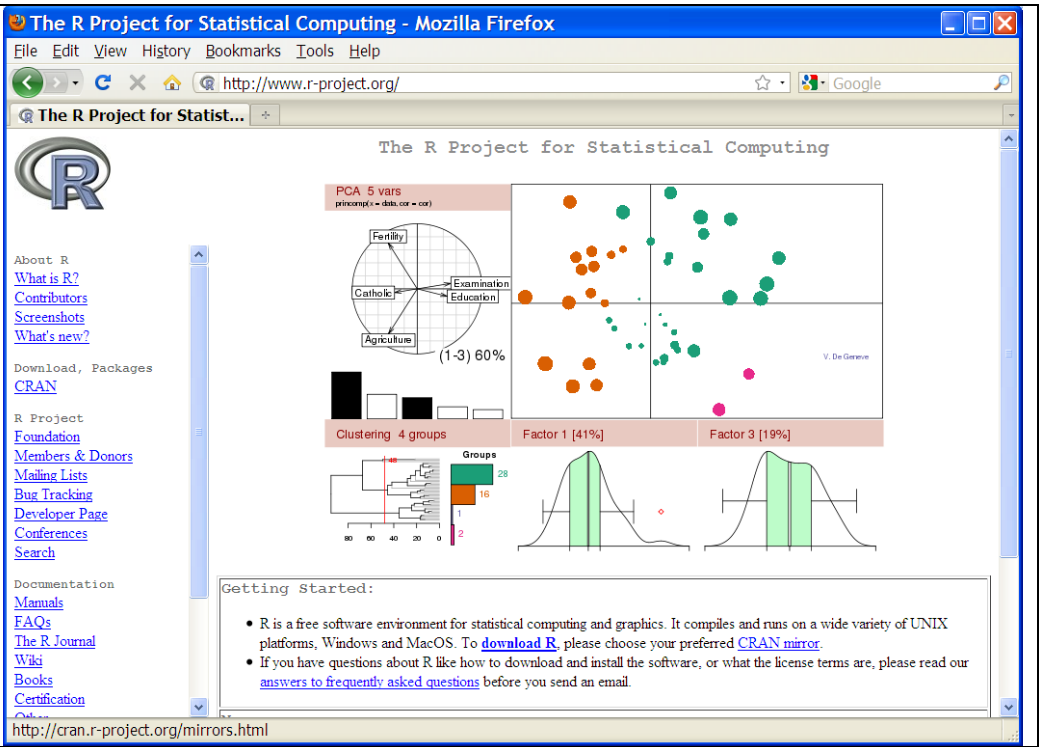
Background: Introducing R

R is a freeware which is a popular tool amongst statisticians and a small community of social science researchers with advanced programming skills. It is an ‘object oriented’ programming language which supports a vast range of statistical analysis routines, and many data management tasks, through its ‘base’ or extension commands. Being ‘object oriented’ is important and means the package appears to behave in a rather different way to the other packages described above. The other packages essentially have one principal quantitative dataset in memory at any one time, plus they store metadata on the matrix and typically some other statistical results in the form other scalars and matrices. In the other packages, commands are automatically applied to the variables of the principal dataset. In R, however, different quantitative datasets (‘data frames’), matrices, vectors, scalars and metadata, are all stored as different ‘objects’, potentially alongside each other. R therefore works by first defining objects, then second performing operations on one or many objects, however defined.

Some researchers are very enthusiastic about R, the common reasons being that it is free and that it often supports exciting statistical models or functions which aren’t available in other packages. However, my perspective is that R isn’t an efficient package for a social survey researcher interested in applied research, as the programming demands to exploit it are very high, and, because it isn’t widely used in applied research, it hasn’t yet developed robust and helpful routines, working interfaces, or documentation standards, to address popular social science data-oriented requirements.

An important concept in R is the ‘extension library’, which is how ‘shortcut’ programmes to undertake many routines are supplied. In fact, you will rarely use R without exploiting extension libraries. The idea here is that R has a ‘base’ set of commands and support, and that many user-contributed programmes have been written in that base language. Those extensions typically provide shortcut routes to useful outcome analyses. A few extension libraries in R are specifically designed to support random effects multilevel model estimation – e.g. the lme package (Bates, 2005; Pinhero & Bates, 2000).

R is installed as freeware and since it is frequently updated it is wise to regularly revisit the distribution site and re-download

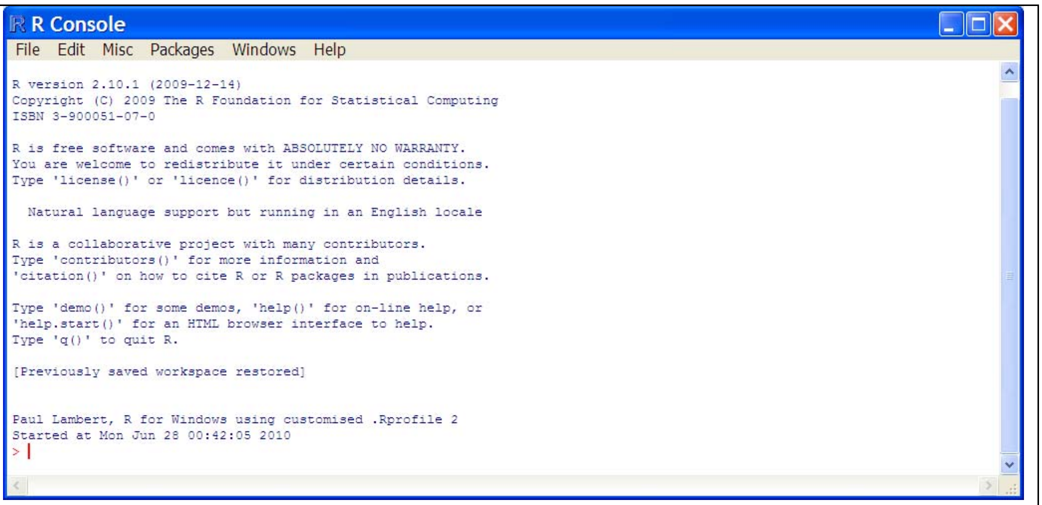


The R Project for Statistical Computing

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

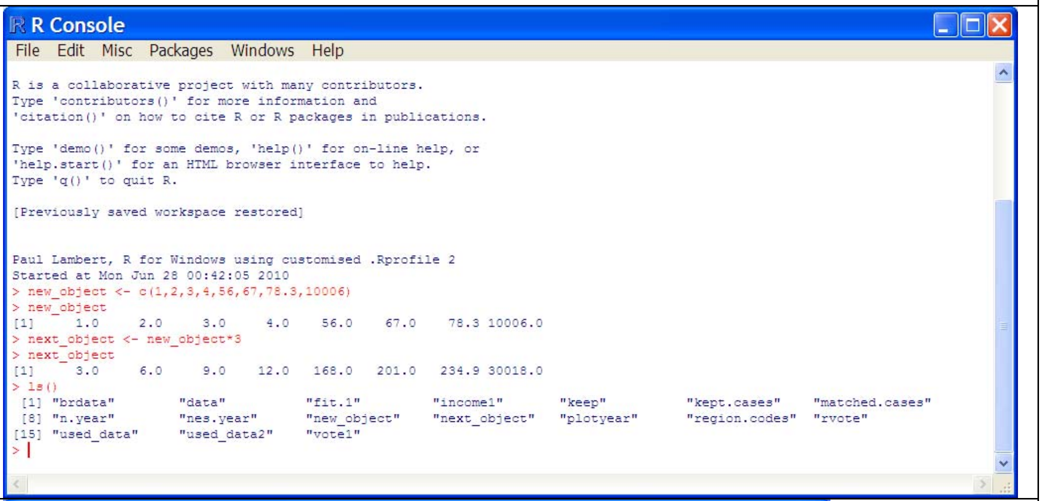
When you open R, you will see something like this ‘R console’



```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]


Paul Lambert, R for Windows using customised .Rprofile 2
Started at Mon Jun 28 00:42:05 2010
> 
```

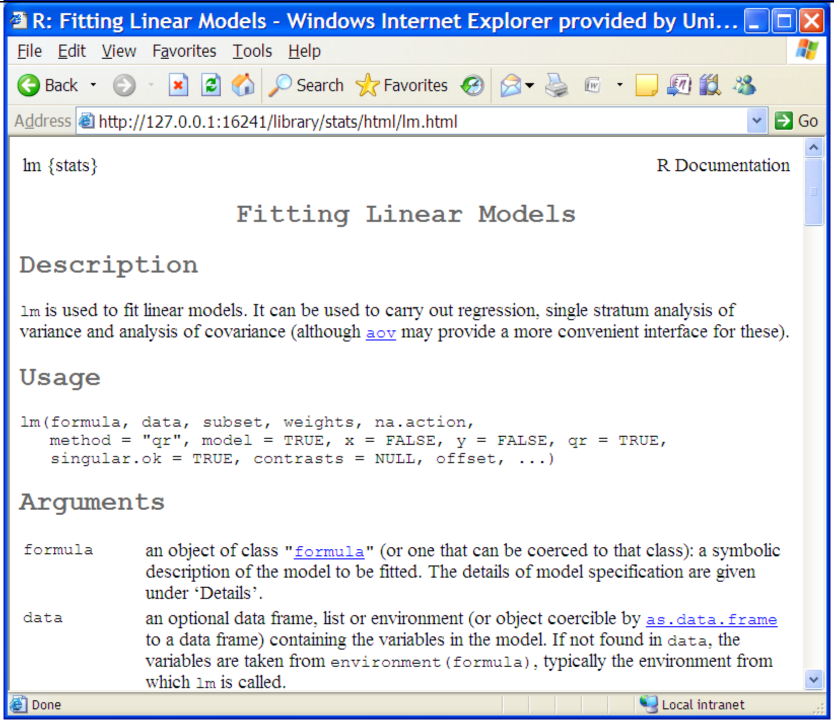
(on my machine I use a ‘.rprofile’ settings file so my starting display is marginally different to the default)

The first few lines show me defining a new object (a short vector) and listing the objects in current memory.



```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]


Paul Lambert, R for Windows using customised .Rprofile 2
Started at Mon Jun 28 00:42:05 2010
> new_object <- c(1,2,3,4,56,67,78.3,10006)
> new_object
[1]     1.0     2.0     3.0     4.0    56.0    67.0    78.3 10006.0
> next_object <- new_object*3
> next_object
[1]     3.0     6.0     9.0    12.0   168.0   201.0   234.9 30018.0
> ls()
 [1] "brdata"        "data"          "fit.1"         "income1"       "keep"          "kept.cases"    "matched.cases"
 [8] "n.year"        "nes.year"      "new_object"    "next_object"   "plotyear"      "region.codes"  "rvote"
[15] "used_data"     "used_data2"    "vote1"
> 
```

R’s basic help functions point to webpages.


R: Fitting Linear Models - Windows Internet Explorer provided by Uni...
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://127.0.0.1:16241/library/stats/html/lm.html
Go
Done
Local intranet

lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

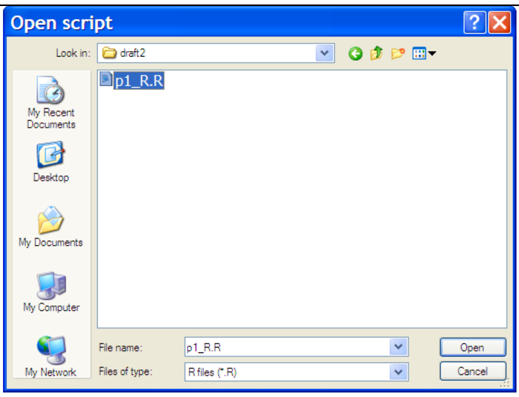
Arguments

`formula` an object of class "`formula`" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under ‘Details’.

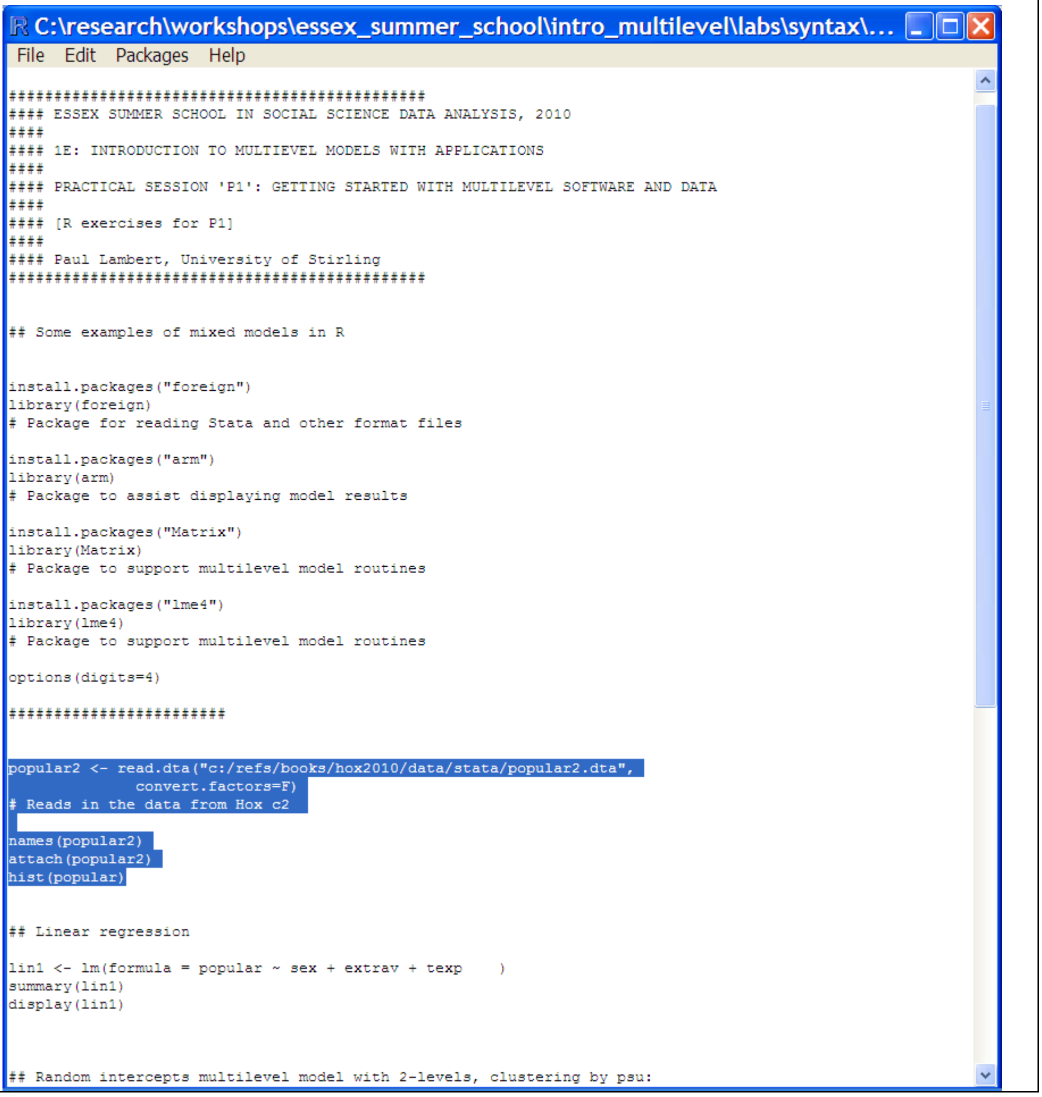
`data` an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `lm` is called.

The general help pages mostly have generic information, and are not in general provided with worked examples. Many R users get their help from other online sources, e.g. http://www.statmethods.net/

In general, with R, the first thing you should do is ask to open a new or existing script and work from that. Scripts in R work in a similar way to a syntax file in Stata or SPSS – highlight a line or lines, and press ‘ctrl+R’.

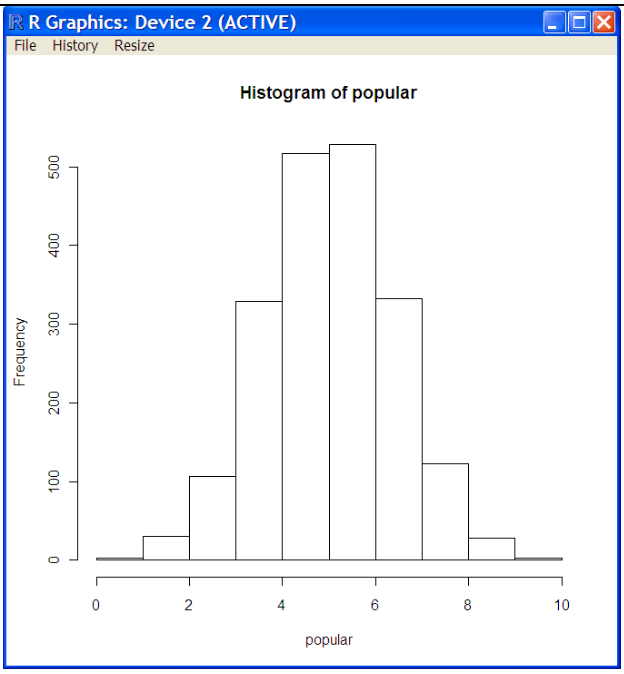


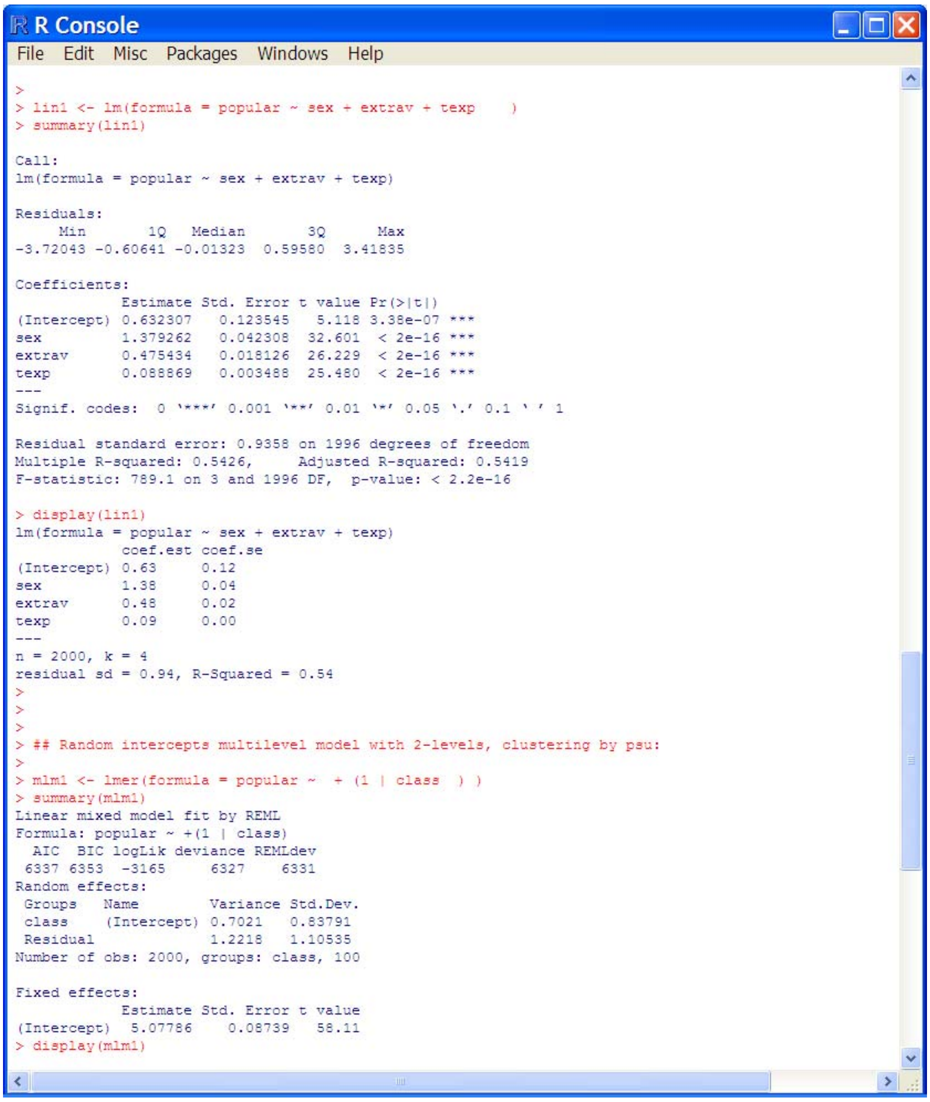
File name: p1_R.R
Files of type: R files (*.R)


C:\research\workshops\essex_summer_school\intro_multilevel\labs\syntax\...
File
Edit
Packages
Help

```
###############################################
#### ESSEX SUMMER SCHOOL IN SOCIAL SCIENCE DATA ANALYSIS, 2010
####
#### 1E: INTRODUCTION TO MULTIEVEL MODELS WITH APPLICATIONS
####
#### PRACTICAL SESSION 'P1': GETTING STARTED WITH MULTILEVEL SOFTWARE AND DATA
####
#### [R exercises for P1]
####
#### Paul Lambert, University of Stirling
###############################################


## Some examples of mixed models in R


install.packages("foreign")
library(foreign)
# Package for reading Stata and other format files

install.packages("arm")
library(arm)
# Package to assist displaying model results

install.packages("Matrix")
library(Matrix)
# Package to support multilevel model routines

install.packages("lme4")
library(lme4)
# Package to support multilevel model routines

options(digits=4)

########################


popular2 <- read.dta("c:/refs/books/hox2010/data/stata/popular2.dta",
              convert.factors=F)
# Reads in the data from Hox c2

names(popular2)
attach(popular2)
hist(popular)


## Linear regression

lin1 <- lm(formula = popular ~ sex + extrav + texp     )
summary(lin1)
display(lin1)



## Random intercepts multilevel model with 2-levels, clustering by psu:
```

After running commands, output is sent either to the main console or a separate graphics window





```
>
> lin1 <- lm(formula = popular ~ sex + extrav + texp    )
> summary(lin1)

Call:
lm(formula = popular ~ sex + extrav + texp)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.72043 -0.60641 -0.01323  0.59580  3.41835 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.632307   0.123545   5.118 3.38e-07 ***
sex         1.379262   0.042308  32.601  < 2e-16 ***
extrav      0.475434   0.018126  26.229  < 2e-16 ***
texp        0.088869   0.003488  25.480  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.9358 on 1996 degrees of freedom
Multiple R-squared: 0.5426,     Adjusted R-squared: 0.5419 
F-statistic: 789.1 on 3 and 1996 DF,  p-value: < 2.2e-16 

> display(lin1)
lm(formula = popular ~ sex + extrav + texp)
            coef.est coef.se
(Intercept) 0.63     0.12   
sex         1.38     0.04   
extrav      0.48     0.02   
texp        0.09     0.00   
---
n = 2000, k = 4
residual sd = 0.94, R-Squared = 0.54
>
>
>
> ## Random intercepts multilevel model with 2-levels, clustering by psu:
>
> mlm1 <- lmer(formula = popular ~  + (1 | class  ) )
> summary(mlm1)
Linear mixed model fit by REML 
Formula: popular ~ +(1 | class) 
  AIC  BIC logLik deviance REMLdev
 6337 6353  -3165     6327    6331
Random effects:
 Groups   Name        Variance Std.Dev.
 class    (Intercept) 0.7021   0.83791 
 Residual             1.2218   1.10535 
Number of obs: 2000, groups: class, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept)  5.07786    0.08739   58.11
> display(mlm1)
```

