WHO-EM/STD/160/E

Introduction to HIV/AIDS and sexually transmitted infection surveillance

MODULE 4

SUPPLEMENT
A guide to using RDS Analyst and NetDraw

World Health Organization
Regional Office for the Eastern Mediterranean

UNAIDS

WHO-EM/STD/160/E

Introduction to HIV/AIDS and sexually transmitted infection surveillance

MODULE 4

SUPPLEMENT
A guide to using RDS Analyst and NetDraw

World Health Organization
Regional Office for the Eastern Mediterranean

UNAIDS

WHO Library Cataloguing in Publication Data

World Health Organization. Regional Office for the Eastern Mediterranean
Introduction to HIV/AIDS and sexually transmitted infection surveillance: Module 4. Supplement: a guide to using RDS analyst and NetDraw / World Health Organization. Regional Office for the Eastern Mediterranean, Joint United Nations Programme on HIV/AIDS
p.
WHO-EM/STD/160/E
1. Acquired Immunodeficiency Syndrome – epidemiology 2. HIV Infections – epidemiology 3. Sexually Transmitted Diseases – epidemiology 4. Data Interpretation, Statistical I. Title II. Regional Office for the Eastern Mediterranean III. Joint United Nations Programme on HIV/AIDS

(NLM Classification: WC 503.4)

© World Health Organization 2014

All rights reserved.

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by the World Health Organization in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by the World Health Organization to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall the World Health Organization be liable for damages arising from its use.

Publications of the World Health Organization can be obtained from Knowledge Sharing and Production, World Health Organization, Regional Office for the Eastern Mediterranean, PO Box 7608, Nasr City, Cairo 11371, Egypt (tel: +202 2670 2535, fax: +202 2670 2492; email: emrgoksp@who.int). Requests for permission to reproduce, in part or in whole, or to translate publications of WHO Regional Office for the Eastern Mediterranean – whether for sale or for noncommercial distribution – should be addressed to WHO Regional Office for the Eastern Mediterranean, at the above address: email: emrgogap@who.int.

Design by Permanent Impressions.

Contents

1	**Step-by-step guide to using RDS Analyst**	**5**
	Introduction	7
	Installing RDS-A	8
	Console and data viewer	10
	Working with data	21
	Analysis of the sample	22
	Analysis of the population	27
	Plots to visually explore RDS data	33
	Conclusion	38
	References for the estimator table	39
2	**Using NetDraw for interpreting RDS data and for problem solving**	**41**
	Getting started with NetDraw	43
	Prepare data for NetDraw	43
	Generating recruitment chains	46

Step-by-step guide to using RDS Analyst

1. Step-by-step guide to using RDS Analyst

Introduction

Why RDS Analyst (RDS-A)?

RDS-A is a software package for the analysis of respondent-driven sampling (RDS) data. RDS-A was conceived and developed by the Hard-to-Reach Population Methods Research Group (HPMRG) as an easier and more useful alternative to other available software for the analysis of data collected through RDS methods. There are several reasons for using RDS-A, including that:

- it is free
- it is intuitive and easy to use with point and click commands
- it allows for programming in the computer package R at the batch/command line
- code is reusable and can be copied from the batch/command line
- it can analyse multiple data sets at once
- it can analyse multiple variables at once
- it easily reads multiple files (SPSS Statistics, STATA, SAS, Microsoft Excel, text in RDSAT format) without having to modify them
- it has the latest estimators
- it allows for many types of analyses including population and sample estimates, population confidence intervals, cross tabs, means, medians, percentiles, standard deviations, design effect, and proportional analyses (bi-variate with confidence intervals)
- it produces diagnostic plots
- it transforms and recodes variables
- data can be viewed and edited in a Microsoft Excel-like spreadsheet.

Overview of the software

In order to provide a free and easy-to-use software alternative to proprietary data analysis software there are several layers to RDS-A.