Lab 1: Pajek exercises

Pajek is a social network analysis software package which has been developed by Vladimar Bataglj and Andrej Mrvar from the University of Ljubljana (hence the name, Slovenian for 'spider'). Some argue the software is not as advanced as competing generalist software such as UCINET or the Social Network package within R, but it has the following benefits:

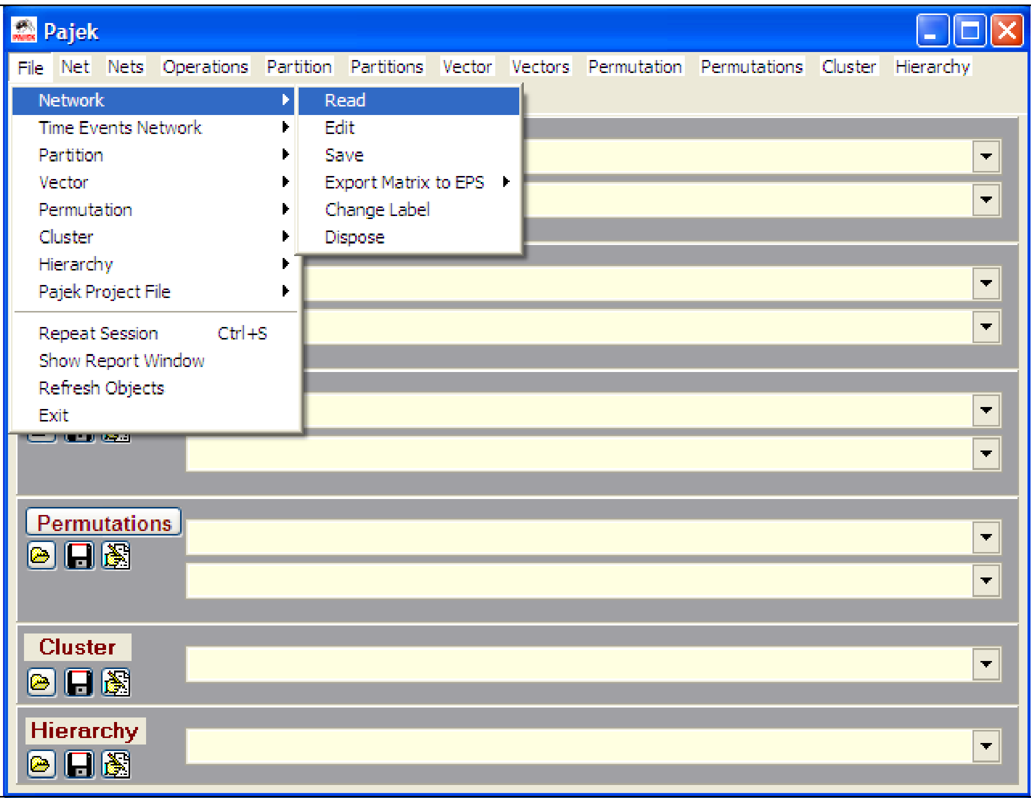
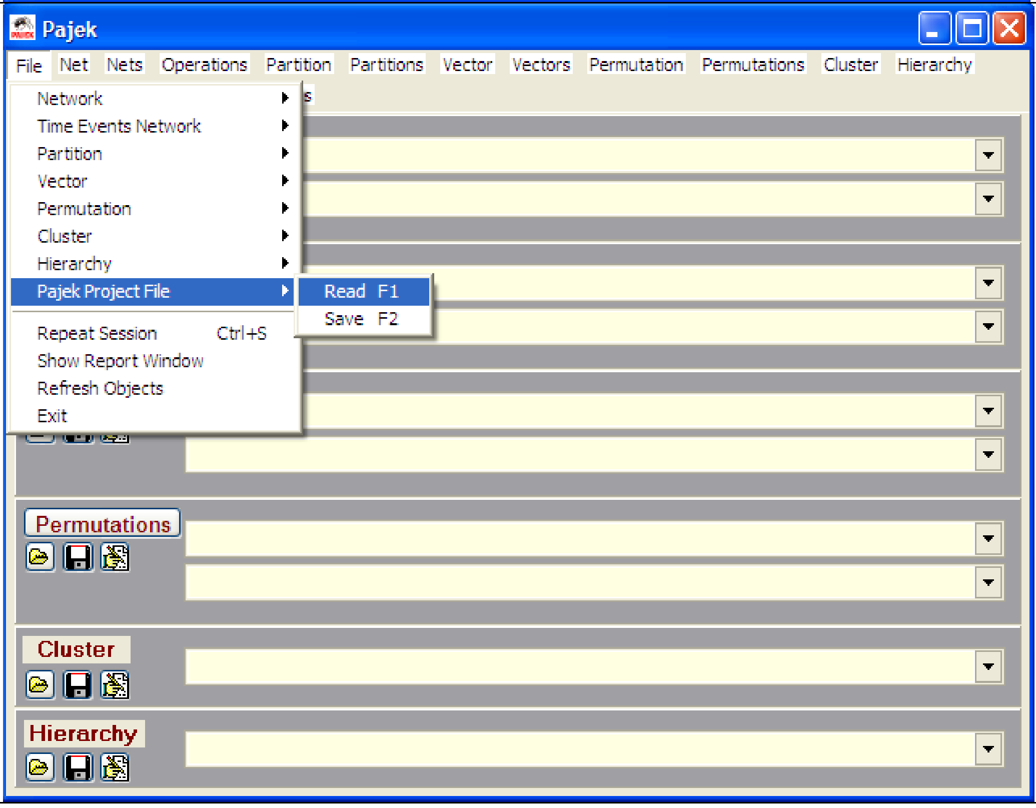
- It is simple and free to install (for non-commercial use)
- It has easy methods for importing data
- It is simple to use and covers most common network commands
- It is more robust than other packages for dealing with very large datasets.

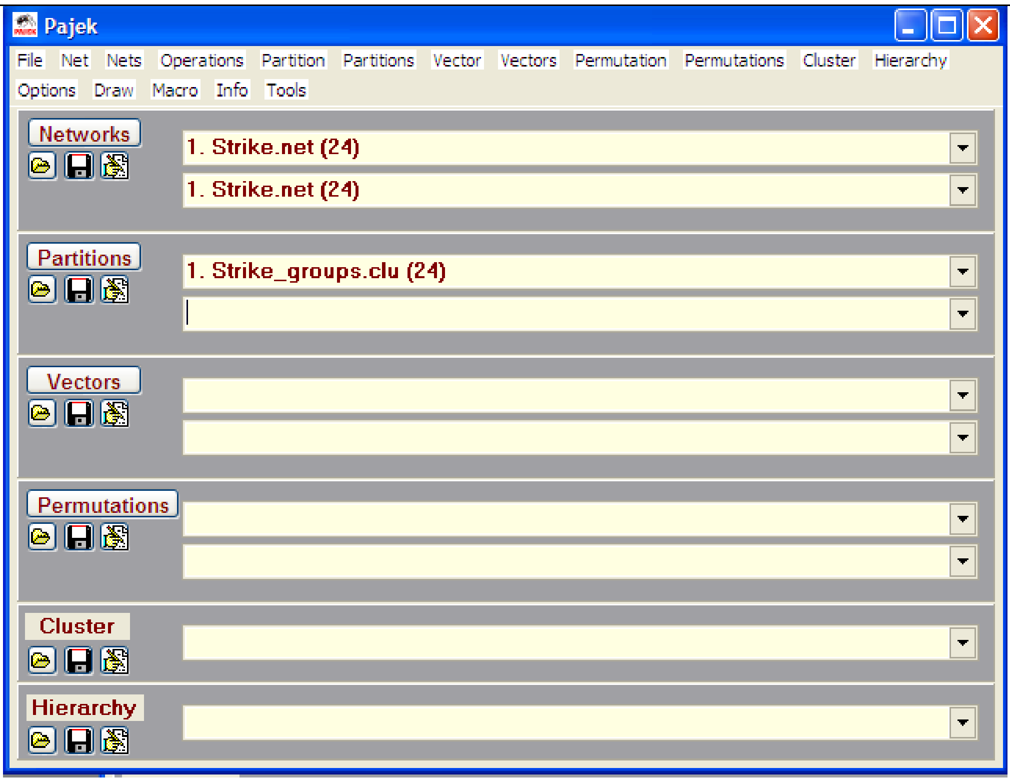
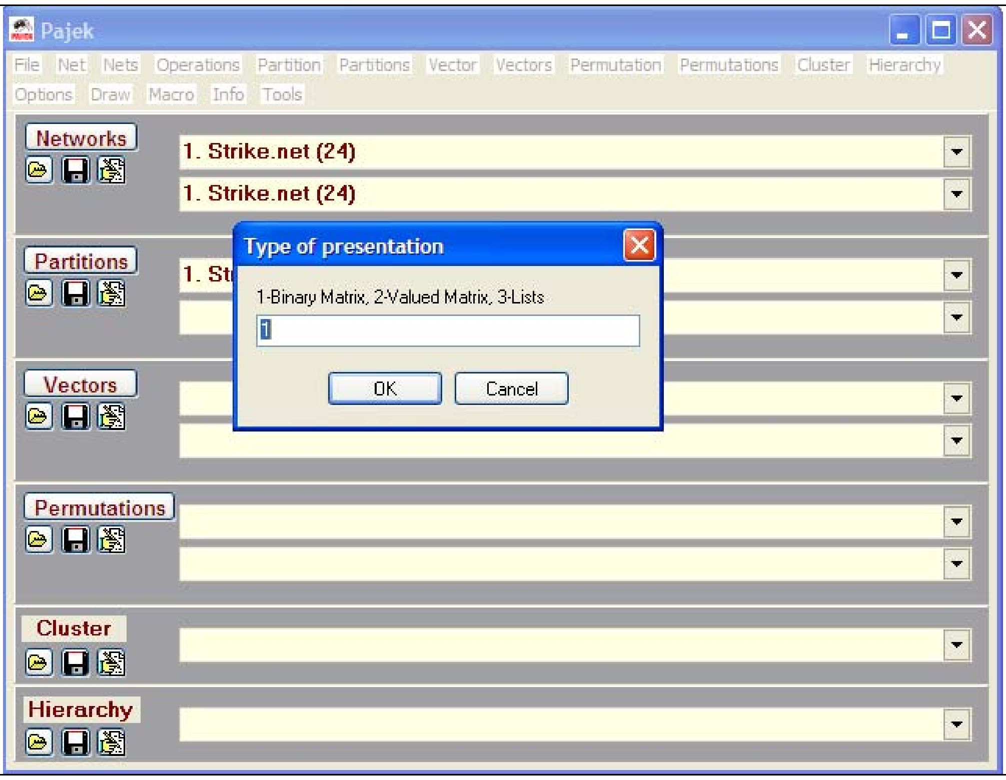
There are several limitations with Pajek. Unlike R there is a requirement to use drop-down menus meaning it is not possible to run syntax files (although all processes can be saved). It cannot perform some of the emerging analyses such as random graph models, forcing users to use SIENA, PNET or other specialist package. However, Pajek performs the basic elements of network analysis in a very user-friendly manner, which makes it the ideal package for people unfamiliar with network methods (and, therefore, less likely to require the more advanced methods central to other packages). It retains sufficient sophistication to be utilised by many experienced researchers. Most other SNA packages (for instance, UCINET, Siena and PNET) have strong links to Pajek and enable data to be readily imported.

Pajek also benefits from having a comprehensive book providing good examples of how to use the software (de Nooy, W., Mrvar, A., & Batagelj. V. (2012) *Exploratory Social Network Analysis with Pajek*. Cambridge: Cambridge University Press. 2nd edition). This book is an excellent introduction to both Pajek and SNA more generally, providing an overview of each method described and working through examples which convey not only how to perform such analysis but also spells out the benefits of each technique.

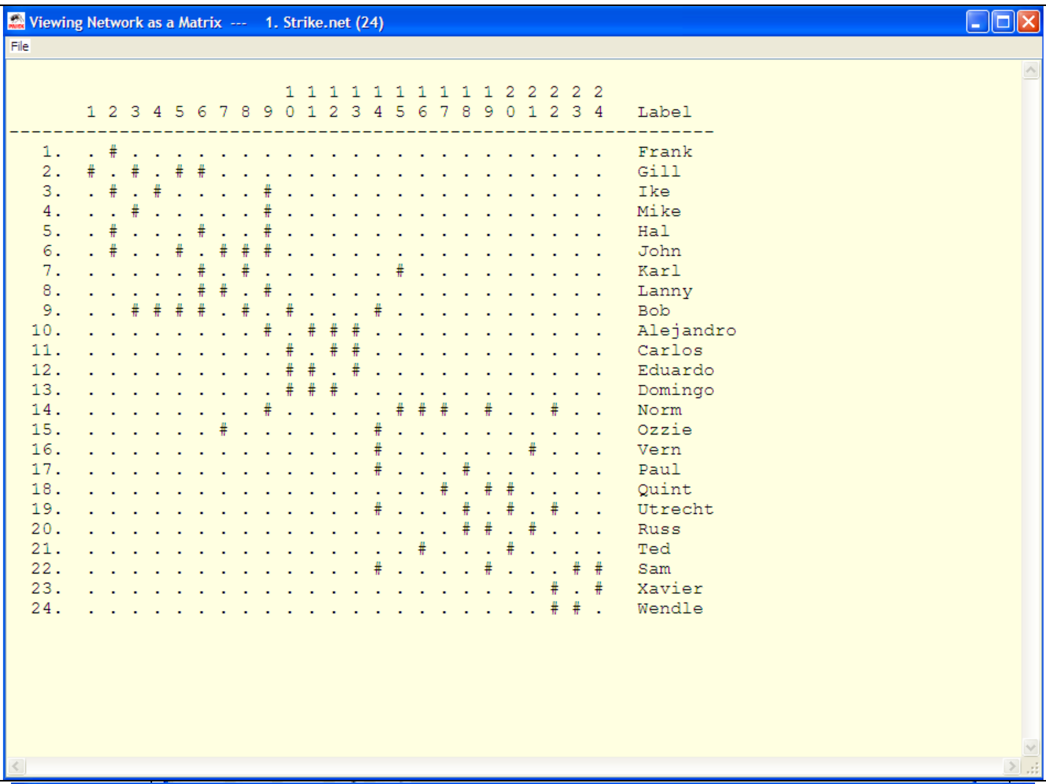
The manual, however, is less helpful if you're unsure of how to use Pajek. It provides detailed information on Pajek but in a manner which assumes prior understanding of the operation. Therefore, it provides many useful resources for experienced users (such as the default colours for vertices and labels for triad censuses), providing in-depth knowledge of the finer points of the package, but the manual is more helpful for advancing your familiarity with the software.

Pajek frequently updates the software (usually fixing tiny glitches, adding new procedures or speeding up processes) so given the ease in installing it's often worth checking you're using the most up-to-date version before starting a piece of work. There is also a dedicated e-mail list which provides rapid answers to complex questions. A new development is the Pajek-XXL programme, which replicates Pajek but operates much faster on huge datasets (tens of thousands of nodes).

When you first open Pajek you need to import your data. You can import in individual networks one at a time using either the drop down menus, or the 'file open' button. You can then open any saved partitions, vectors and so on in the same way.	 Pajek File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy Network Time Events Network Partition Vector Permutation Cluster Hierarchy Pajek Project File Repeat Session Ctrl+S Show Report Window Refresh Objects Exit Read Edit Save Export Matrix to EPS Change Label Dispose Permutations Cluster Hierarchy
Alternatively, you could open a 'Pajek Project File' which imports all saved data relevant to the piece of work.	 Pajek File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy Network Time Events Network Partition Vector Permutation Cluster Hierarchy Pajek Project File Repeat Session Ctrl+S Show Report Window Refresh Objects Exit Read F1 Save F2 Permutations Cluster Hierarchy

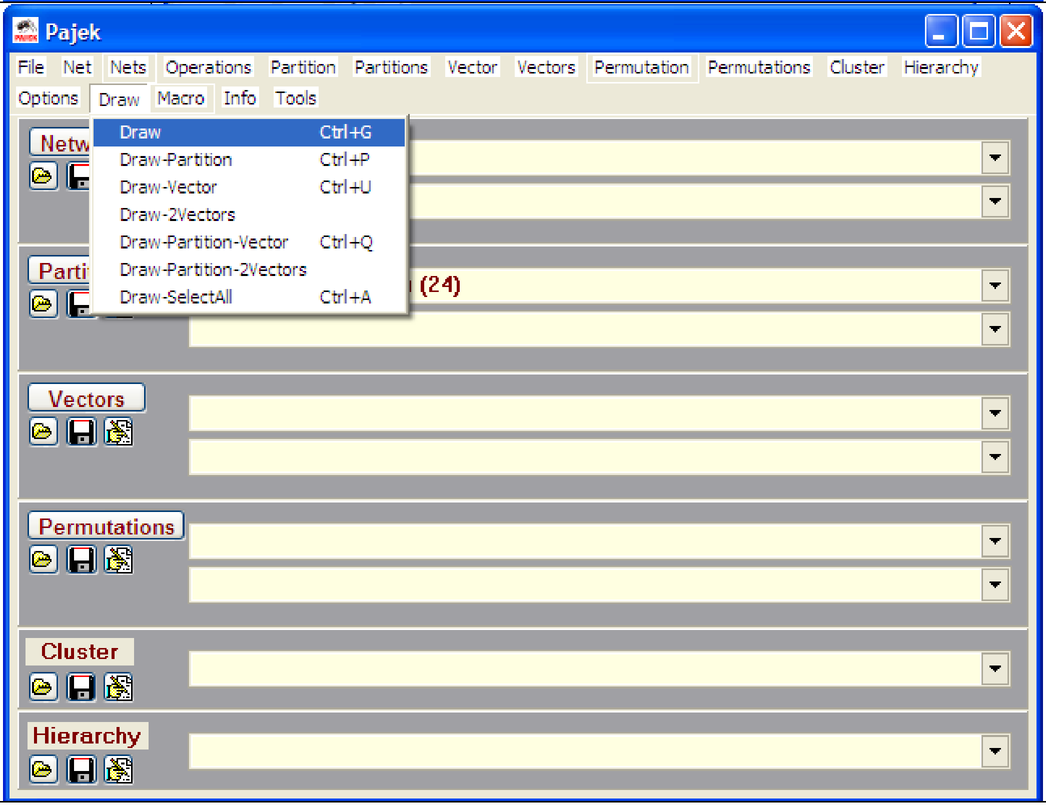
This provides the data useful for the research. We can simply add any new networks, partitions etc. to this file by simply opening additional individual data as before. However, opening a new 'Pajek Project File' will remove all the data from the package. Multiple files can be open at once in Pajek. The labels for each file have the structure for a number for the item in the drop list, the file name (or method of construction) and finally the number of cases in brackets.	 Pajek File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy Options Draw Macro Info Tools Networks Partitions Vectors Permutations Cluster Hierarchy 1. Strike.net (24) 1. Strike.net (24) 1. Strike_groups.clu (24)
We can view the matrix which is providing all the information Pajek needs to operate. To do this, click on the actual name on the network in the yellow drop-drop section of the networks tab (the line marked "1. Strike.net (24)" in this example). This produces a dialogue box enabling us to see whether a binary matrix (# marks a link), a valued list (showing the numeric value of the link) or a list (a list of the ties which are formed).	 Pajek File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy Options Draw Macro Info Tools Networks Partitions Vectors Permutations Cluster Hierarchy OK Cancel 1. Strike.net (24) 1. Strike.net (24) 1. St Type of presentation 1-Binary Matrix, 2-Valued Matrix, 3-Lists 1