- The base layer of RDS-A is R, a free, open-source and powerful computer package that can be downloaded at: http://www.r-project.org/. If you know R statistical language, you can make your own commands at the batch/command line.
- The next layer is JGR (Java graphical user interface [GUI]) designed for use with R (for more information see: http://rforge.net/JGR/).

- The next layer is Deducer (for more information see: http://www.deducer.org/pmwiki/index.php?n=Main.DeducerManual). This layer provides an intuitive GUI for R which allows non-technical users to learn and perform analyses by pointing and clicking on commands.
- The final layer is the RDS-A point-and-click GUI.

Installing RDS-A

HPMRG web site

To download RDS-A, go to the HPMRG web site (www.hpmrg.org) (see Figure 1.1).

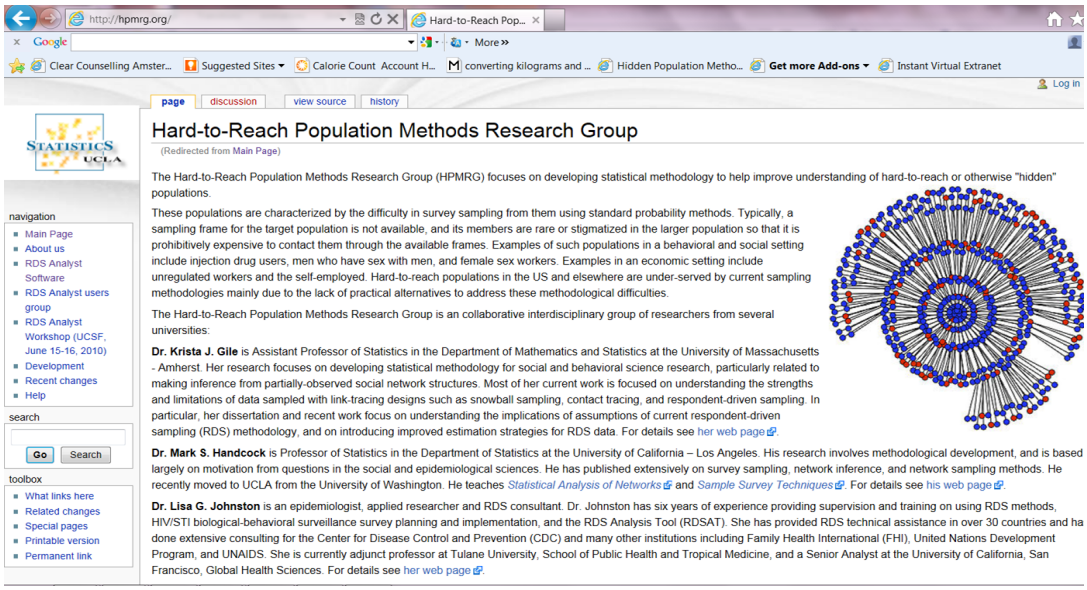


Figure 1.1 Example of the HPMRG web site for downloading RDS-A

At the left hand side is a navigation bar. Click on "RDS Analyst Software". This will take you to the page shown in Figure 1.2.[1]

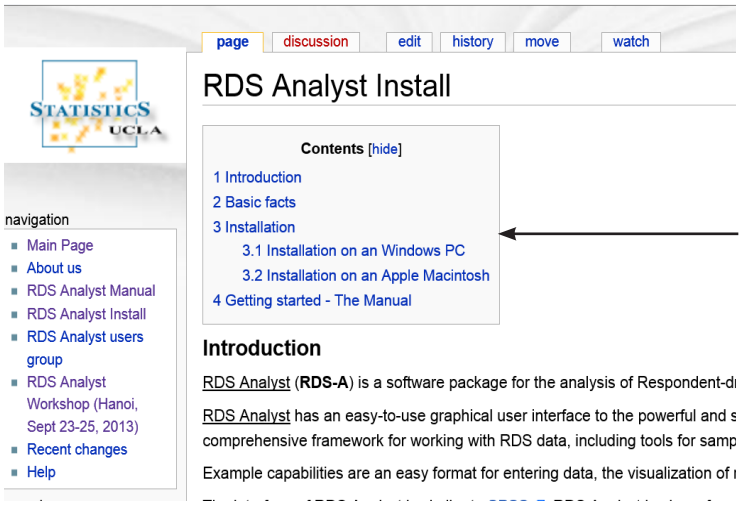


Figure 1.2 Page for downloading RDS-A

[1] You may have to log in to access the software package.