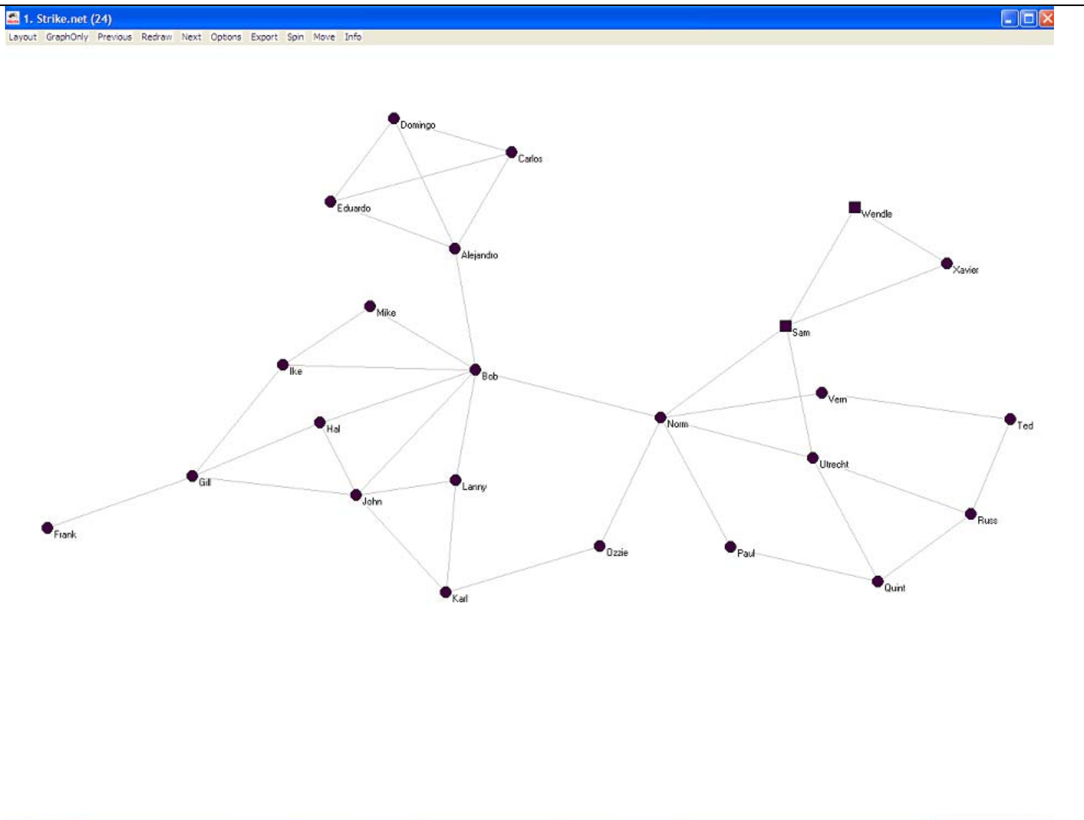
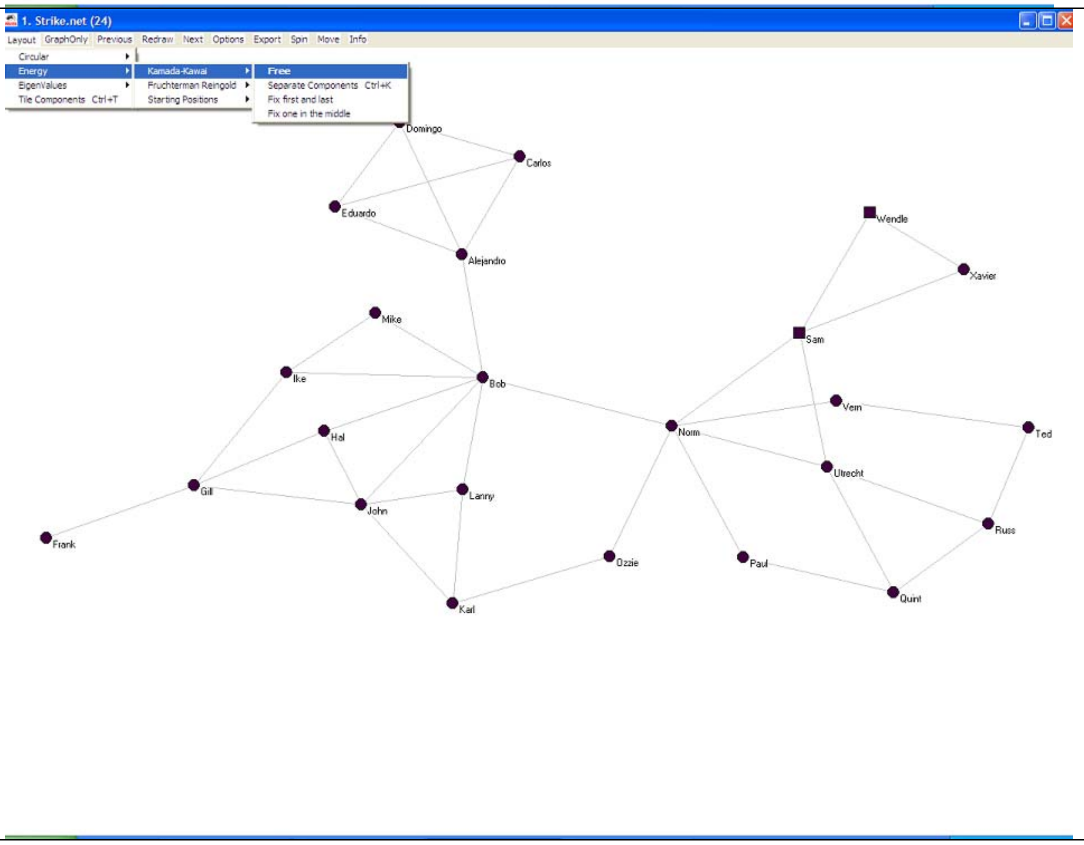
As this is non-numeric data, we have selected a binary matrix (1). This shows the presence of links between actors. The actors in the rows (which are labelled) send a link to the actors in the columns (which are in the same order). In this case, the links must be reciprocated (i.e., if A speaks to B, B speaks to A). This is not always the case (i.e., if A likes B, that doesn't necessarily mean B likes A).



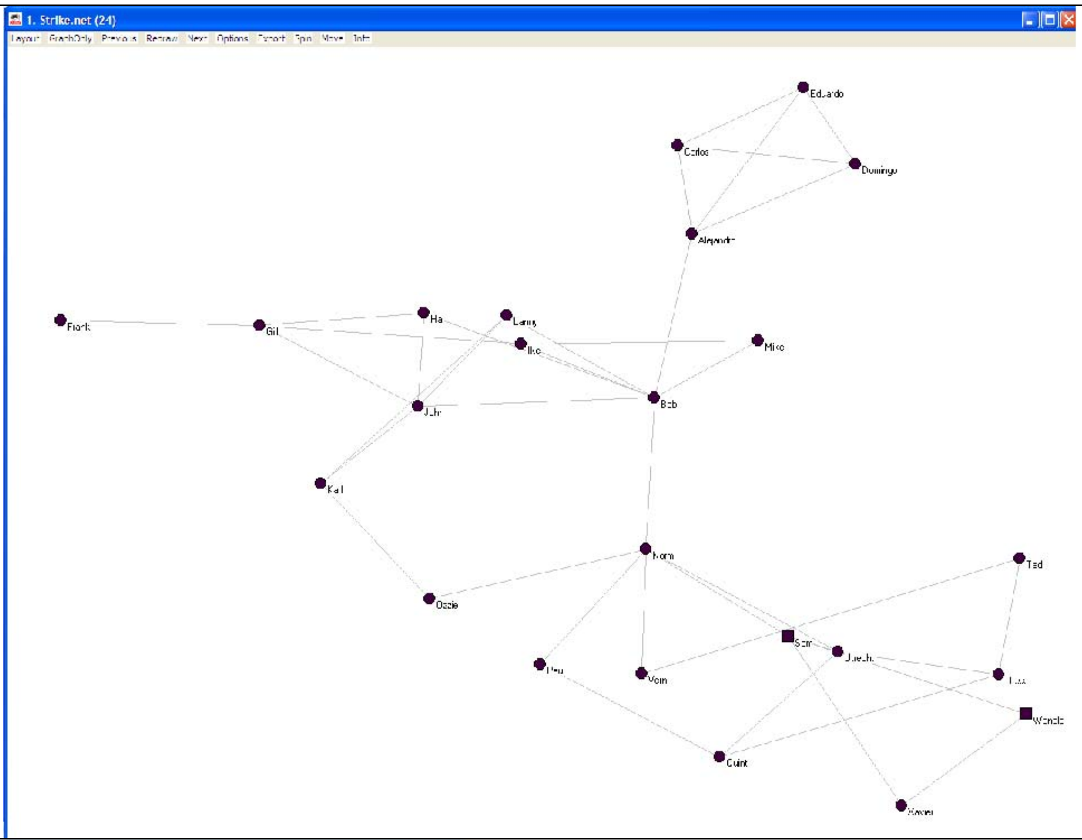
```
                          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2
        1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4   Label
-----------------------------------------------------------------
   1.   . # . . . . . . . . . . . . . . . . . . . . . .   Frank
   2.   # . # . # # . . . . . . . . . . . . . . . . . .   Gill
   3.   . # . # . . . . # . . . . . . . . . . . . . . .   Ike
   4.   . . # . . . . . # . . . . . . . . . . . . . . .   Mike
   5.   . # . . . # . . # . . . . . . . . . . . . . . .   Hal
   6.   . # . . # . # # # . . . . . . . . . . . . . . .   John
   7.   . . . . . # . # . . . . . . # . . . . . . . . .   Karl
   8.   . . . . . # # . # . . . . . . . . . . . . . . .   Lanny
   9.   . . # # # # . # . # . . . # . . . . . . . . . .   Bob
  10.   . . . . . . . . # . # # # . . . . . . . . . . .   Alejandro
  11.   . . . . . . . . . # . # # . . . . . . . . . . .   Carlos
  12.   . . . . . . . . . # # . # . . . . . . . . . . .   Eduardo
  13.   . . . . . . . . . # # # . . . . . . . . . . . .   Domingo
  14.   . . . . . . . . # . . . . . # # # . # . . # . .   Norm
  15.   . . . . . . # . . . . . . # . . . . . . . . . .   Ozzie
  16.   . . . . . . . . . . . . . # . . . . . . # . . .   Vern
  17.   . . . . . . . . . . . . . # . . . # . . . . . .   Paul
  18.   . . . . . . . . . . . . . . . . # . # # . . . .   Quint
  19.   . . . . . . . . . . . . . # . . . # . # . # . .   Utrecht
  20.   . . . . . . . . . . . . . . . . . # # . # . . .   Russ
  21.   . . . . . . . . . . . . . . . # . . . # . . . .   Ted
  22.   . . . . . . . . . . . . . # . . . . # . . . # #   Sam
  23.   . . . . . . . . . . . . . . . . . . . . . # . #   Xavier
  24.   . . . . . . . . . . . . . . . . . . . . . # # .   Wendle
```

You can also visual the data as a network. Firstly, we will look at the basic structure of the network. This can be done through either using the 'DRAW' drop-down menu, or CTRL+G.

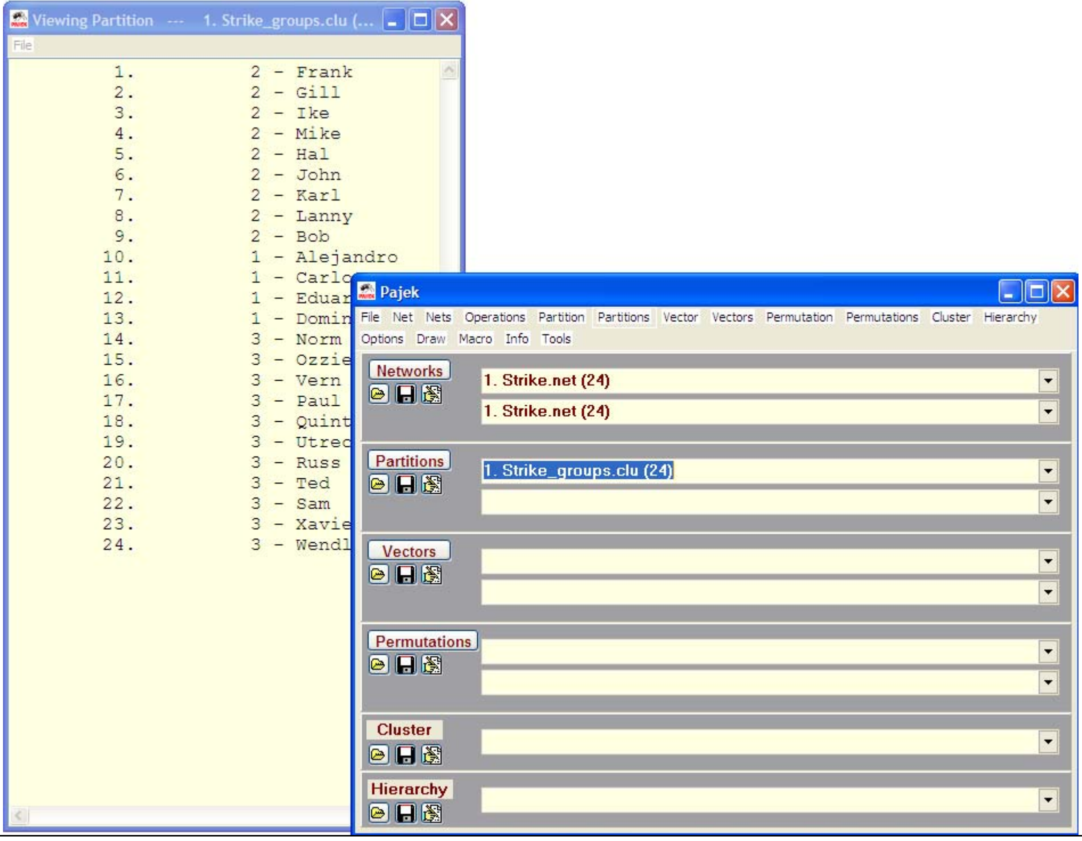


This produces a basic visualisation (in this case, who speaks with whom). Pajek places the nodes in what it believes is the most appropriate position. This is based upon a calculation which can be different each time. Sometimes the graph is displayed nicely the first time, as occurred for me in this example. The graph will be positioned slightly differently each time you generate it.	 1. Strike.net (24) Layout GraphOnly Previous Redraw Next Options Export Spin Move Info Domingo Carlos Eduardo Alejandro Wendle Xavier Mike Sam Ike Bob Vern Ted Hal Norm Utrecht Gill Lanny John Russ Frank Ozzie Paul Quint Karl
Using the layout drop-down menu, you can use ‘energy’, then ‘Kamada-Kawai’, then ‘separate components’ (or CTRL+K) to rearrange the network using the same criteria as originally if you want a slightly different visualisation. The links will remain the same, but the nodes will be positioned slightly differently as it tries again to produce the most appropriate layout.	 1. Strike.net (24) Layout GraphOnly Previous Redraw Next Options Export Spin Move Info Circular Energy EigenValues Tile Components Ctrl+T Kamada-Kawai Fruchterman Reingold Starting Positions Free Separate Components Ctrl+K Fix first and last Fix one in the middle Domingo Carlos Eduardo Alejandro Wendle Xavier Mike Sam Ike Bob Vern Hal Norm Ted Utrecht Gill Lanny John Russ Frank Ozzie Paul Quint Karl

This is an example of how the network can look if repositioned using the Kamada-Kawai equation. Substantially there is no difference between the networks, but the first was spaced more nicely and made the links easier to read. Sometimes the layout is not optimal, therefore it is always useful to press CTRL+K a few times to see a few representations of the data.



This data thus far has not distinguished between actors.. We might have some characteristics of the actors we wish to group them by. This project file contains a partition. Clicking on the name of the partition (highlighted on the Pajek window), brings up a list of the partitions and labels. We can see, therefore, these individuals are split into three groups.


Viewing Partition --- 1. Strike_groups.clu (...
File
Pajek
File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy
Options Draw Macro Info Tools
Networks
Partitions
Vectors
Permutations
Cluster
Hierarchy

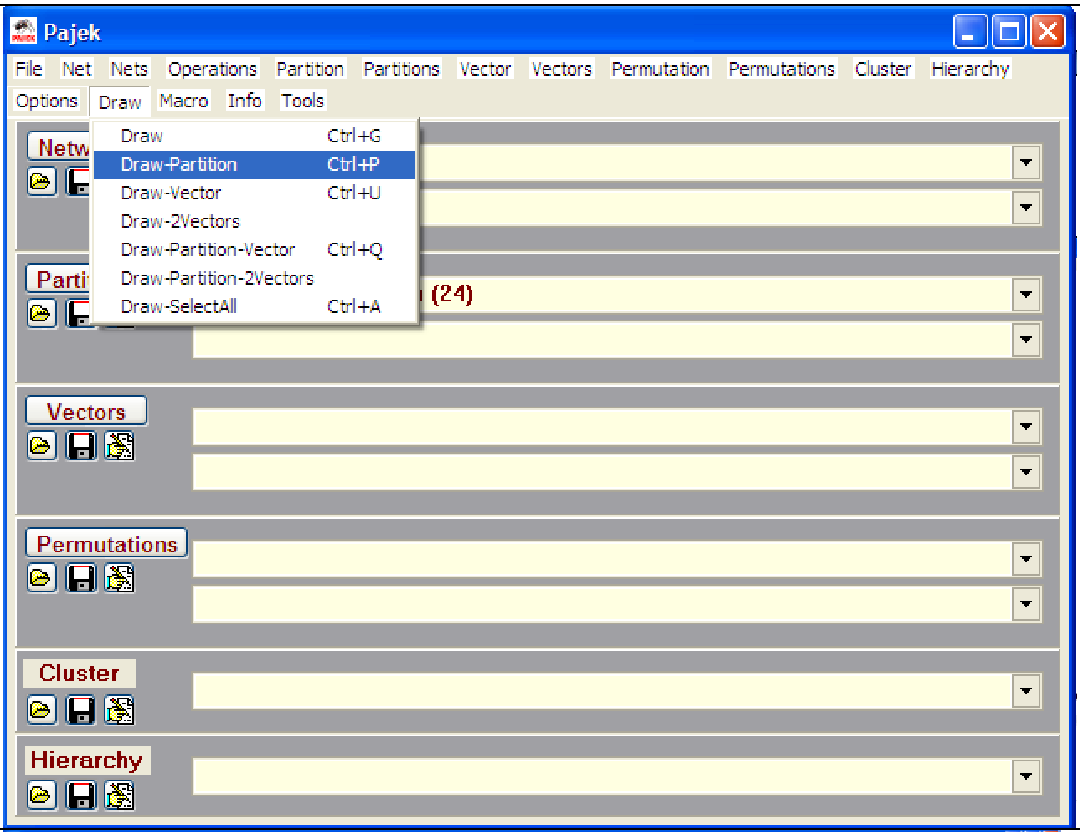
```
  1.         2 - Frank
  2.         2 - Gill
  3.         2 - Ike
  4.         2 - Mike
  5.         2 - Hal
  6.         2 - John
  7.         2 - Karl
  8.         2 - Lanny
  9.         2 - Bob
 10.         1 - Alejandro
 11.         1 - Carlo
 12.         1 - Eduar
 13.         1 - Domin
 14.         3 - Norm
 15.         3 - Ozzie
 16.         3 - Vern
 17.         3 - Paul
 18.         3 - Quint
 19.         3 - Utrec
 20.         3 - Russ
 21.         3 - Ted
 22.         3 - Sam
 23.         3 - Xavie
 24.         3 - Wendl
```

Networks: 1. Strike.net (24)

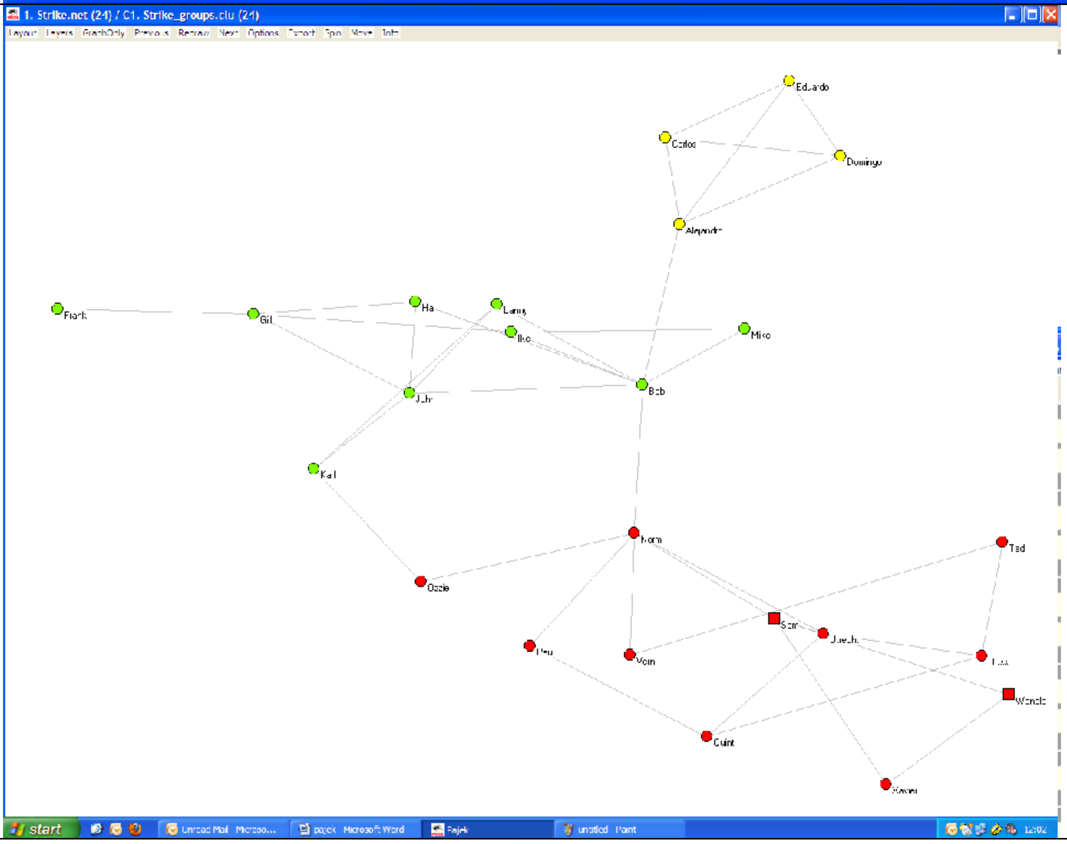
1. Strike.net (24)

Partitions: 1. Strike_groups.clu (24)

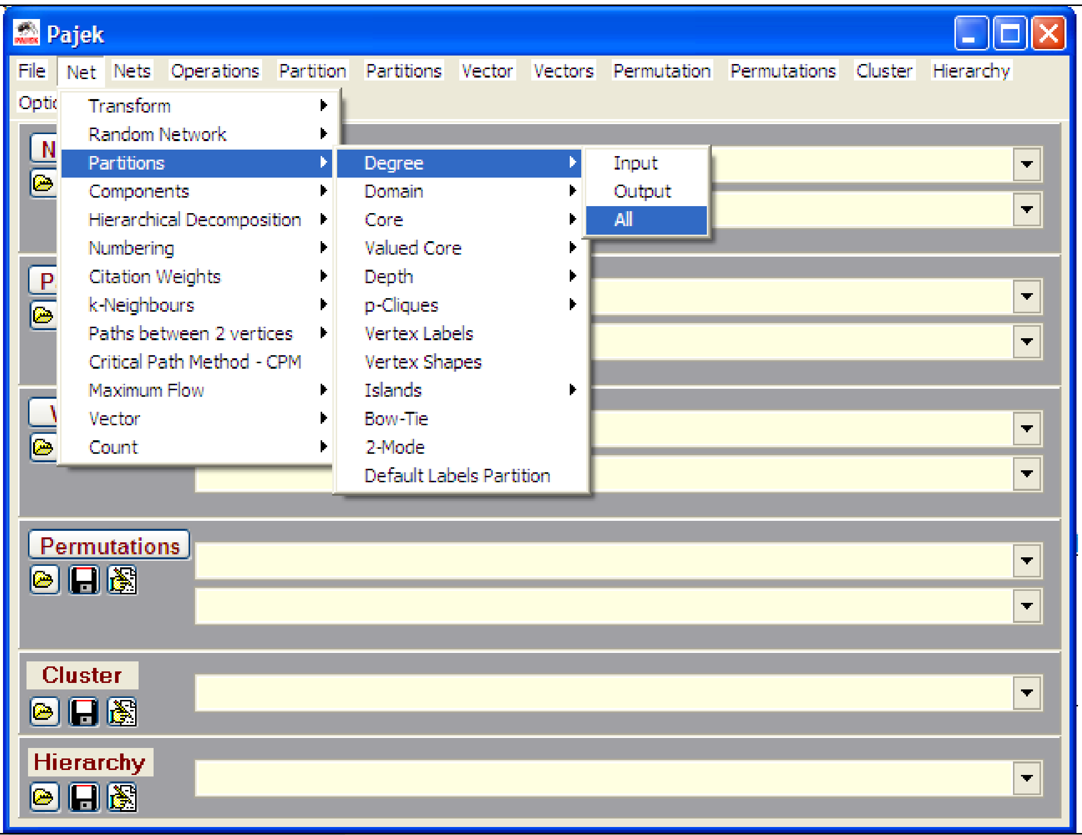
Therefore, we may wish to know how well the grouping within this partition corresponds with the network we've created. We can again go to the draw menu, but select 'draw-partition' (or CTRL+P). This shades the nodes dependent on their partition.



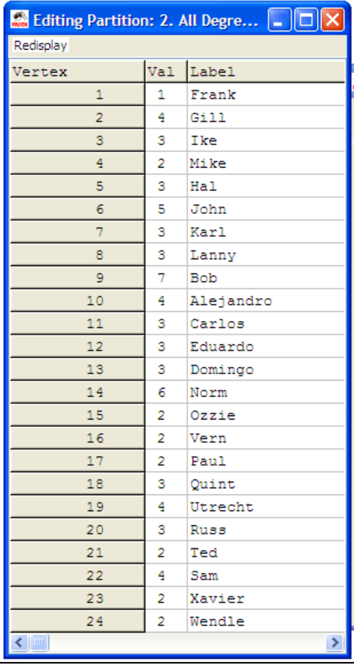
This graph shows we can explain very well the network positions by the characteristics used in creating the partition. In this example, the upper grouping are Spanish speakers, the middle grouping are younger English speakers and the lower group older English speakers. The hypothesis that age and language influences communication networks is clearly supported here.



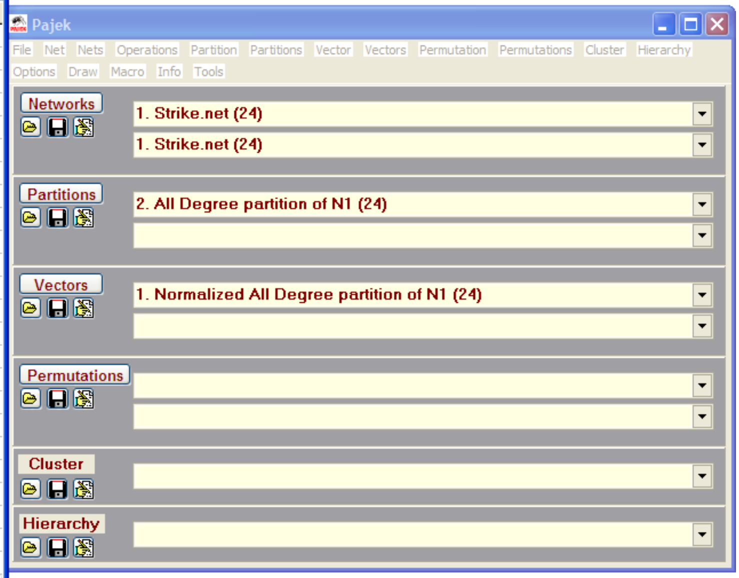
Visualisation is not all that Pajek can do. We might want to gather some statistics about our network. We can see who the most central actors are. We can do this by creating a new partition of the degree for each actor. *(Note: we have the option of input (incoming ties), output (outgoing ties) or all – as ties must be reciprocated within this network there are no differences, but if they were not reciprocated by design the choice would be important.)*



This creates a new partition. Clicking on the name of the new partition (2. All Degree partition of N1 (24)) enables us to see the number of ties each actor has. If we are interested in degree centrality (the number of people each individual speaks to in this case) we can get the information here.



Vertex	Val	Label
1	1	Frank
2	4	Gill
3	3	Ike
4	2	Mike
5	3	Hal
6	5	John
7	3	Karl
8	3	Lanny
9	7	Bob
10	4	Alejandro
11	3	Carlos
12	3	Eduardo
13	3	Domingo
14	6	Norm
15	2	Ozzie
16	2	Vern
17	2	Paul
18	3	Quint
19	4	Utrecht
20	3	Russ
21	2	Ted
22	4	Sam
23	2	Xavier
24	2	Wendle

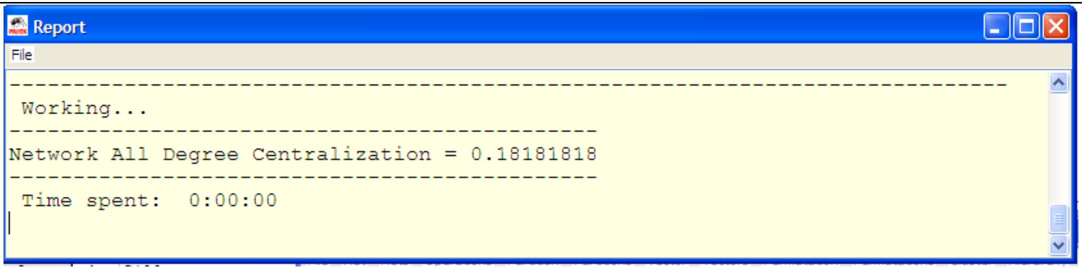


Networks: 1. Strike.net (24)
1. Strike.net (24)

Partitions: 2. All Degree partition of N1 (24)

Vectors: 1. Normalized All Degree partition of N1 (24)

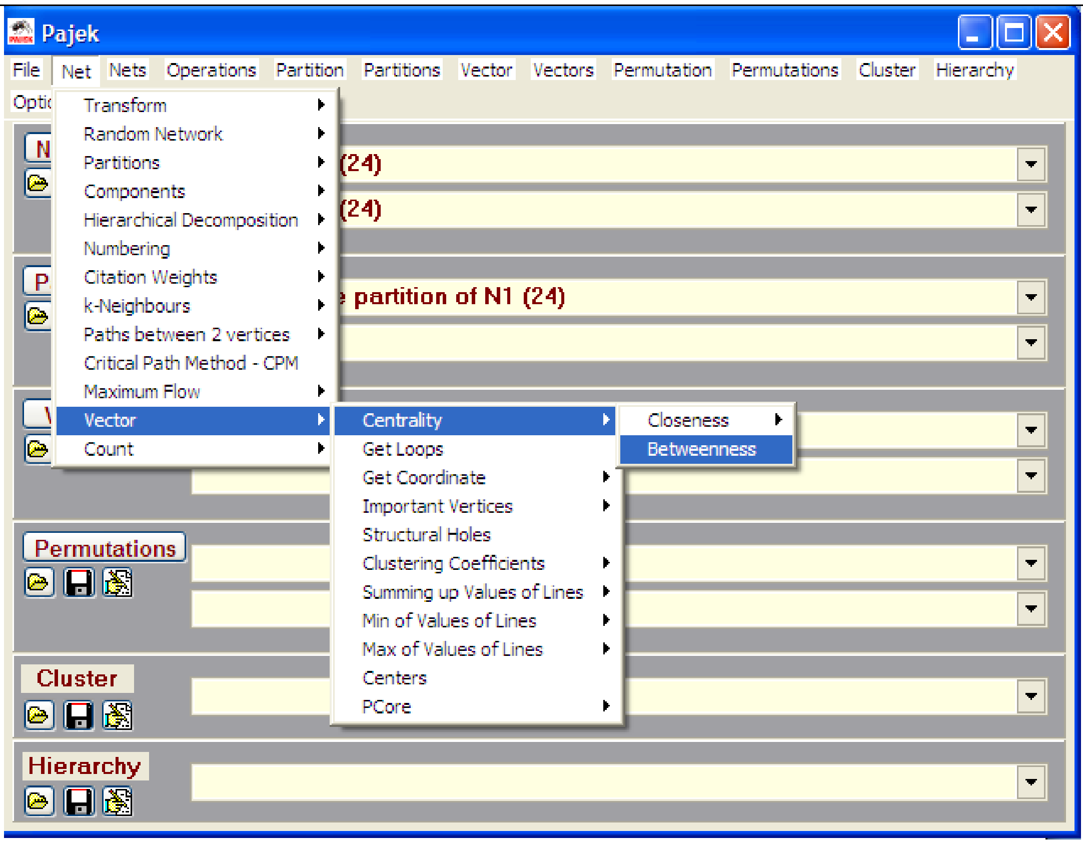
Also created is a report window. This explains what we have done, but also gives us the degree centralization of the network (i.e., the percentage of possible ties which were created). In this case it is 0.18, showing that



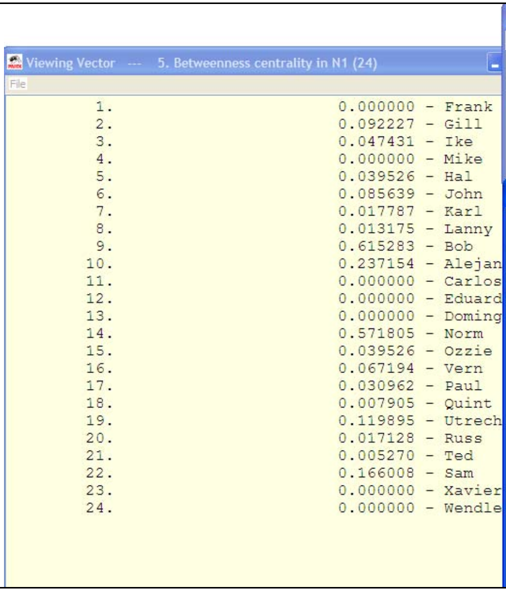
```
-----------------------------------------------------------------------------------
 Working...
----------------------------------------------------------
Network All Degree Centralization = 0.18181818
----------------------------------------------------------
 Time spent:  0:00:00
```

18% of all possible connections (i.e., where everyone speaks to everyone else) have been formed.

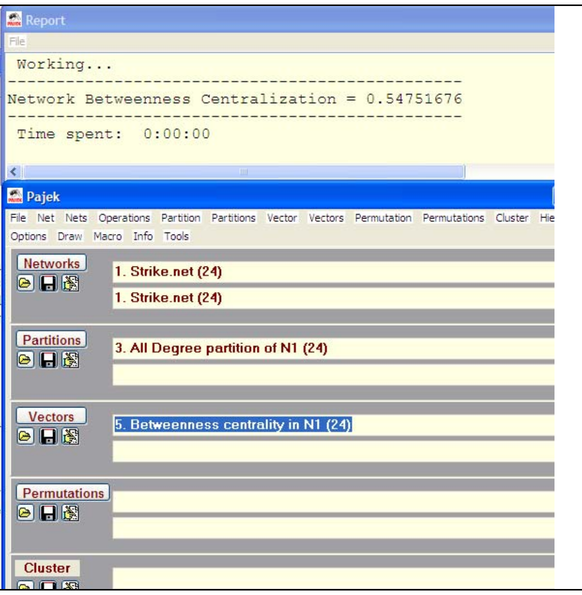
We might be interested in other forms of centrality. Betweenness centrality measures how often an actor is part of a path between two other actors (i.e., where A and B can only communicate though C, whether directly or through others). The higher the betweenness centrality, the increased opportunity for actors to influence and control information which flows around a network. This can be found from the 'net', 'centrality', 'betweenness' drop down menu.



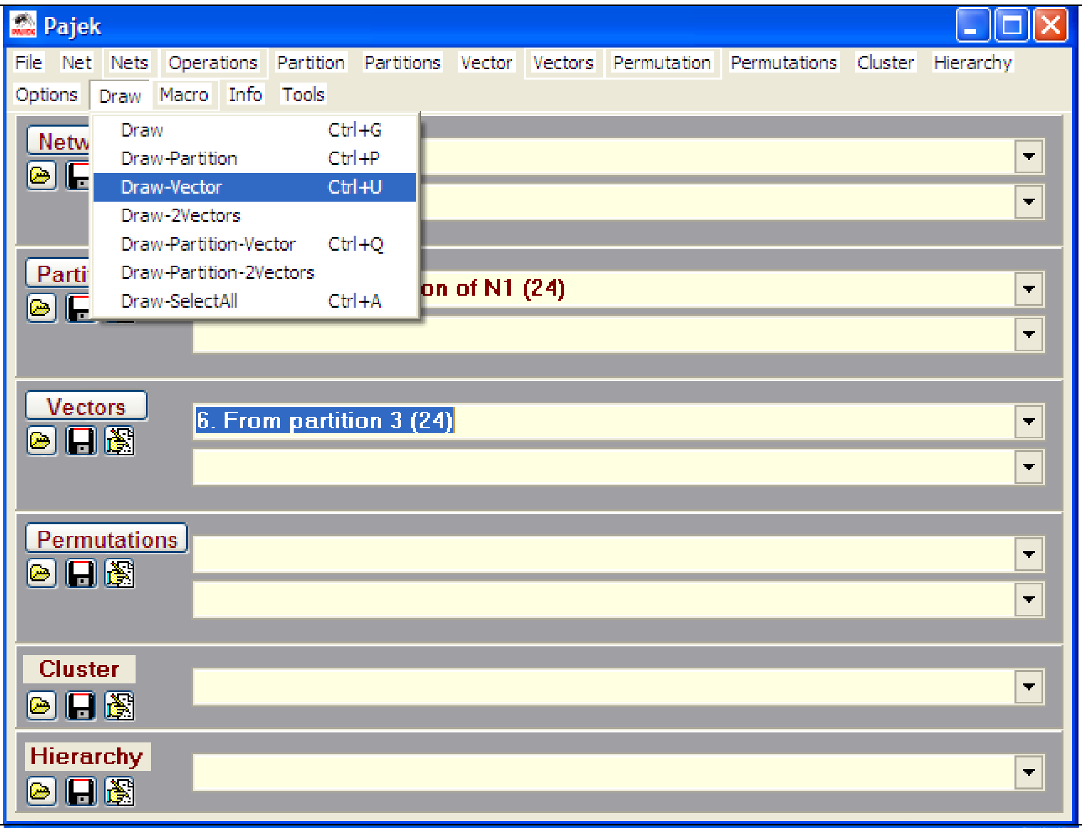
Again, the report window gives us a score for the overall network (.548) which can be compared to other similar networks. Clicking on the highlighted vector enables us to see the scores for each individual. Frank, in the first row, has a score of 0 as he never connects people. Bob (9) has the highest value of 0.61.



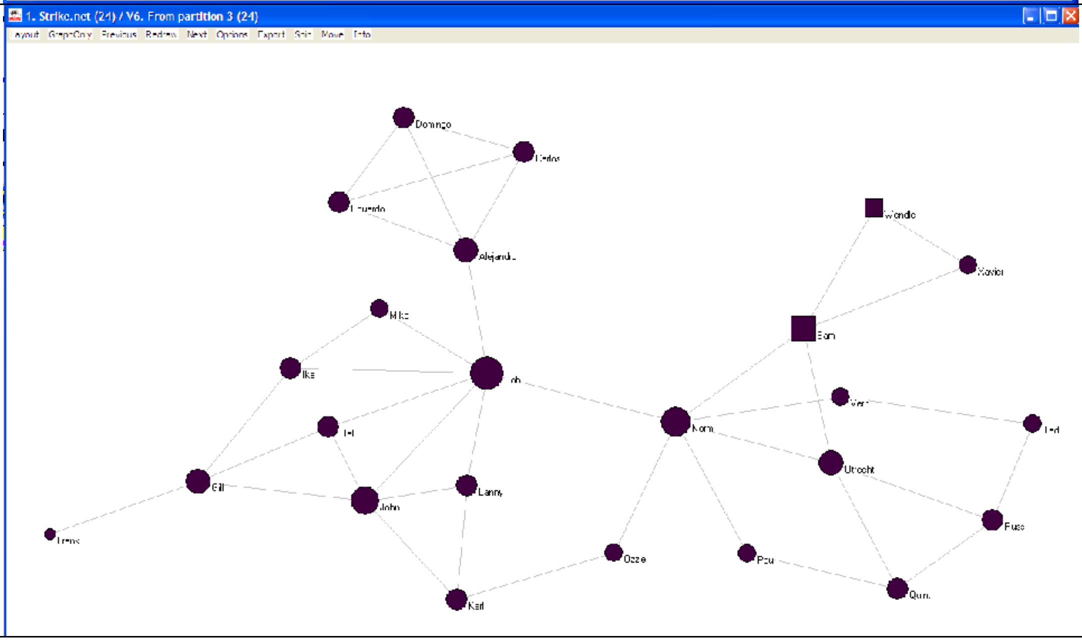
```
       1.                      0.000000 - Frank
       2.                      0.092227 - Gill
       3.                      0.047431 - Ike
       4.                      0.000000 - Mike
       5.                      0.039526 - Hal
       6.                      0.085639 - John
       7.                      0.017787 - Karl
       8.                      0.013175 - Lanny
       9.                      0.615283 - Bob
      10.                      0.237154 - Alejan
      11.                      0.000000 - Carlos
      12.                      0.000000 - Eduard
      13.                      0.000000 - Doming
      14.                      0.571805 - Norm
      15.                      0.039526 - Ozzie
      16.                      0.067194 - Vern
      17.                      0.030962 - Paul
      18.                      0.007905 - Quint
      19.                      0.119895 - Utrech
      20.                      0.017128 - Russ
      21.                      0.005270 - Ted
      22.                      0.166008 - Sam
      23.                      0.000000 - Xavier
      24.                      0.000000 - Wendle
```



```
 Working...
-------------------------------------------------------
Network Betweenness Centralization = 0.54751676
-------------------------------------------------------
 Time spent:  0:00:00
```

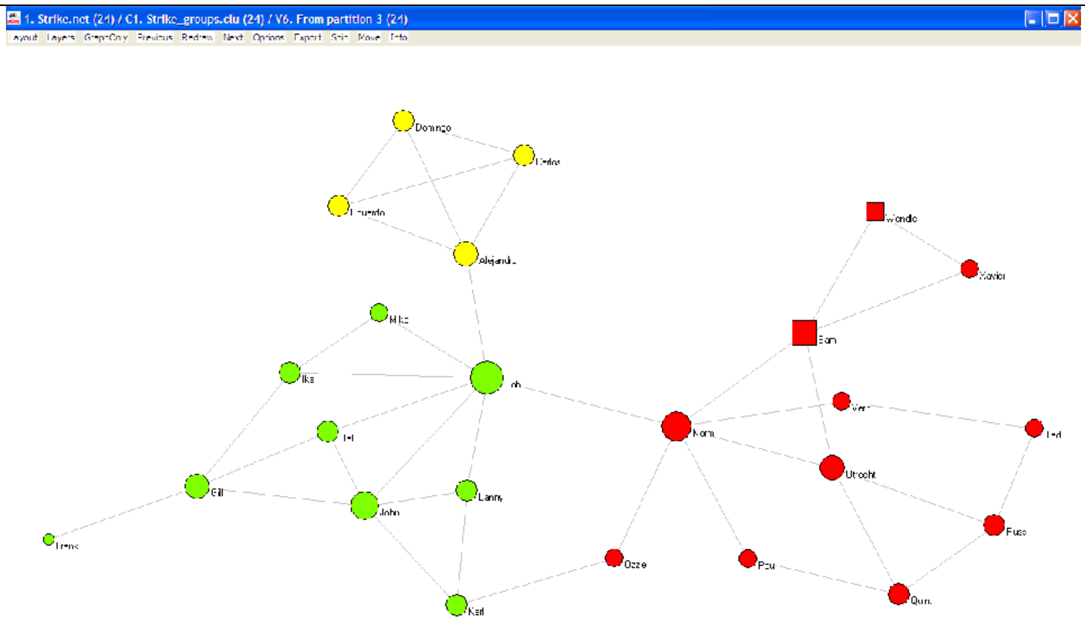
<table>
<tr>
<td>We might want to visualise this to understand the importance of betweenness in terms of the network. We can do this as we did with the partitions, but selecting ‘Draw-Vector’ (or CTRL+U).</td>
<td>


on of N1 (24)

6. From partition 3 (24)
</td>
</tr>
<tr>
<td>Draw-Vector sizes the node by the scale of the vector. This demonstrates that, for instance, Gill has a reasonable level of influence despite being on the left of the diagram away from the more central actors. This is because they alone can communicate to Frank and manipulate his opinions to the group.</td>
<td>

1. Strike.net (24) / V6. From partition 3 (24)

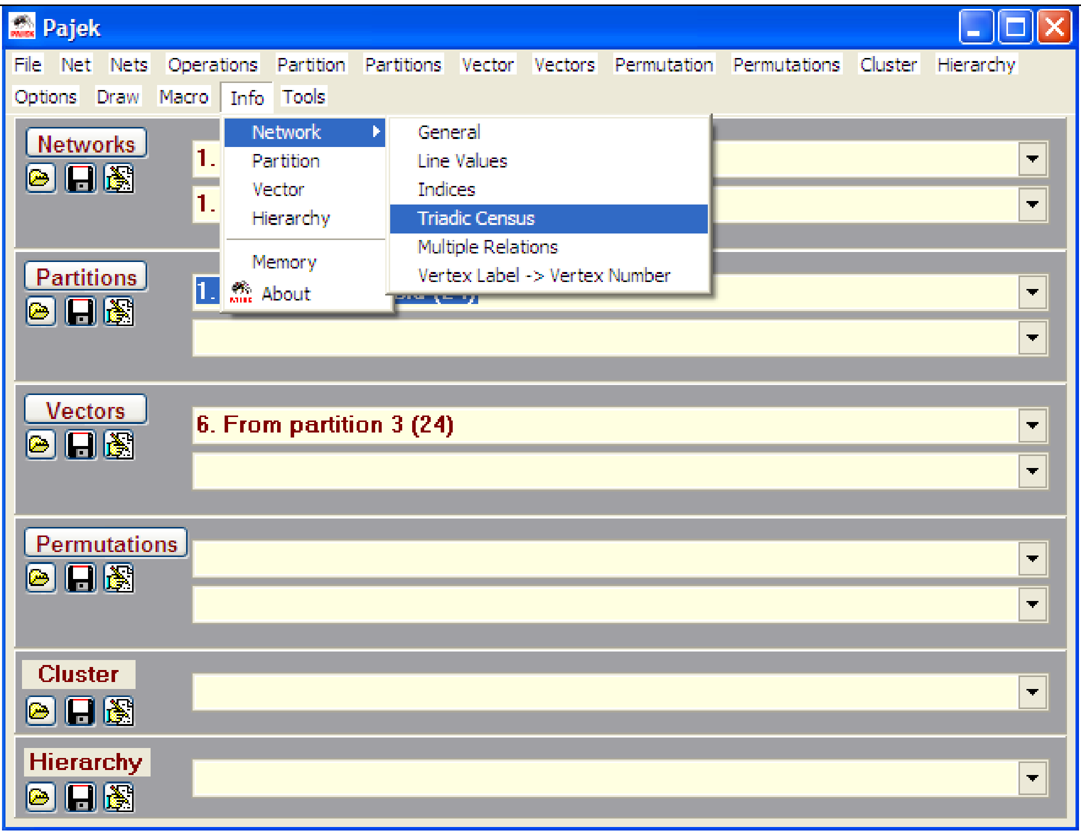
</td>
</tr>
</table>

We can mix together different ways of visualising the data. We could look at both partition (in this instance the three groups) and a vector (in this case betweenness) to see how the groups differ. Whilst the upper group of Spanish speakers appears marginalised, using betweenness values we can see they have as much potential influence as the other groups.