Download RDS Analyst

The current development form of RDS-A is for Microsoft Windows and Apple Macintosh. Click on "Installation on a Windows PC" or "Installation on an Apple Macintosh" (see Figure 1.2) depending on your needs. Once you click on either of the installations in the contents, you can click on the installers directly to download RDS-A (see Figure 1.3).

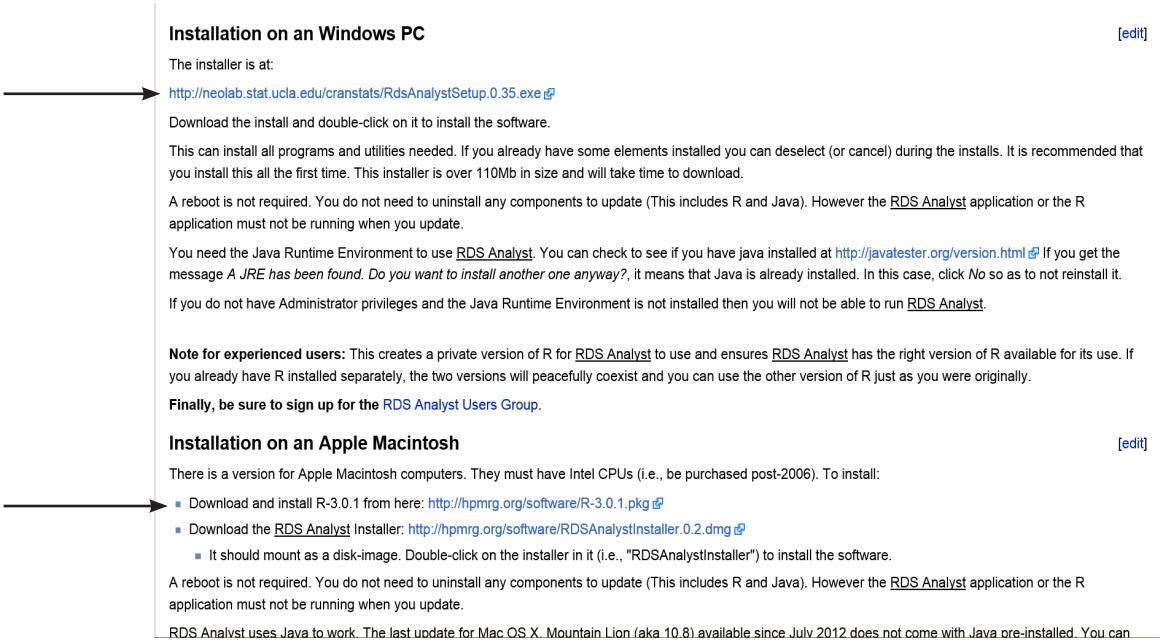


Figure 1.3 RDS-A installation

Below are detailed descriptions of downloading RDS-A for Microsoft Windows or Apple Macintosh.

RDS-A for Microsoft Windows

The Windows installer is at: http://hpmrg.org/software/RDSAnalystSetup.0.1.exe

Download the installer and double-click on it to install the software. To ensure your installation is the latest version of the packages, use the updater at:

http://hpmrg.org/software/RDSAnalystUpdater.0.1.exe

and for the latest and secure version of Java:

http://hpmrg.org/software/jre-7u11-windows-i586.exe

This will install all programs and utilities needed. If you already have some elements installed you can deselect (or cancel) during the installation. It is recommended that you install the entire package the first time you install RDS-A.[1] This installer is over 110Mb in size and will take time to download. A reboot is not required. You do not need to uninstall any components in order to update (this includes R and Java). However, the RDS-A application or the R application must not be running when you update.