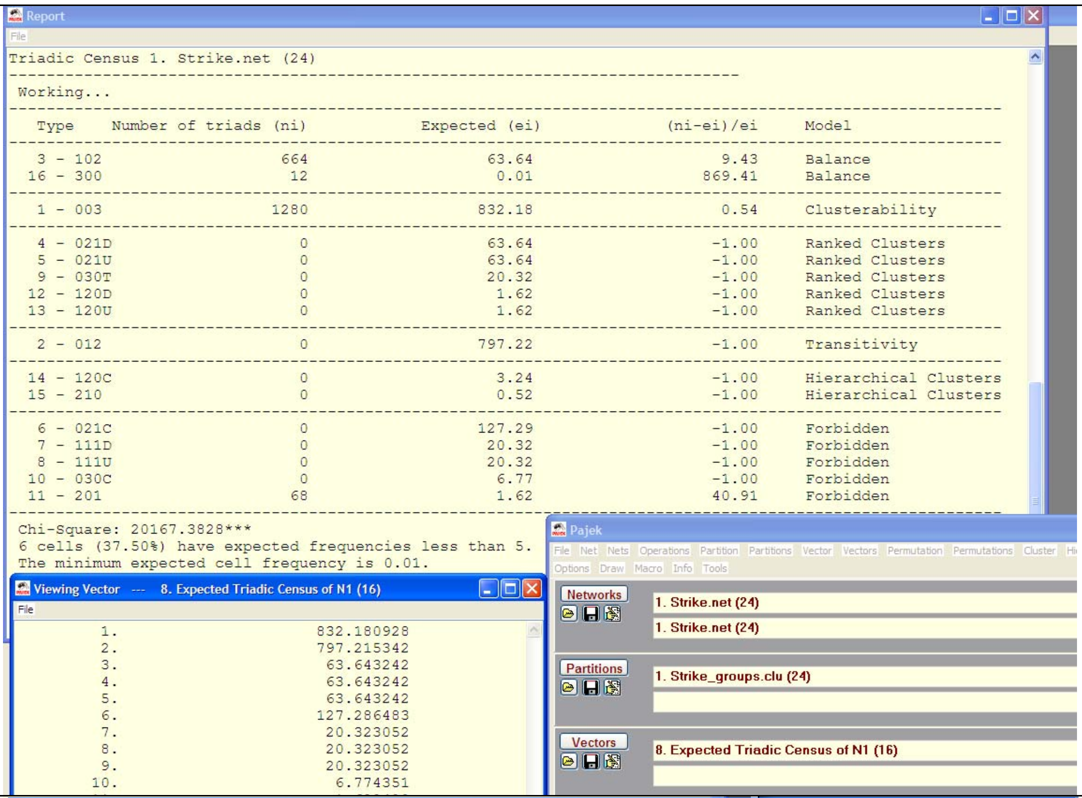

1. Strike.net (24) / C1. Strike_groups.clu (24) / V6. From partition 3 (24)
Layout Layers GraphOnly Previous Redraw Next Options Export Spin Move Info
Domingo
Carlos
Eduardo
Alejandro
Wendle
Xavier
Mike
Sam
Ike
Bob
Vern
Ted
Hal
Norm
Utrecht
Gill
Lanny
John
Frank
Russ
Ozzie
Paul
Quint
Karl

We can look at other elements of the network. We might be interested in a triadic census of the data. This can be performed using the info, network, triadic census drop down menu.

A triadic census takes every possible combination of three actors and looks at the structures between them.



This produces a report of the number of each triad we have observed and the expected number given the number of actors and links. The triads are also labelled according to what they show. A chi-square test is conducted, which in this case shows a difference from what would be expected. There is a large percentage of triads which all connect to each other, showing there is balance within the network (i.e., my friends' friend is my friend).



```
Triadic Census 1. Strike.net (24)
-----------------------------------------------------------------------------
 Working...
-------------------------------------------------------------------------------------------------------------
   Type    Number of triads (ni)              Expected (ei)              (ni-ei)/ei      Model
-------------------------------------------------------------------------------------------------------------
   3 - 102                  664                     63.64                      9.43      Balance
  16 - 300                   12                      0.01                    869.41      Balance
-------------------------------------------------------------------------------------------------------------
   1 - 003                 1280                    832.18                      0.54      Clusterability
-------------------------------------------------------------------------------------------------------------
   4 - 021D                   0                     63.64                     -1.00      Ranked Clusters
   5 - 021U                   0                     63.64                     -1.00      Ranked Clusters
   9 - 030T                   0                     20.32                     -1.00      Ranked Clusters
  12 - 120D                   0                      1.62                     -1.00      Ranked Clusters
  13 - 120U                   0                      1.62                     -1.00      Ranked Clusters
-------------------------------------------------------------------------------------------------------------
   2 - 012                    0                    797.22                     -1.00      Transitivity
-------------------------------------------------------------------------------------------------------------
  14 - 120C                   0                      3.24                     -1.00      Hierarchical Clusters
  15 - 210                    0                      0.52                     -1.00      Hierarchical Clusters
-------------------------------------------------------------------------------------------------------------
   6 - 021C                   0                    127.29                     -1.00      Forbidden
   7 - 111D                   0                     20.32                     -1.00      Forbidden
   8 - 111U                   0                     20.32                     -1.00      Forbidden
  10 - 030C                   0                      6.77                     -1.00      Forbidden
  11 - 201                   68                      1.62                     40.91      Forbidden
-------------------------------------------------------------------------------------------------------------
 Chi-Square: 20167.3828***
 6 cells (37.50%) have expected frequencies less than 5.
 The minimum expected cell frequency is 0.01.
```

Viewing Vector --- 8. Expected Triadic Census of N1 (16)

```
        1.                  832.180928
        2.                  797.215342
        3.                   63.643242
        4.                   63.643242
        5.                   63.643242
        6.                  127.286483
        7.                   20.323052
        8.                   20.323052
        9.                   20.323052
       10.                    6.774351
```

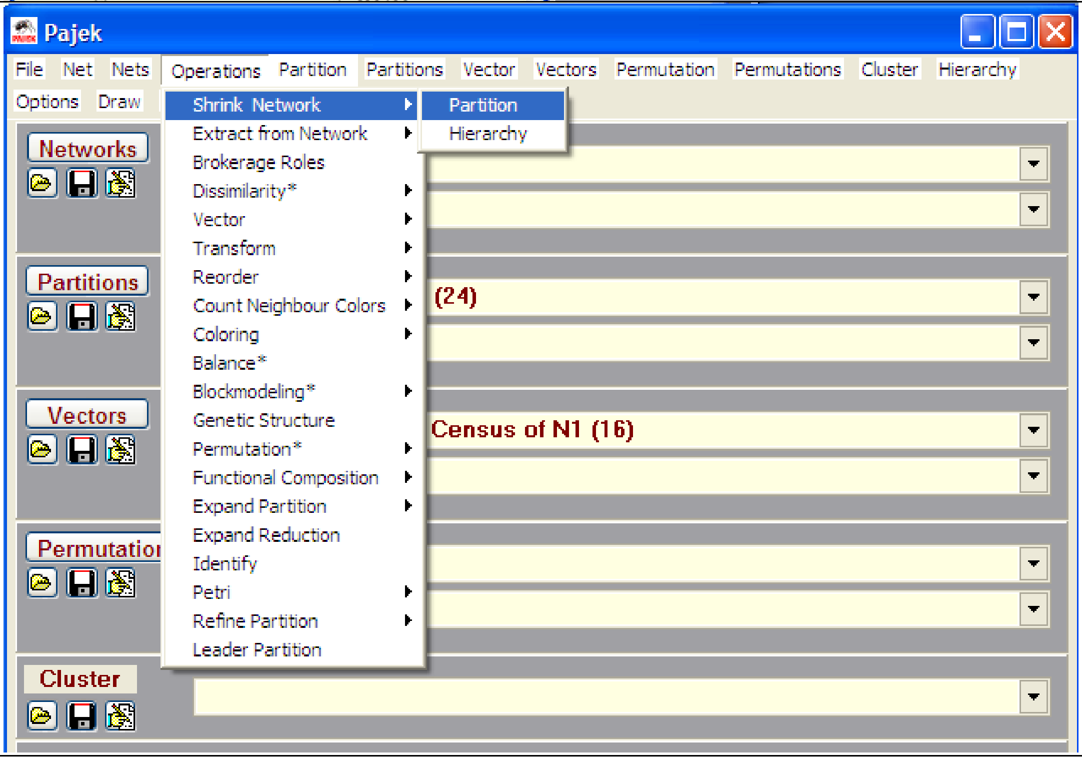
Networks: 1. Strike.net (24)
1. Strike.net (24)

Partitions: 1. Strike_groups.clu (24)

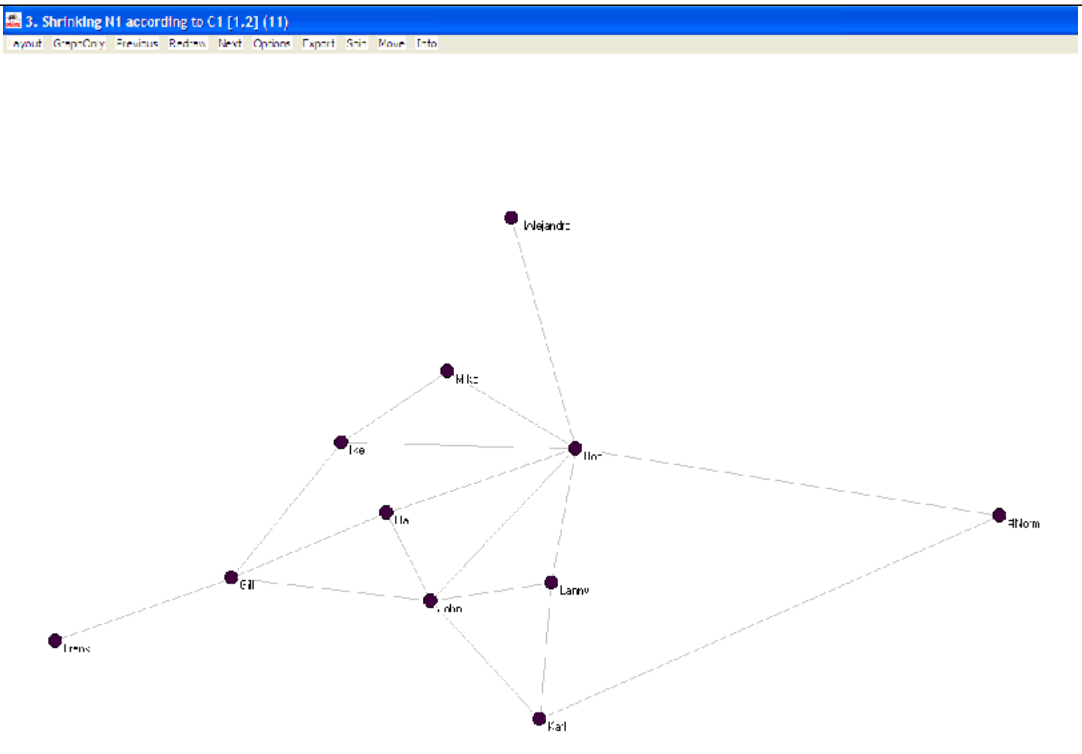
Vectors: 8. Expected Triadic Census of N1 (16)

Report
File
Pajek
File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hi
Options Draw Macro Info Tools

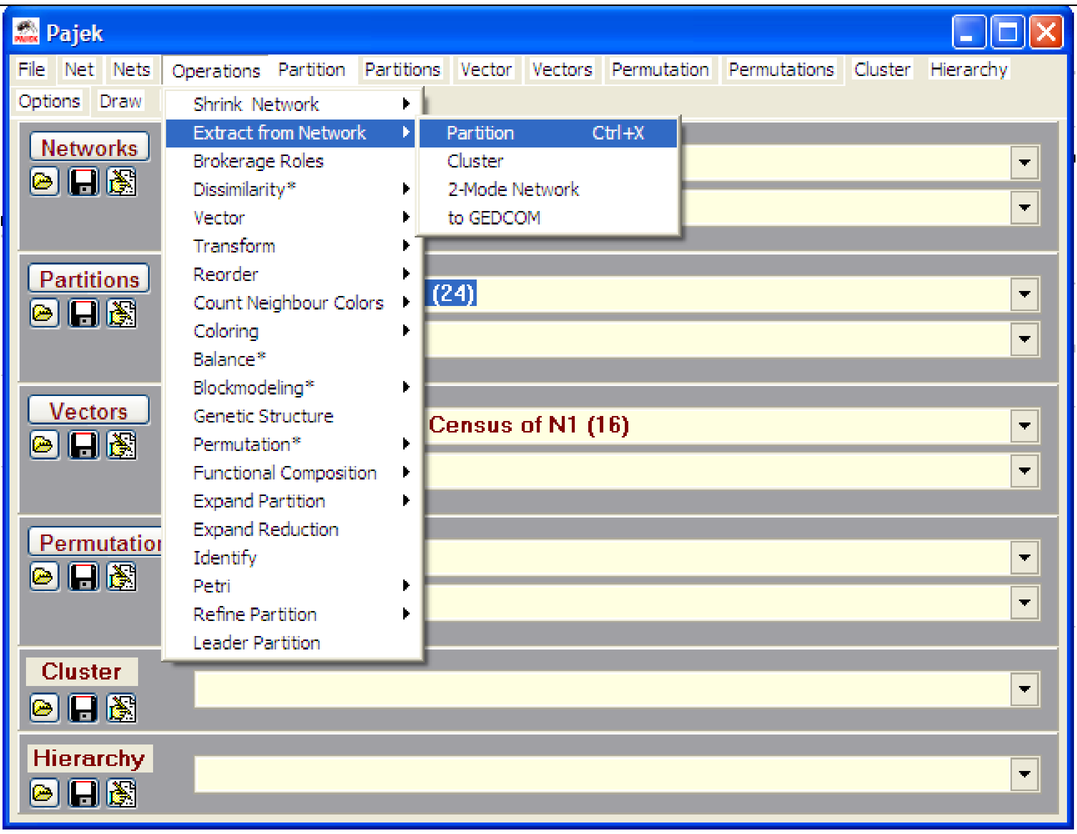
We can also perform analysis grouping together particular clusters. For instance, we can shrink the network into partition. This produces a dialogue box asking for the minimum number of ties to connect to another cluster (default of 1), before asking which partition(s) should not be shrunk. Once selected, it reduces all other clusters to a single actor (assuming the analysed cluster hold sufficient number of links to its members).



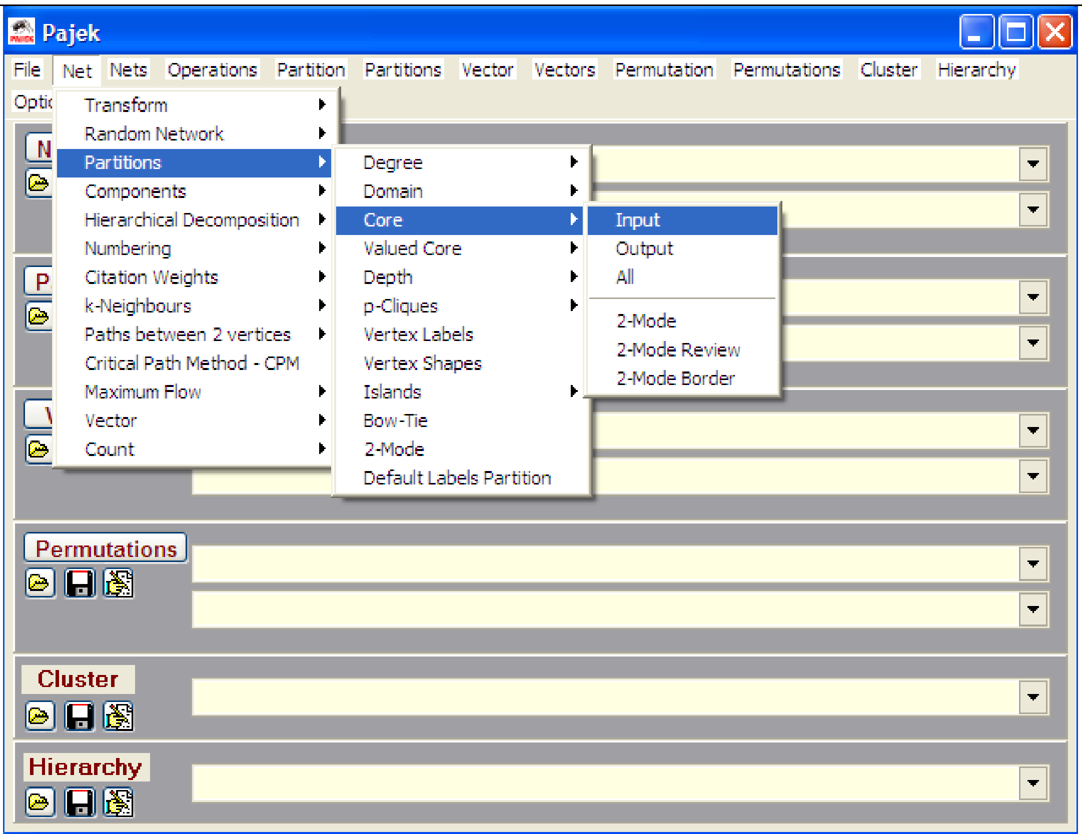
Here the network is shown, but only for the younger English cohort. The older English cohort have been shrunk into one group, and the Spanish cohort into another (labelled by #). We can now rerun any of our analysis (such as centrality methods) viewing the other clusters as sub-groups rather than their individual actors.



Alternatively, we can just decide to extract a particular partition from the network (by drop down menu or CTRL+X) which enables us to perform analysis irrespective of the wider network the actors are part of. (Note: this is particular useful if you have, for instance, international trade networks and you wish to look at Europe in isolation, or European countries by the other European countries and continents they trade with).

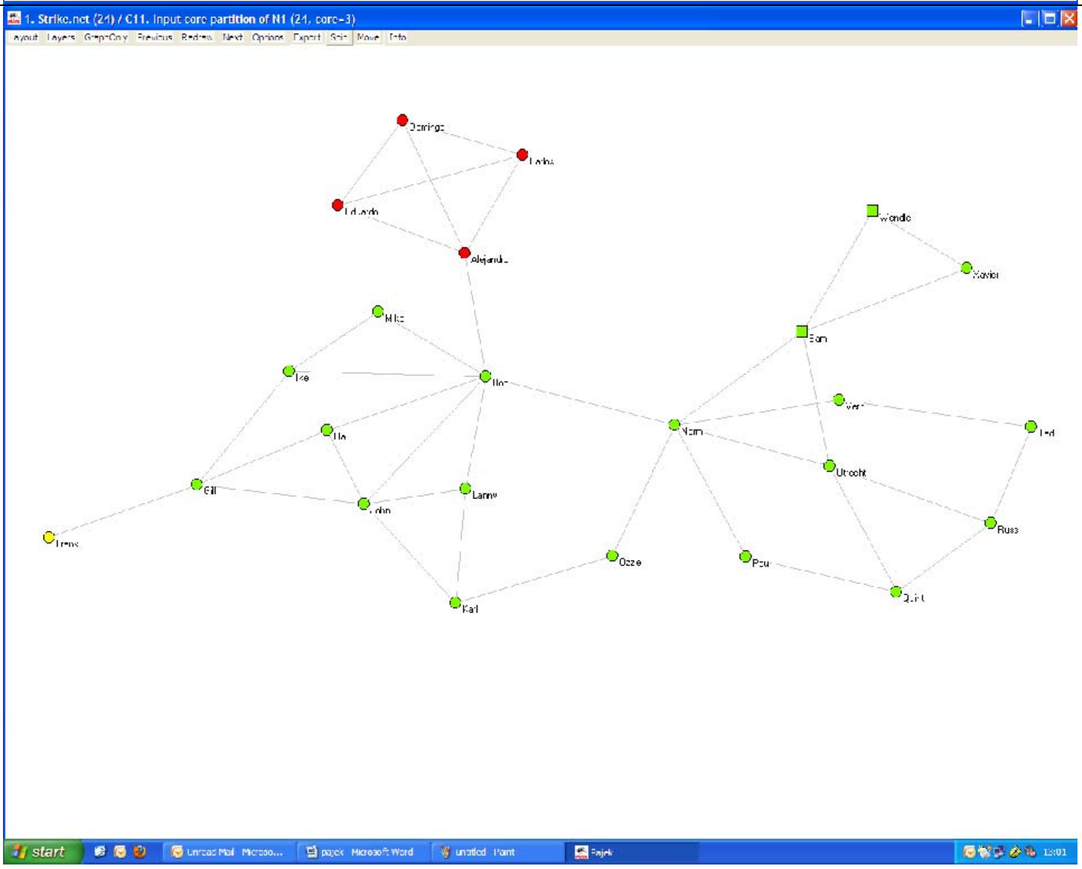


We could also look at the whole network and decide to look at the cores. This produces sub-graphs whereby each actor is connected to X actors who all have at least Y links. The core for each actor is the largest possible Y that they can be part of a core of.



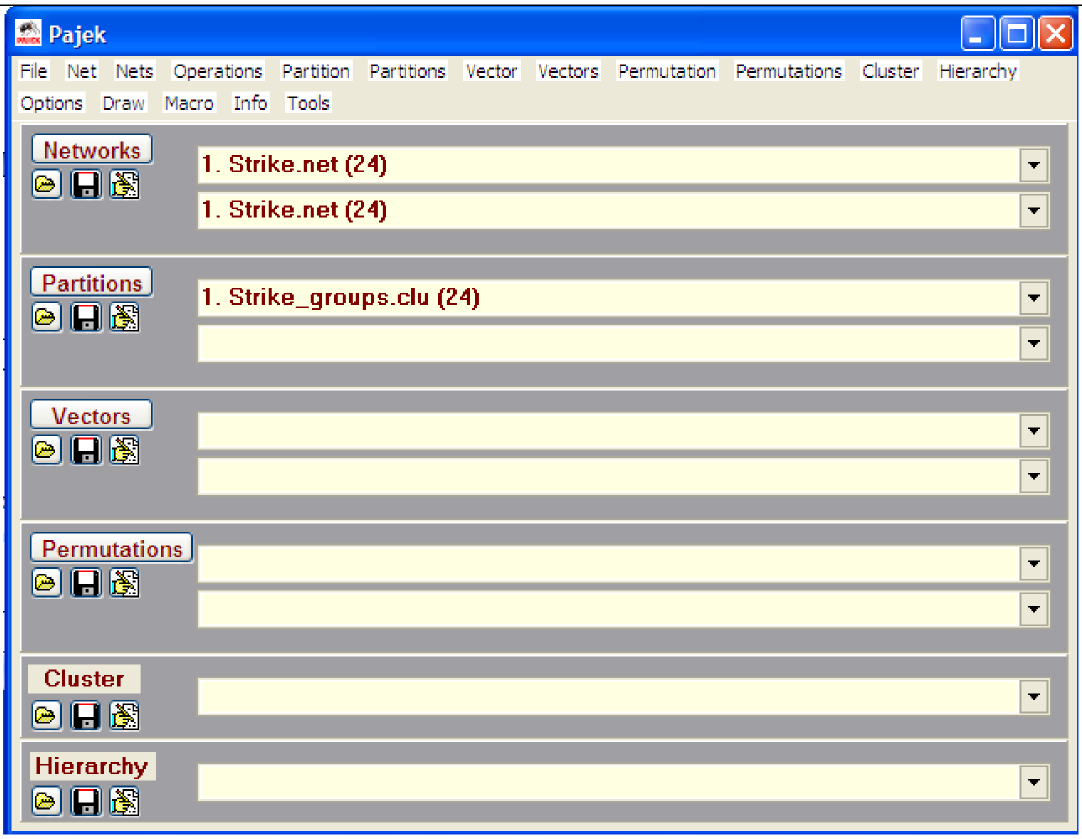
Therefore, we see that Frank can only be incorporated in a core measured by one link. The English speakers, Frank aside, all form part of the two core. The three Spanish speakers form a three-core.

Whilst Karl has three connections, Ozzy does not. Therefore, removing Ozzy from the 3-core removes Karl, which in turn removes Lanny. Losing Mike loses Ike, which removes Gill and quickly the English structure falls apart. Therefore, this is ideal for spotting well-constructed elements of networks and identify cores (even if of non-central actors).



Saving data in Pajek involves using the file buttons under each tab (i.e., networks, partitions etc.), or the ‘file’ drop down menu. Each file you’ve used needs to be saved separately. You can move between them using the right-hand arrows. Alternatively, use ‘file, Pajek project file, save’ to save everything in one large file.

Note: Pajek keeps every file you’ve opened in its list which can become large if working on multiple tasks. Therefore, it is often beneficial to close and reopen the window to avoid confusion.

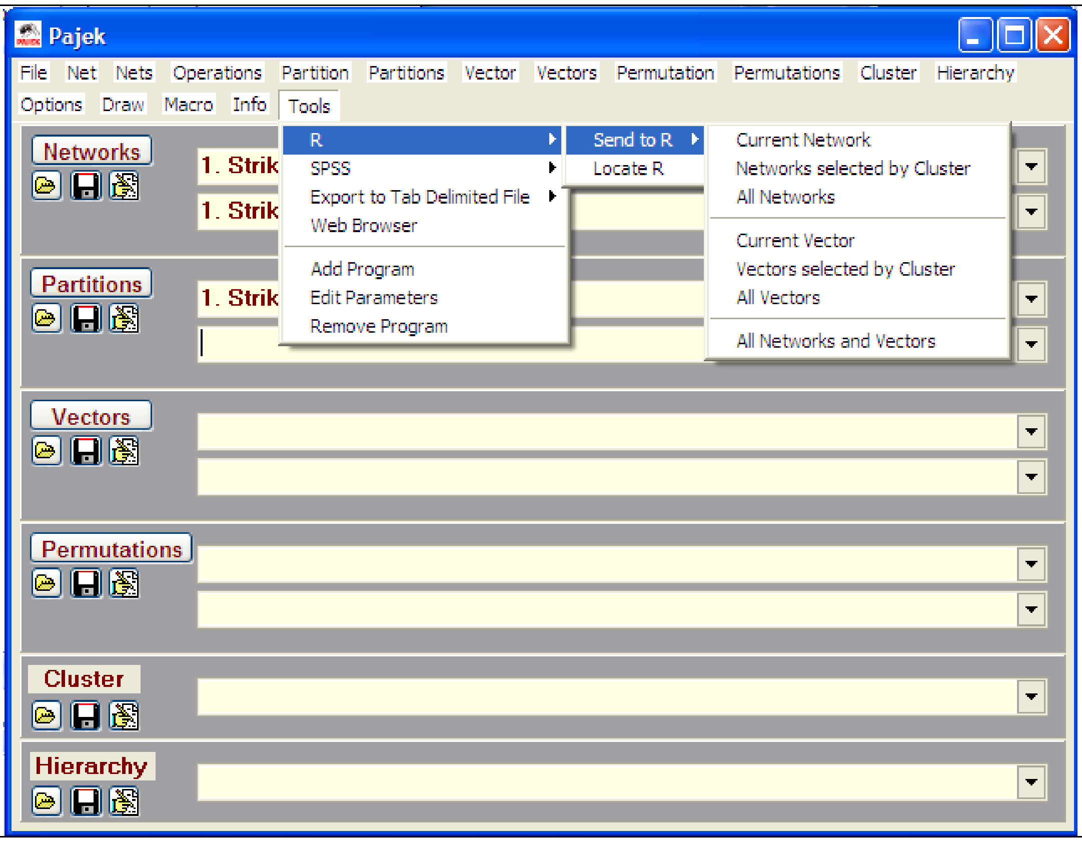

Pajek
File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy
Options Draw Macro Info Tools
Networks
Partitions
Vectors
Permutations
Cluster
Hierarchy

1. Strike.net (24)

1. Strike.net (24)

1. Strike_groups.clu (24)

Pajek also has the options to export data directly to R and SPSS.


Pajek
File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy
Options Draw Macro Info Tools
R
SPSS
Export to Tab Delimited File
Web Browser
Add Program
Edit Parameters
Remove Program
Send to R
Locate R
Current Network
Networks selected by Cluster
All Networks
Current Vector
Vectors selected by Cluster
All Vectors
All Networks and Vectors
Networks
Partitions
Vectors
Permutations
Cluster
Hierarchy

1. Strik

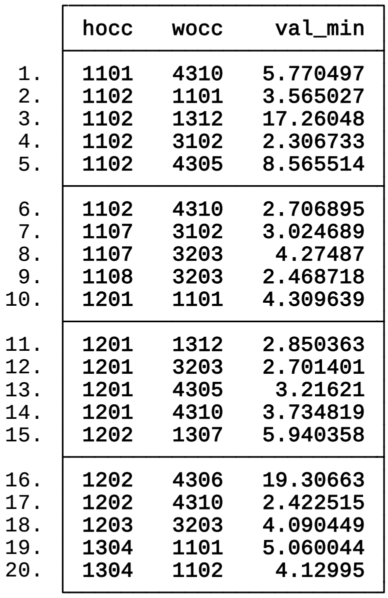
1. Strik

1. Strik

Preparing data for Pajek

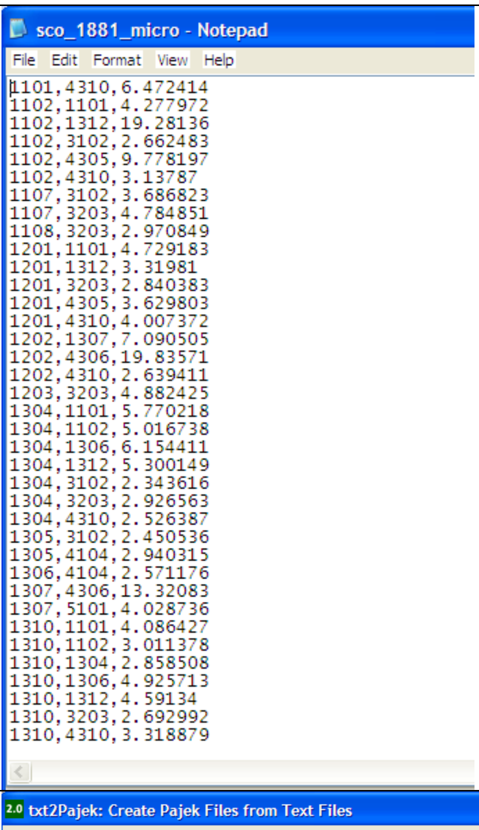
The Pajek website (http://pajek.imfm.si) offers a range of useful datasets for exploring network theory. It also offers the Excel2Pajek and Text2Pajek tools for formatting data. Converting a dataset into Pajek format is simple.

Firstly, produce a dataset which has two columns of nodes, for instance, male and female occupations within married couples. A value for strength of line can also be saved. This can be produced in any format.



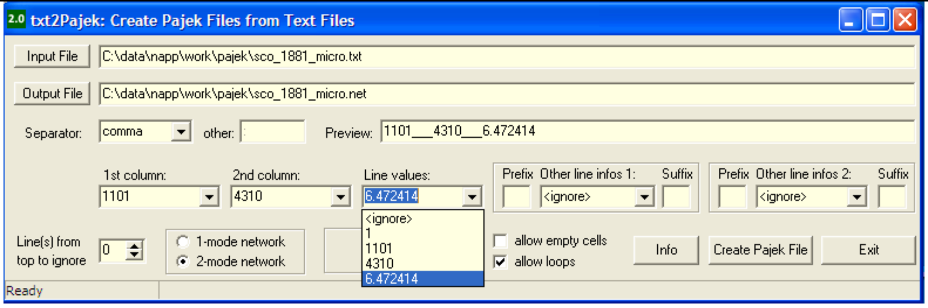
```
       hocc   wocc    val_min

  1.   1101   4310   5.770497
  2.   1102   1101   3.565027
  3.   1102   1312   17.26048
  4.   1102   3102   2.306733
  5.   1102   4305   8.565514

  6.   1102   4310   2.706895
  7.   1107   3102   3.024689
  8.   1107   3203    4.27487
  9.   1108   3203   2.468718
 10.   1201   1101   4.309639

 11.   1201   1312   2.850363
 12.   1201   3203   2.701401
 13.   1201   4305    3.21621
 14.   1201   4310   3.734819
 15.   1202   1307   5.940358

 16.   1202   4306   19.30663
 17.   1202   4310   2.422515
 18.   1203   3203   4.090449
 19.   1304   1101   5.060044
 20.   1304   1102    4.12995
```

To use Text2Pajek, save the list as a text file. In this example it's a comma-separated version. Just the rows of data are required, with no additional information.



```
1101,4310,6.472414
1102,1101,4.277972
1102,1312,19.28136
1102,3102,2.662483
1102,4305,9.778197
1102,4310,3.13787
1107,3102,3.686823
1107,3203,4.784851
1108,3203,2.970849
1201,1101,4.729183
1201,1312,3.31981
1201,3203,2.840383
1201,4305,3.629803
1201,4310,4.007372
1202,1307,7.090505
1202,4306,19.83571
1202,4310,2.639411
1203,3203,4.882425
1304,1101,5.770218
1304,1102,5.016738
1304,1306,6.154411
1304,1312,5.300149
1304,3102,2.343616
1304,3203,2.926563
1304,4310,2.526387
1305,3102,2.450536
1305,4104,2.940315
1306,4104,2.571176
1307,4306,13.32083
1307,5101,4.028736
1310,1101,4.086427
1310,1102,3.011378
1310,1304,2.858508
1310,1306,4.925713
1310,1312,4.59134
1310,3203,2.692992
1310,4310,3.318879
```

We can then open the txt2Pajek software. Firstly, we use the 'input file' button to select the file. It then defaults to save the file in the same folder with the same name but



Input File: C:\data\napp\work\pajek\sco_1881_micro.txt

Output File: C:\data\napp\work\pajek\sco_1881_micro.net

Separator: comma   other: Preview: 1101___4310___6.472414

1st column: 1101   2nd column: 4310   Line values: 6.472414

<ignore>
1
1101
4310
6.472414

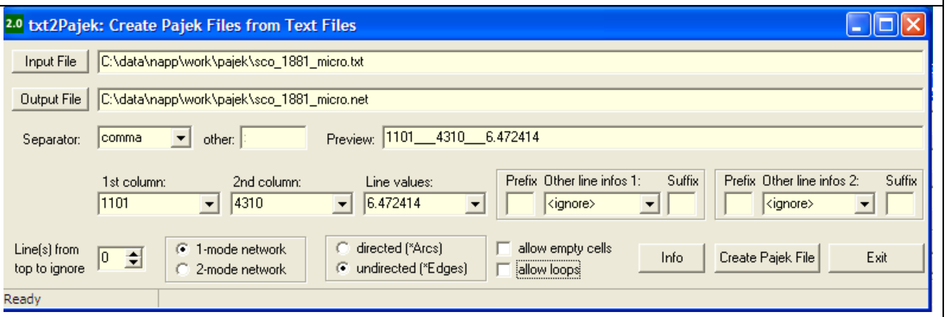
Prefix   Other line infos 1: <ignore>   Suffix

Prefix   Other line infos 2: <ignore>   Suffix

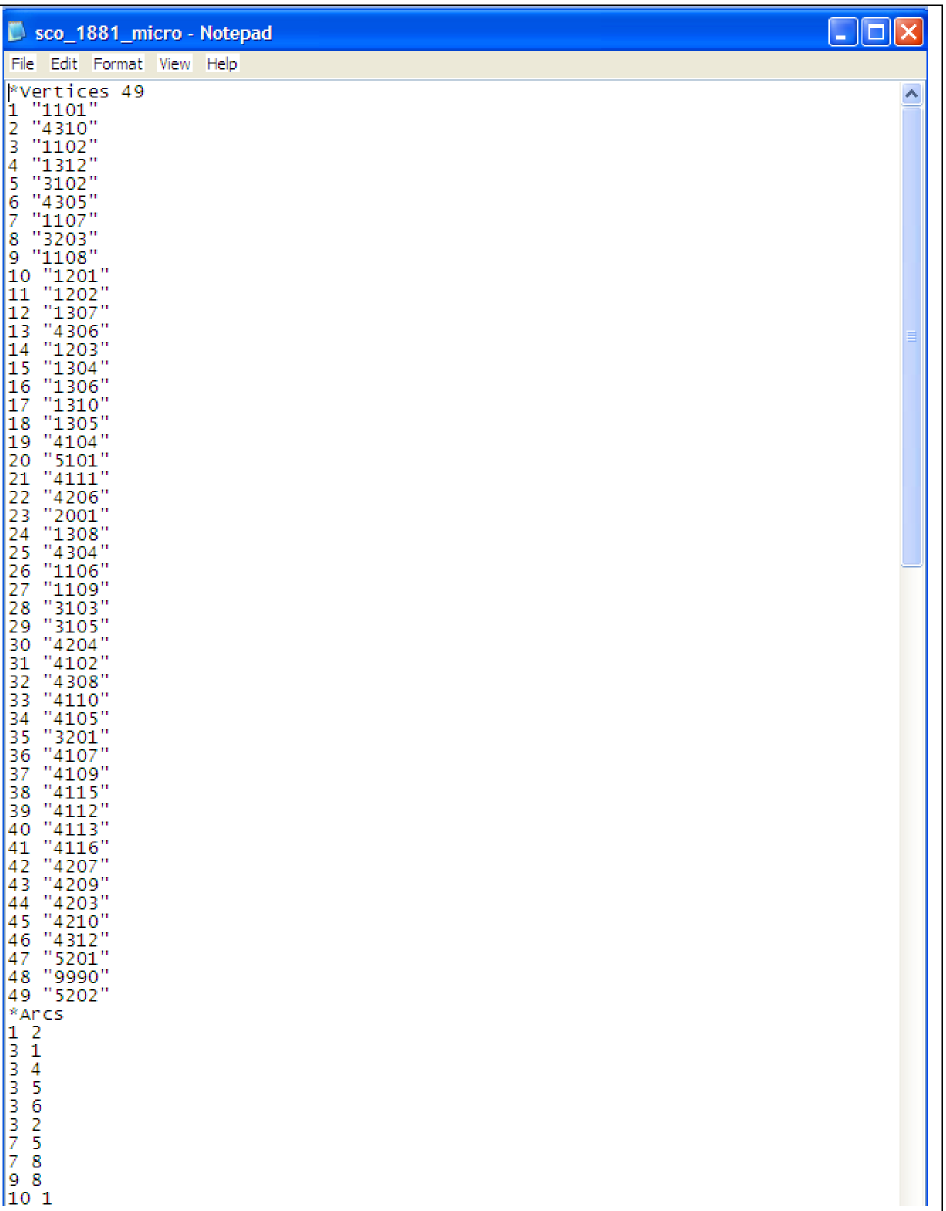
Line(s) from top to ignore: 0

○ 1-mode network
● 2-mode network

☐ allow empty cells
☑ allow loops

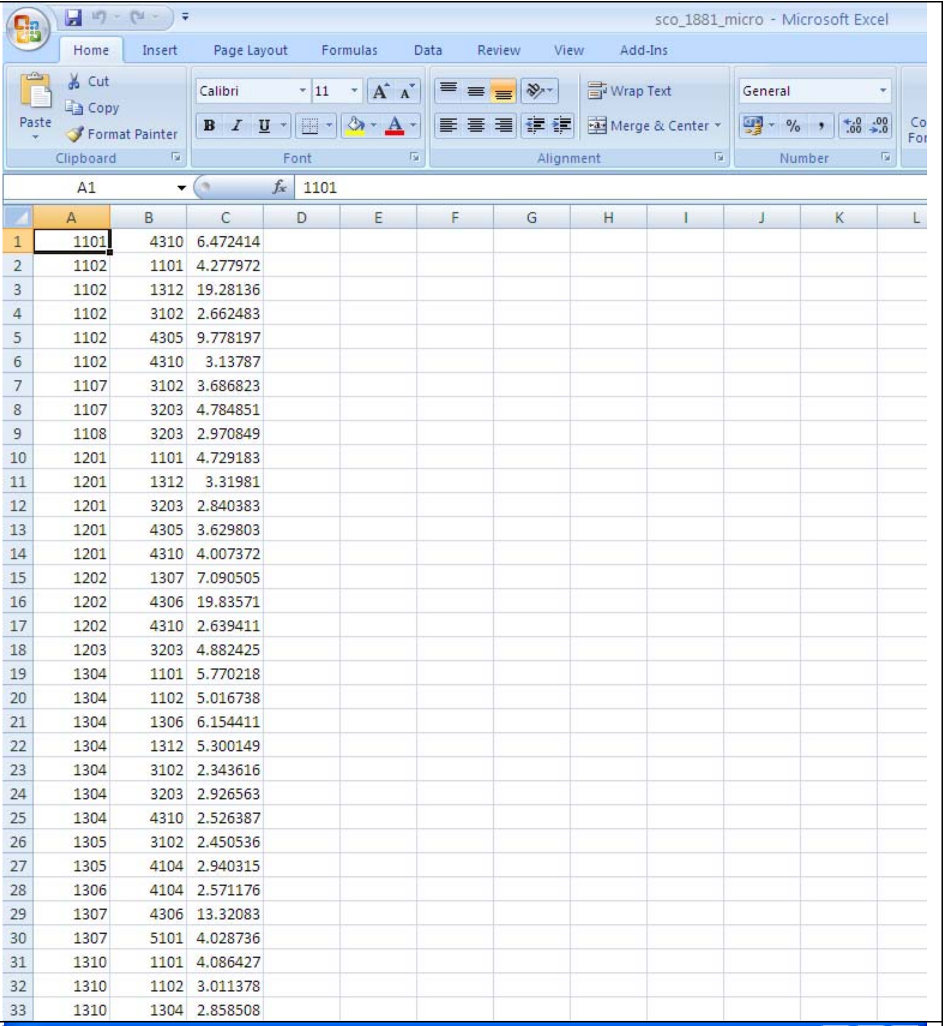
different suffix (clicking on the output name allows you to change it). We can specify the separator (comma) and select the 1st and 2nd columns from the available columns. We can include or ignore line values.	
We can specify if it is a one-mode (the same actors in both columns, such as jobs) or two-mode (different actors in each column, such as employees-employers) and whether the network is directed (i.e., if ties are assumed to always be replicated). We then click on 'Create Pajek File'.	 2.0 txt2Pajek: Create Pajek Files from Text Files Info Create Pajek File Exit Input File: C:\data\napp\work\pajek\sco_1881_micro.txt Output File: C:\data\napp\work\pajek\sco_1881_micro.net Separator: comma other: Preview: 1101__4310__6.472414 1st column: 1101 2nd column: 4310 Line values: 6.472414 Prefix Other line infos 1: <ignore> Suffix Prefix Other line infos 2: <ignore> Suffix Line(s) from top to ignore: 0 ● 1-mode network ○ 2-mode network ○ directed (*Arcs) ● undirected (*Edges) □ allow empty cells □ allow loops Ready

This produces a .net file which we need for Pajek. It starts off by specifying how many vertices exist, then giving a number to each label (as we saw above with the matrix). It also produces a list of the arcs/edges (links, whether directed or undirected), just showing the two numbers which are connected. In this example the labels are the occupational numbers. The data could be exported with the labels instead, which might in some cases be beneficial if they are to be shown in the graphs (obviously, with limitations on the size of what will be readable).



```
*vertices 49
1 "1101"
2 "4310"
3 "1102"
4 "1312"
5 "3102"
6 "4305"
7 "1107"
8 "3203"
9 "1108"
10 "1201"
11 "1202"
12 "1307"
13 "4306"
14 "1203"
15 "1304"
16 "1306"
17 "1310"
18 "1305"
19 "4104"
20 "5101"
21 "4111"
22 "4206"
23 "2001"
24 "1308"
25 "4304"
26 "1106"
27 "1109"
28 "3103"
29 "3105"
30 "4204"
31 "4102"
32 "4308"
33 "4110"
34 "4105"
35 "3201"
36 "4107"
37 "4109"
38 "4115"
39 "4112"
40 "4113"
41 "4116"
42 "4207"
43 "4209"
44 "4203"
45 "4210"
46 "4312"
47 "5201"
48 "9990"
49 "5202"
*Arcs
1 2
3 1
3 4
3 5
3 6
3 2
7 5
7 8
9 8
10 1
```

Using the Excel2Pajek tool is just as simple. Start off with the data saved as an Excel file. Again, how two columns showing the linkages which form part of the network. You have the option to have a third variable for the strength of the line. It does not matter if there is additional information stored in the Excel file, as you will select the columns you wish. Therefore, it would be possible to have both the occupational codes and labels in the same file, which you could export as two different networks (if required).