You need the Java Runtime Environment to use RDS-A. If you get the messages "a JRE has

[1] Note for experienced users: This creates a private version of R for RDS Analyst to use and ensures RDS-A has the correct version of R available for its use. If you already have R installed separately, the two versions will peacefully coexist and you can still use the other version of R.

been found" and "Do you want to install another one anyway?", it means that Java is already installed. In this case, click "No" if you do not want to reinstall it. After you install, you should ensure your installation is the latest version of the packages by downloading the updater: http://hpmrg.org/software/RDSAnalystUpdater.0.1.exe.

RDS-A for Apple Macintosh

For Apple Macintosh users, it is necessary to have Intel central processing units purchased after 2006. Download and install R-2.15.2 from: http://hpmrg.org/software/R-2.15.2.pkg. Then download the RDS-A installer at: http://hpmrg.org/software/RDSAnalystInstaller.0.1.dmg. It should load as a disk-image. Double-click on the installer in it (i.e. "RDSAnalystInstaller") to install the software. A reboot is not required. You do not need to uninstall any components to update (this includes R and Java). However, the RDS-A application or the R application must not be running when you update. RDS-A uses Java to work. [1] The RDS-A application and R will be in your applications folder. To run RDS-A, double-click on it in the applications folder.

After you install, use the updater to ensure your installation is the latest version of the packages. Download the RDS-A updater at: http://hpmrg.org/software/RDSAnalystUpdater.0.1.dmg. It should load as a disk-image. Double-click on the installer in it (i.e. "RDSAnalystUpdater") to install the software. This just installs the core packages (that is, anything that has changed since the full installation was made). It will typically be a few Mb in size.

Desk icon

Once you download RDS-A, you will have a desk icon that looks like Figure 1.4:

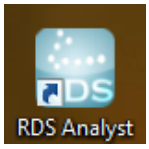


Figure 1.4 Desk icon

Click this icon to launch RDS-A. Alternatively, if you did not install this desktop icon, you will need to go to programs to launch RDS-A.

Console and data viewer

Introduction

RDS-A has two main windows, the "Console" (Figure 1.5) and the "Data viewer" (Figure 1.6), which should open when you download the program and/or when you launch the program using either the desktop icon or from the programs folder.

[1] To use RDS-A, you need Java installed on your Apple Macintosh. If you are using Mac OS X 10.6 and below, Java comes pre-installed. If you are using Mac OS X 10.7 (Lion) and above, then Java is not pre-installed. You can check to see if you have Java installed at http://javatester.org/version.html. To install Java version 6, open the "Java Preferences.app" located in the Applications > Utilities folder on your Apple Macintosh. It will ask if you want to install Java if it is not already there. Accept its invitation.

Console

The primary window is the Console. The menus at the top allow you to open, save and edit files and to do statistical analysis, plots and diagnostics. Each of the menu items are discussed in more detail below. The panel along the left side is a navigation panel. You can use it to quickly go from one statistical output to another. You can also remove output that you do not want by selecting it in the navigation panel and pressing the "Remove" button at the bottom of the navigation panel. The panel along the bottom is a command console. You can type in commands in the R language here.

The panel in the upper right, which takes up most of the screen, is an output window. By default, RDS-A commands that are submitted for execution show up as red text and the generated output shows up as black text. There are two tabs at the top of the output panel. The "Console View" shows all the output, like one long ream of paper. The "Element View" shows you just a single output element at a time, without any red RDS-A commands. You can click between these to get a view of the history of the commands and output. The Console also records a log of all the commands and output produced from them.