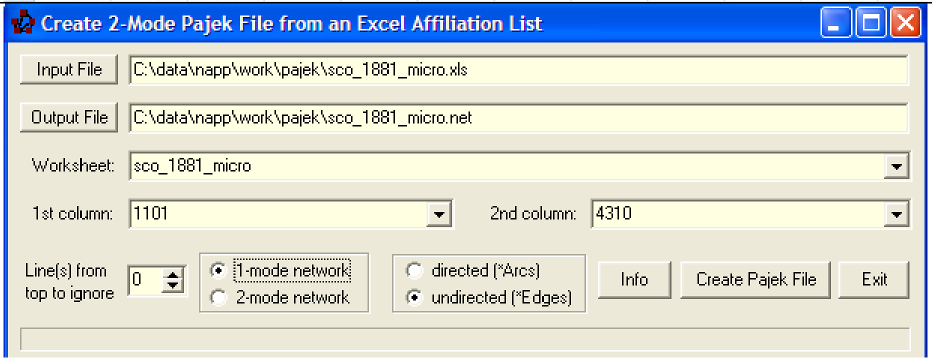
sco_1881_micro - Microsoft Excel

Calibri 11 General

A1 fx 1101

	A	B	C
1	1101	4310	6.472414
2	1102	1101	4.277972
3	1102	1312	19.28136
4	1102	3102	2.662483
5	1102	4305	9.778197
6	1102	4310	3.13787
7	1107	3102	3.686823
8	1107	3203	4.784851
9	1108	3203	2.970849
10	1201	1101	4.729183
11	1201	1312	3.31981
12	1201	3203	2.840383
13	1201	4305	3.629803
14	1201	4310	4.007372
15	1202	1307	7.090505
16	1202	4306	19.83571
17	1202	4310	2.639411
18	1203	3203	4.882425
19	1304	1101	5.770218
20	1304	1102	5.016738
21	1304	1306	6.154411
22	1304	1312	5.300149
23	1304	3102	2.343616
24	1304	3203	2.926563
25	1304	4310	2.526387
26	1305	3102	2.450536
27	1305	4104	2.940315
28	1306	4104	2.571176
29	1307	4306	13.32083
30	1307	5101	4.028736
31	1310	1101	4.086427
32	1310	1102	3.011378
33	1310	1304	2.858508

Again, we can select the input file and it defaults in the same way. We can decide which worksheet we want to use, and which columns are important. We can save as a 1- or 2-mode network, and decide to ignore the top line (if it is merely column labels). Click on 'Create Pajek file' and the file will be ready to be opened in Pajek.



Create 2-Mode Pajek File from an Excel Affiliation List

Input File: C:\data\napp\work\pajek\sco_1881_micro.xls

Output File: C:\data\napp\work\pajek\sco_1881_micro.net

Worksheet: sco_1881_micro

1st column: 1101 2nd column: 4310

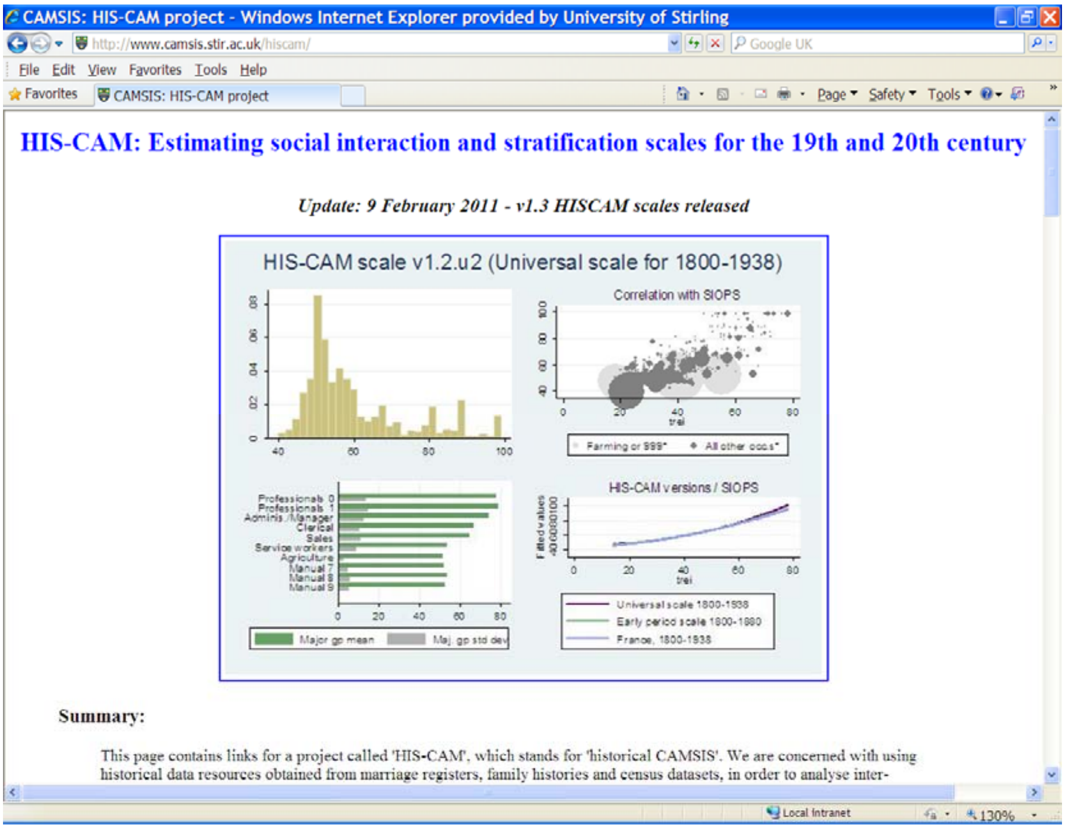
Line(s) from top to ignore: 0

● 1-mode network
○ 2-mode network

○ directed (*Arcs)
● undirected (*Edges)

Lab 2: Historical data on occupations and its analysis using SNA and SID approachs

This lab features some examples focussed upon using large scale occupational data on social connections. Examples include operatinalising key measures using occupational records, and worked examples of SID and SNA techniques. The Stata and R exercises can be found in the accompanying command files, whilst the Pajek instructions are given below (the Pajek exercise begins with a brief use of Stata, though that can be skipped if necessary).



HIS-CAM: Estimating social interaction and stratification scales for the 19th and 20th century

Update: 9 February 2011 - v1.3 HISCAM scales released

Summary:

This page contains links for a project called 'HIS-CAM', which stands for 'historical CAMSIS'. We are concerned with using historical data resources obtained from marriage registers, family histories and census datasets, in order to analyse inter-

Screenshot of the HIS-CAM website, a useful resource for data on occupations

Lab 2: Pajek exercises

Using SNA to analyse occupational stratification in the past

In this lab we explore ways we can use social network analysis to understand more about occupational stratification. Research in this area always benefits by consideration of the occupational structure and the national context of vocations. Research using social connections, similarly, can benefit from understanding underlying patterns of social interactions.

This analysis involves processes from both Stata and Pajek. Stata is utilised to generate the data which can be utilised by network analysis, which is processed within Pajek.

Our first step is to find a dataset consisting of pairs of occupations. For this lab, a dataset has been created consisting of within household combinations of microclasses. Only male ties with an age-gap of 16 years, aged between 16 and 75, are included. Combinations within the same microclass have been excluded.

```
. use $path9\scot81.dta, clear

. codebook, compact

Variable      Obs Unique      Mean   Min     Max  Label
--------------------------------------------------------------------------
serial     258740 136853  403449.1     2  784262  Household index number
pernum1    258740     67  3.802242     1      87  Person index within ...
age1       258740     48  52.48311    28      75  Age
sex1       258740      1         1     1       1  Sex
occgb1     258740    384   210.265     1     413  Occupation, Britain ...
hocc       258740     64  4164.652  1101    9990
pernum     258740     76  6.289681     1     110  Person index within ...
age        258740     48  21.85543    12      59  Age
sex        258740      1         1     1       1  Sex
occgb      258740    382  199.2149     1     413  Occupation, Britain ...
wocc       258740     64  4061.158  1101    9990
pershh     258740     47  7.602937     2      49
pairshh    258740     47  62.56172     1    1176
aged       258740     48  30.62768    16      63
--------------------------------------------------------------------------
```

Next we run a syntax file which automatically generates some information on pairs of occupations.

```
do http://www.camsis.stir.ac.uk/sonocs/do/pajek.do
```

This provides 16 variables. hocc is the older cohort occupation and wocc the younger cohort. freq is the number of connections for each combination. ewocc is the expected number of ties if the data was random. val_min is the value of over-representation (taking standard errors into account). These are the most important variables.

```
Variable      Obs Unique      Mean        Min        Max  Label
------------------------------------------------------------------------------
hocc         3131     64  3458.181       1101       9990
wocc         3131     64  3401.974       1101       9990
freq         3131    395  82.63813          1      25594  (count) freq
tot          3131      1    258740     258740     258740  total number in ...
nhocc        3131     64  4849.383          4      39522  total number of ...
nwocc        3131     64  4828.742          5      33491  total number of ...
phocc        3131     64  .0187423   .0000155   .1527479  percentage of me...
pwocc        3131     64  .0186625   .0000193   .1294388  percentage of wo...
ewocc        3131   3129  79.17888   .0091598   5115.681  expected number ...
prop         3131    395  .0003194   3.86e-06   .0989178
staner       3131    395  .0000229   3.86e-06   .0005869  Standard error f...
pro_obs      3131    395  .0003194   3.86e-06   .0989178  Observed proport...
pro_exp      3131   3129   .000306   3.54e-08   .0197715  Expected proport...
pro_min      3131    395  .0002965   7.28e-12   .0983309  Lower confidence...
pro_max      3131    395  .0003423   7.73e-06   .0995048  Higher confidenc...
value        3131   3116  1.592005   .0257491    109.173  Observed value o...
val_min      3131   3131  .9458267   4.85e-08   24.38851  Value of represe...
val_max      3131   3131  2.238184   .0503272   218.3458  Value of represe...
------------------------------------------------------------------------------
```

Next we need some criteria for which ties are needed. We are interested in cases occurring at least once in every 10,000 cases (therefore, frequency of at least 28) and which occur at least twice as often as we would expect.

```
. keep if freq>=27
(1942 observations deleted)

. keep if val_min>=2
(1081 observations deleted)
```

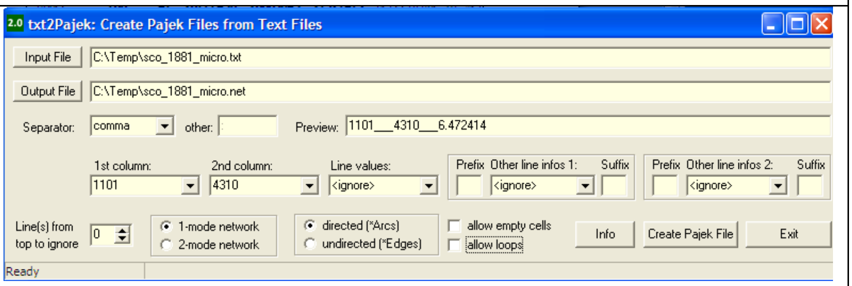
This produces 108 combinations which are both frequently constructed and occur more commonly than expected. As the val_min shows, these occur up to 20 times more than we would anticipate. There are 41 different microclasses for the older cohort and 37 for the younger cohort.

```
Variable     Obs Unique      Mean        Min        Max   Label
------------------------------------------------------------------------
hocc         108     41  2992.852       1101       9990
wocc         108     37  3223.398       1101       9990
freq         108     85  466.1852         29      25594   (count) freq
tot          108      1    258740     258740     258740   total number in s...
nhocc        108     41  4435.796        146      39522   total number of m...
nwocc        108     37  4578.713        281      33491   total number of f...
phocc        108     41  .0171438   .0005643   .1527479   percentage of men...
pwocc        108     37  .0176962    .001086   .1294388   percentage of wom...
ewocc        108    108  114.8665   3.142614   5115.681   expected number o...
prop         108     85  .0018018   .0001121   .0989178
staner       108     85  .0000512   .0000208   .0005869   Standard error fo...
pro_obs      108     85  .0018018   .0001121   .0989178   Observed proporti...
pro_exp      108    108  .0004439   .0000121   .0197715   Expected proporti...
pro_min      108     85  .0017506   .0000913   .0983309   Lower confidence ...
pro_max      108     85  .0018529   .0001329   .0995048   Higher confidence...
value        108    108  4.659531   2.122135   19.83571   Observed value of...
val_min      108    108  4.150986   2.011253   19.30663   Value of represen...
val_max      108    108  5.168075    2.22231   21.30224   Value of represen...
------------------------------------------------------------------------
```

We can then export the data as a comma-separated text file, showing hocc wocc and val_min

```
outsheet hocc wocc val_min using ///
"$path9\sco_1881_micro.txt", ///
comma nonames nolabel replace
```

We can then use txt2Pajek to convert the data into a Pajek file. We select the input file, which then defaults to saving the output file to the same folder. We specify it is comma separated and select the two microclass labels. We then assert it is a one-mode directed network before clicking on 'create Pajek file'.



txt2Pajek: Create Pajek Files from Text Files

Input File: C:\Temp\sco_1881_micro.txt

Output File: C:\Temp\sco_1881_micro.net

Separator: comma other:

Preview: 1101___4310___6.472414

1st column: 1101

2nd column: 4310

Line values: <ignore>

Prefix □ Other line infos 1: <ignore> Suffix □

Prefix □ Other line infos 2: <ignore> Suffix □

Line(s) from top to ignore: 0

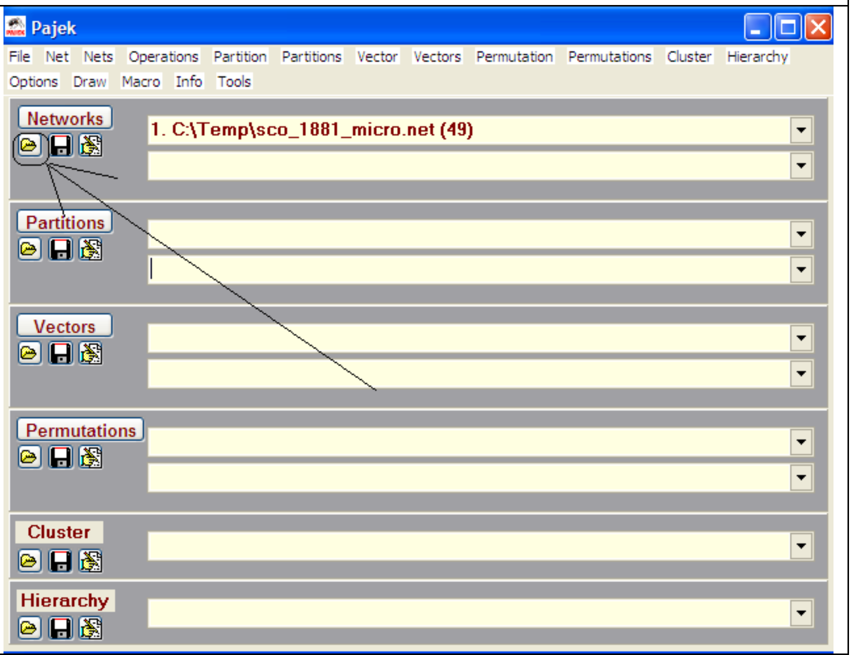
● 1-mode network
○ 2-mode network

● directed (*Arcs)
○ undirected (*Edges)

□ allow empty cells
□ allow loops

Ready

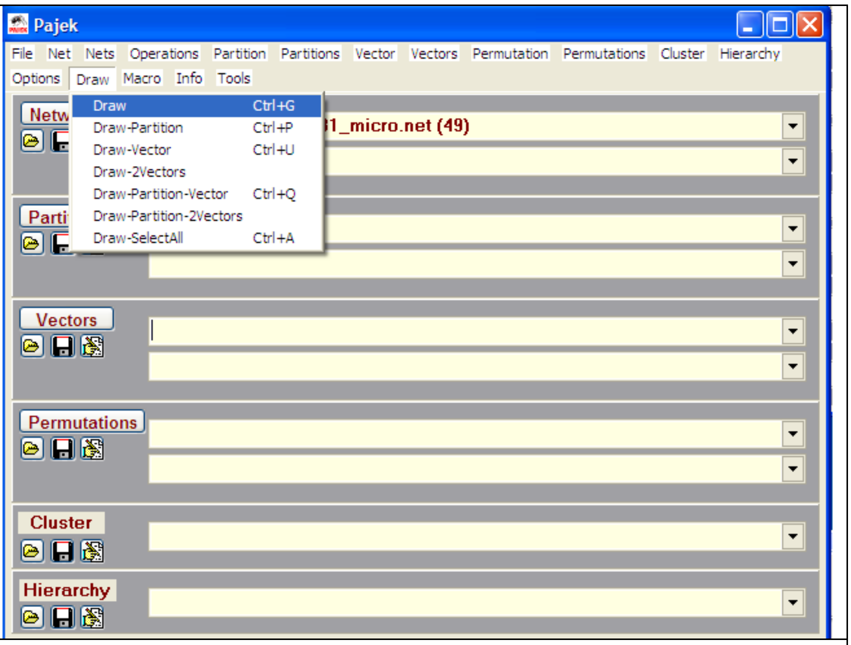
We can then open the file in Pajek, using either the file button below network, or the drop-down menu.



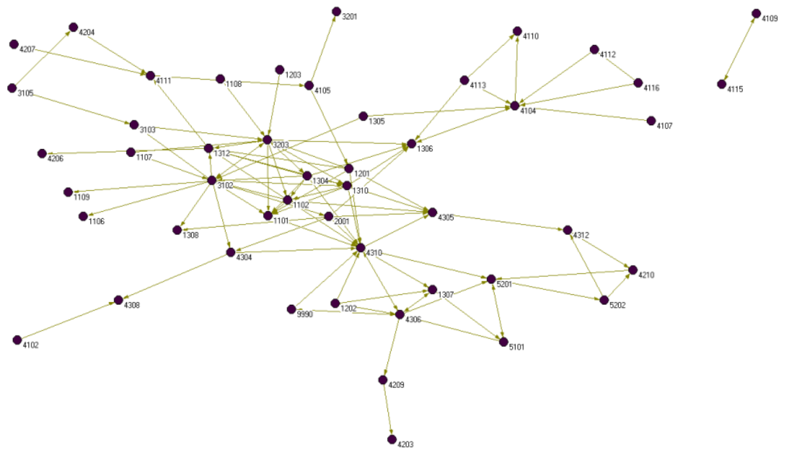
Pajek

Networks: 1. C:\Temp\sco_1881_micro.net (49)

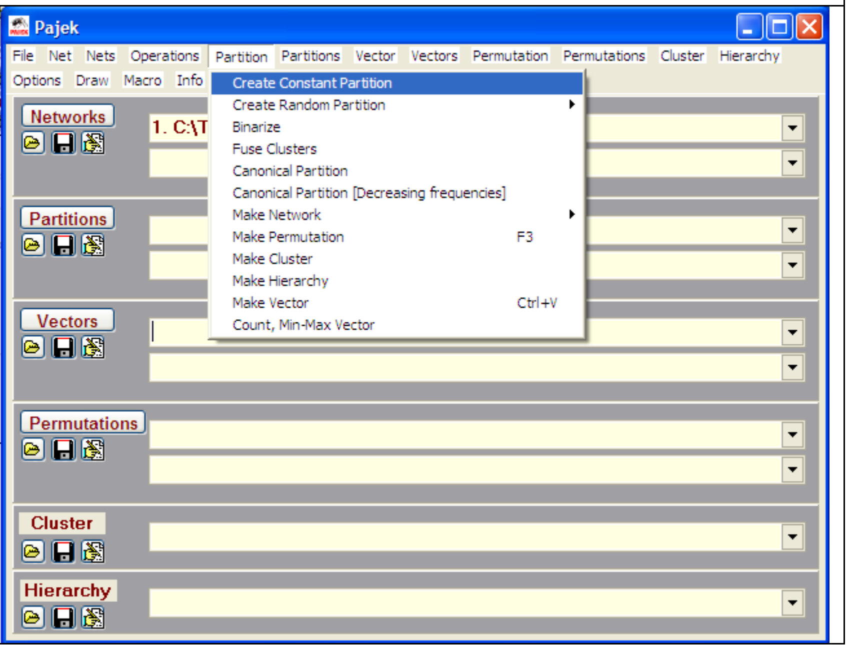
We can then select ‘draw’ or CTRL+G to visual the network.



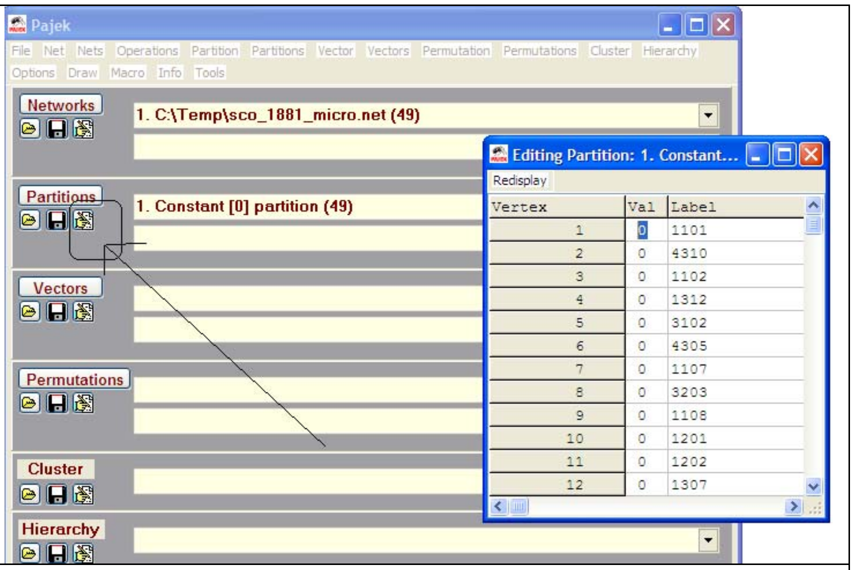
This provides a diagram of the network. The nodes are distributed where the software believes they are best placed. Sometimes it can be a little strange, so pressing CTRL+K will regenerate the network. Clicking on an individual node enables it to be moved around.



We can change the colour of the nodes to see differences more clearly. If we close the graph window we can click on ‘partition’, ‘create constant partition’ to create a blank partition. The first dialogue box regards the size of the partition (which defaults to the number of nodes, as it needs to be). The second dialogue box asks for the constant term to be used (default is 0).



If we click on the last of the three buttons we can edit the partition. This enables us to group the nodes. In this example, we will place the first number of each label as the partition (which enables us to code the nodes by macroclass).

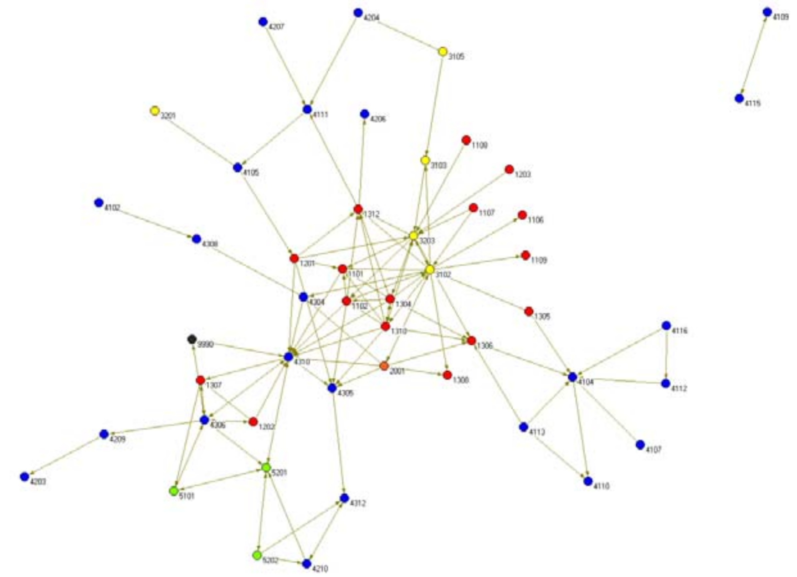

Pajek
File Net Nets Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy
Options Draw Macro Info Tools
Networks
Partitions
Vectors
Permutations
Cluster
Hierarchy
Editing Partition: 1. Constant...
Redisplay

1. C:\Temp\sco_1881_micro.net (49)

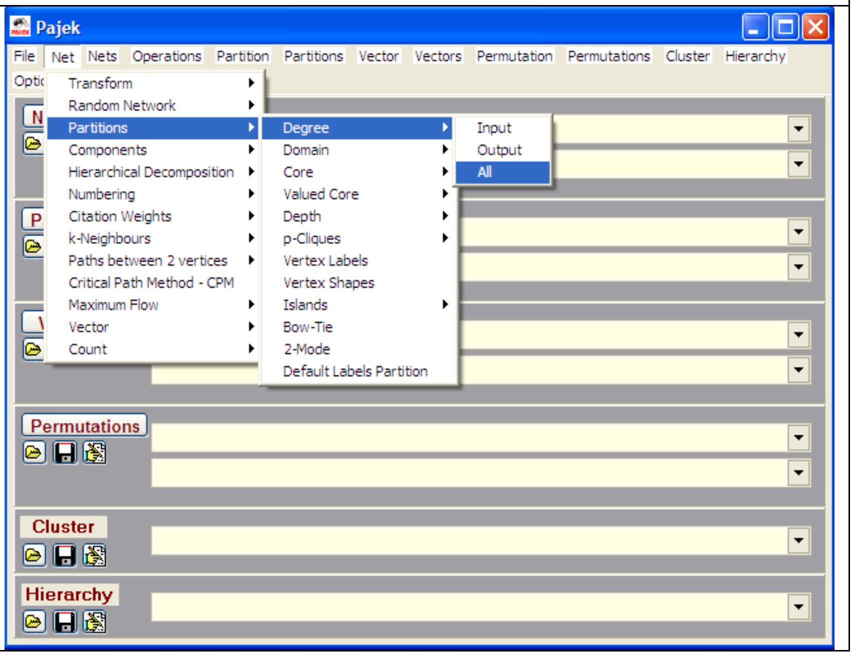
1. Constant [0] partition (49)

Vertex	Val	Label
1	0	1101
2	0	4310
3	0	1102
4	0	1312
5	0	3102
6	0	4305
7	0	1107
8	0	3203
9	0	1108
10	0	1201
11	0	1202
12	0	1307

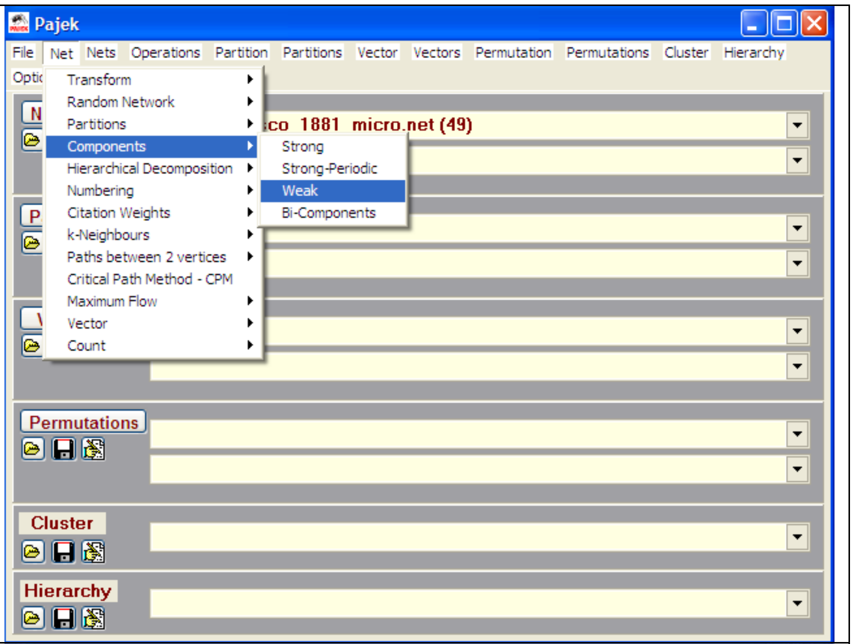
This enables us to see the differences between the macroclasses more clearly. The colours of the partitions can be altered using the 'options', 'colors', 'partition colors', 'of vertices' drop down menu within the draw window. Therefore, the colours might not be the same as displayed.



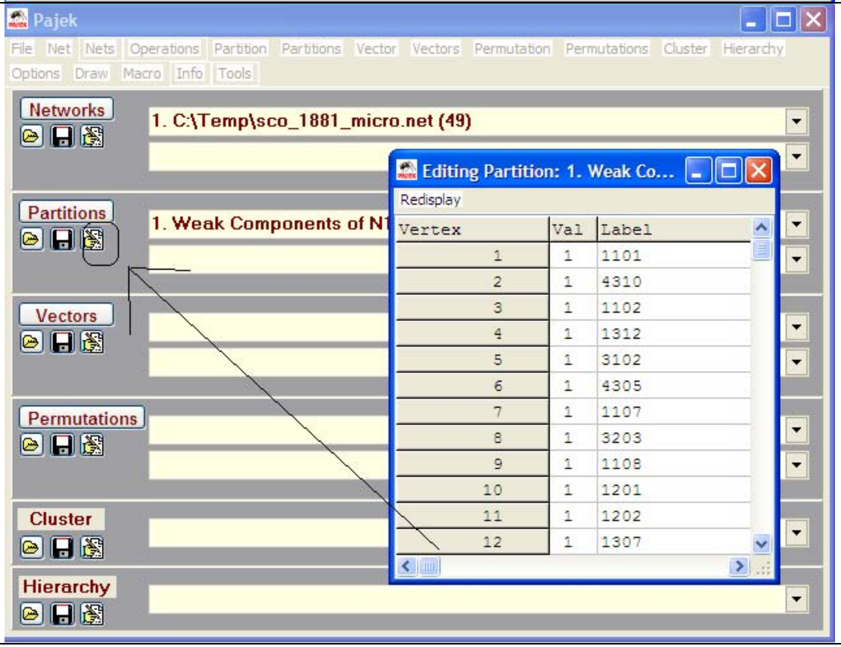
We can also find out statistics about the network. Using 'net', 'partition', 'degree', 'all' we can see the degree centrality of the network (as well as getting the degree for each occupation).



We can see how many components exist, using 'net', 'components', 'weak'. From the dialogue box we select '1' in this example (this is just about the strength of ties needed to form part of the community. We are interested here in nodes which have at least one tie to a member of the sub-population which are connected).



We can then reduce our analysis to just the largest component. This allows us to perform some further analysis. Firstly, we click on the partition named 'components' to see which is largest.

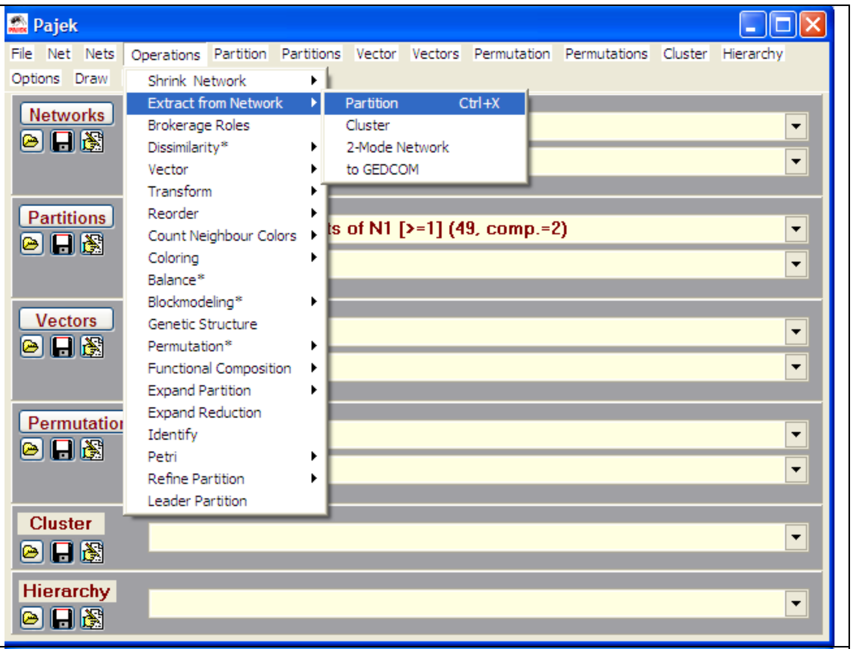


1. C:\Temp\sco_1881_micro.net (49)

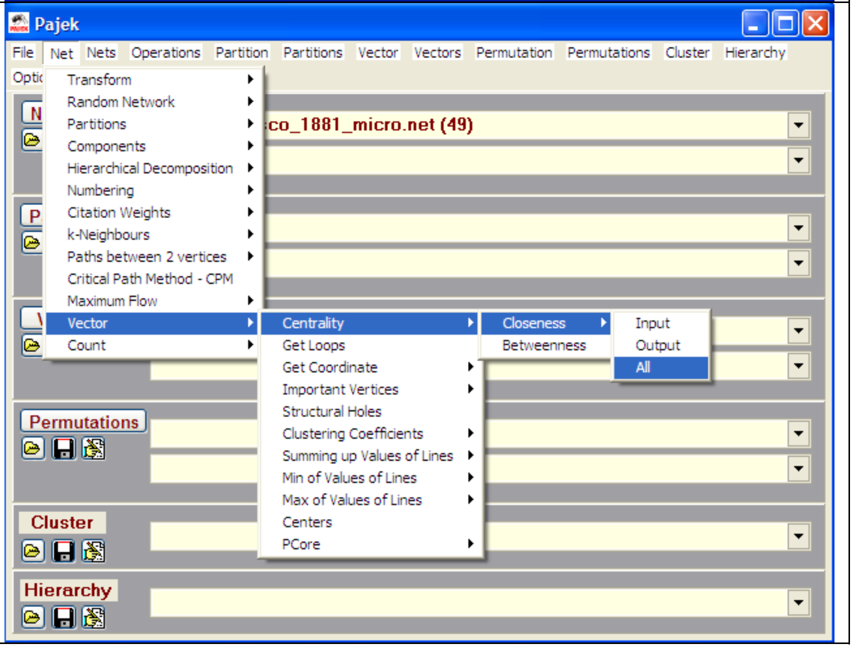
1. Weak Components of N1

Vertex	Val	Label
1	1	1101
2	1	4310
3	1	1102
4	1	1312
5	1	3102
6	1	4305
7	1	1107
8	1	3203
9	1	1108
10	1	1201
11	1	1202
12	1	1307

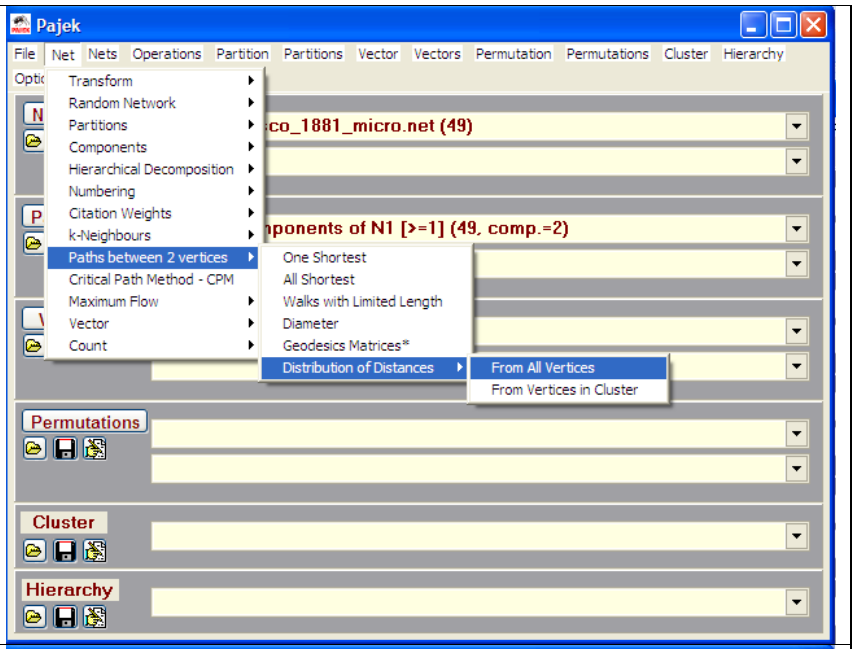
We then click on ‘operations’, ‘extract from network’, ‘partition’ (or CTRL+X) and decide to keep only ‘1’. This removes those cases which do not connect to the main component.



We can use ‘net’, ‘vector’, ‘centrality’, ‘closeness’, ‘all’ to see the closeness centrality.



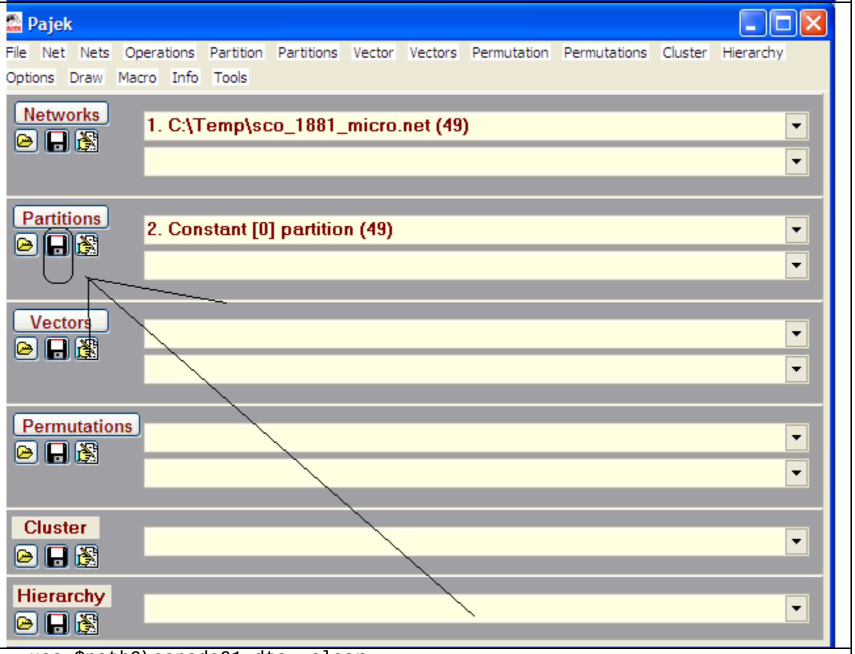
We can use ‘net’, ‘Paths between 2 vertices’, ‘distribution of distances’, ‘from all vertices’ to see the average distance between the nodes.



co_1881_micro.net (49)

nponents of N1 [>=1] (49, comp.=2)

We can then save the partition we created to ensure we have the data next time we analyse it.



1. C:\Temp\sco_1881_micro.net (49)

2. Constant [0] partition (49)

We can analyse more than just microclasses. In this second example we will look at Canada 1891 data, looking at religion. Again, a dataset has been created.

```
. use $path9\canada91.dta, clear

. codebook, compact

Variable      Obs Unique      Mean       Min     Max  Label
--------------------------------------------------------------------------
serial      21699   7675  51919.54        12   76715  Household index ...
age1        21699     48   51.6469        28      75  Age
sex1        21699      1         1         1       1  Sex
hisco1      21699    312  60062.34      1110   98990
microclass1 21699     71  3658.492      1101    9990
age         21699     46  23.66455        12      57  Age
sex         21699      1         1         1       1  Sex
hisco       21699    337  62402.54      1110   98990
microclass  21699     70  3820.816      1101    9990
religion1   21699     66  3595.099      1100    9997  Religion, first ...
religion    21699     61  3609.406      1100    9997  Religion, first ...
hiscam1     21699     68  58.69302  45.69142      99  (mean) hiscam
hiscam      21699     68  57.72059  45.69142      99  (mean) hiscam
--------------------------------------------------------------------------
```

From the religion data, we can dichotomise the microclasses by placing a 1 in front of the microclasses for Catholics.

```
. capture drop cath*

. gen cath=religion==1100

. gen cath1=religion1==1100

. capture drop hocc

. gen hocc=microclass

. replace hocc=microclass+10000 if cath==1
(30927 real changes made)

. capture drop wocc

. gen wocc=microclass1

. replace wocc=microclass1+10000 if cath1==1
(29994 real changes made)
```

This enables us to run the analysis as above.

```
do http://www.camsis.stir.ac.uk/sonocs/do/pajek.do
```

This provides us with over 4,000 combinations, although many of these are of little relevance to us.

```
Variable      Obs Unique      Mean       Min       Max  Label
---------------------------------------------------------------------
hocc         4471    137  7575.914      1101     19990
wocc         4471    138  7164.323      1101     19990
freq         4471     77   4.66115         1      1111  (count) freq
tot          4471      1     20840     20840     20840  total number in ...
nhocc        4471    107   265.926         1      1566  total number of ...
nwocc        4471    102  309.9512         1      2662  total number of ...
phocc        4471    107  .0127604   .000048   .075144  percentage of me...
pwocc        4471    102  .0148729   .000048  .1277351  percentage of wo...
ewocc        4471   3832  3.305061  .0004319  200.0332  expected number ...
prop         4471     77  .0002237   .000048  .0533109
staner       4471     77  .0000831   .000048  .0015562  Standard error f...
pro_obs      4471     77  .0002237   .000048  .0533109  Observed proport...
pro_exp      4471   3818  .0001586  2.07e-08  .0095985  Expected proport...
pro_min      4471     77  .0001406  1.15e-09  .0517548  Lower confidence...
pro_max      4471     77  .0003067   .000096  .0548671  Higher confidenc...
value        4471   3968  5.717776  .0432525  2315.556  Observed value o...
val_min      4471   4217  1.115635  1.04e-06  81.03288  Value of represe...
val_max      4471   4214  10.31992  .0691135  4631.055  Value of represe...
---------------------------------------------------------------------
```

We set limits of frequency of 5 and val_min of 2.

```
.
. keep if freq>=5
(3546 observations deleted)

. keep if val_min>=2
(617 observations deleted)
```

This provides us with over 300 combinations of microclass_religion linkages.

```
Variable     Obs Unique      Mean       Min       Max  Label
---------------------------------------------------------------------
hocc         308     92  8767.581      1101     15202
wocc         308     91  8542.578      1101     15202
freq         308     57  22.03571         5      1111  (count) freq
tot          308      1     20840     20840     20840  total number in s...
nhocc        308     85  273.1526        14      1566  total number of m...
nwocc        308     77    330.25        12      2662  total number of f...
phocc        308     85  .0131071  .0006718   .075144  percentage of men...
pwocc        308     77  .0158469  .0005758  .1277351  percentage of wom...
ewocc        308    304  4.442368  .0735605  200.0332  expected number o...
prop         308     57  .0010574  .0002399  .0533109
staner       308     57  .0001857  .0001073  .0015562  Standard error fo...
pro_obs      308     57  .0010574  .0002399  .0533109  Observed proporti...
pro_exp      308    304  .0002132  3.53e-06  .0095985  Expected proporti...
pro_min      308     57  .0008717  .0001326  .0517548  Lower confidence ...
pro_max      308     57  .0012431  .0003472  .0548671  Higher confidence...
value        308    304   9.38363  2.268401  109.2243  Observed value of...
val_min      308    306  6.628953  2.010274  81.03288  Value of represen...
val_max      308    306  12.13831  2.521529  137.4158  Value of represen...
---------------------------------------------------------------------
```

We can then export the data as above.

```
outsheet hocc wocc val_min using ///
"$path9\ca_1891_micro_cath.txt", ///
comma nonames nolabel replace
```

We can then rerun the analysis in Pajek, as above.

Follow the steps from the txt2pajek example, using the Canadian data. The practices in Pajek are the same, regardless of the type of data.

(Note: When we create the partition to distinguish between Catholics and non-Catholics, if you ask it to set a constant partition with a value of 0, you can simply change the value to a '1' if it is a five-digit number (i.e., it has a one added to the front).

If you wish, you might want to code the macroclass additionally (the first number if 4 digits, first two if 5 digits). This will create potential values of 1-5, 9, 11-15 & 9. Using 'options', 'colors', 'partition colors',

	'of vertices' in the graph window, you can click on the colors and assign them a new value. This would enable a 'bluescale' system for Catholics and a 'redscale' for non-Catholics akin to the usual 'greyscale' we see of the macroclasses being represented by brightness and the religion represented by colours).

Selected references

Stata

Kohler, H. P., & Kreuter, F. (2009). *Data Analysis Using Stata, 2nd Ed* College Station, Tx: Stata Press.

Leckie, G., & Charlton, C. (2011). *runmlwin: Running MLwiN from within Stata.* Bristol: University of Bristol, Centre for Multilevel Modelling, http://www.bristol.ac.uk/cmm/software/runmlwin/ [accessed 1.6.2011].

Long, J. S. (2009). *The Workflow of Data Analysis Using Stata*. Boca Raton: CRC Press.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and Longitudinal Modelling Using Stata, Second Edition.* . College Station, Tx: Stata Press.

Treiman, D. J. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. New York: Jossey Bass.

Web: http://www.longitudinal.stir.ac.uk/Stata_support.html

R

Bates, D. M. (2005). Fitting linear models in R using the lme4 package. *R News, 5*(1), 27-30.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression, 2nd Ed.* London: Sage.

Gelman, A., & Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Spector, P. (2008). *Data Manipulation with R (Use R)*. Amsterdam: Springer.

(There is also a useful guide to using R, within SPSS, in the body of Levesque & SPSS Inc 2010, cited above).

Web: http://www.ats.ucla.edu/stat/r/

Pajek

de Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory Social Network Analysis with Pajek*. Cambridge: Cambridge University Press.

Other references of relevance

Crouchley, R., Stott, D., & Pritchard, J. (2009) *Multivariate generalised linear mixed models via sabreStata (Sabre in Stata).* Lancaster: Centre for e-Science, Lancaster University.

Dale, A. (2006). Quality Issues with Survey Research. *International Journal of Social Research Methodology, 9*(2), 143-158.

Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods and Research, 36*(2), 153-171.

Rafferty, A., & Watham, J. (2008). *Working with survey files: using hierarchical data, matching files and pooling data*. Manchester: Economic and Social Data Service, and http://www.esds.ac.uk/government/resources/analysis/.