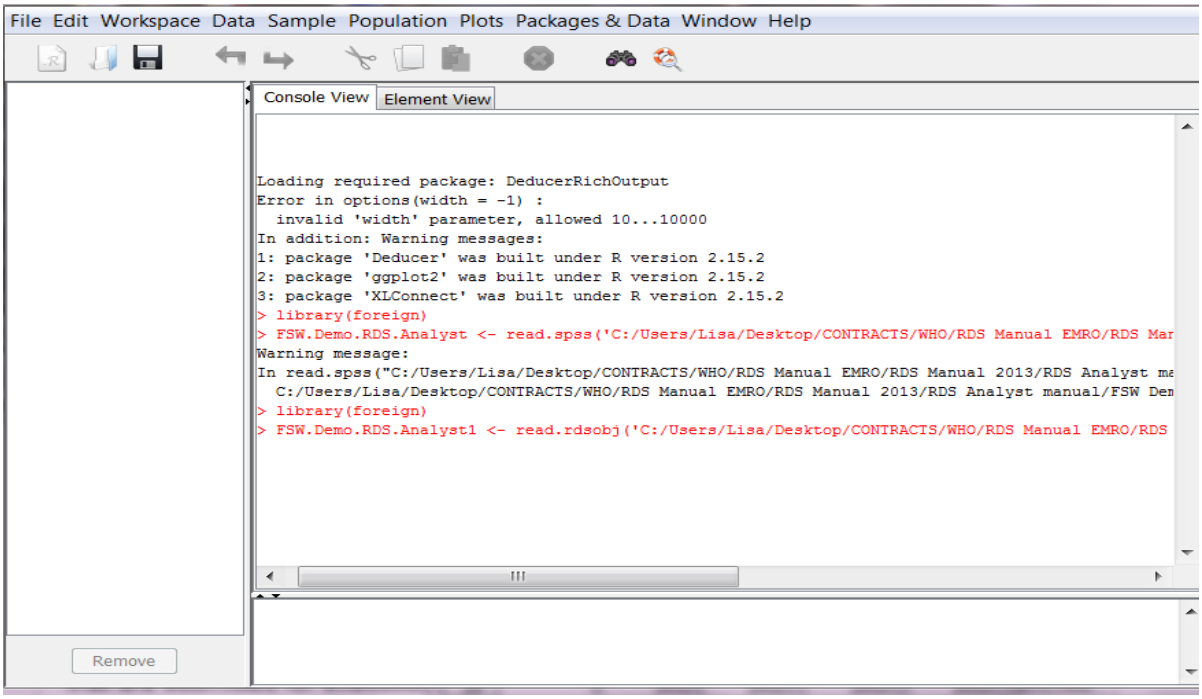


Figure 1.5 RDS-A Console

Learning the tabs on the Console

The Console has ten tabs: File, Edit, Workspace, Data, Sample, Population, Plots, Packages and Data, Window and Help. The tab features identified in bold in Table 1.1 are featured in more detail in this guide. Many of the tab features are intuitive and do not need further instruction.

Table 1.1 RDS-A Console

TAB NAME	FEATURES UNDER TAB	MAIN PURPOSE OF TAB
File	**Open Data**, **Save Data**, New Document, Open Document, **Save**, Set Working Directory, Quit	Open and save data sets, set default location for data and save output using the Set Working Directory.
Edit	Undo, Redo, Cut, Copy, Copy Special, Paste, Delete, Select All, Find, Find Next, Clear Console, Increase or Decrease Font Size, Preferences	Edit features. "Preferences" allows you to change defaults such as output styles.
Workspace	Open, Save, Save As, Clear All	Objects created in RDS-A are held in computer memory as a workspace and are not saved on disk unless told to do so. To save objects when RDS-A is closed, save your workspace when exiting. Save and open (previously saved) workspaces from this menu. For multiple projects, save entire workspace for each project in a separate file. Then open them from this menu. "Clear All" empties workspace (i.e. removes all objects).
Data	Edit Factor, **Recode Variables**, Transform, **Compute Weights**, Reset Row Names, Sort, Edit Meta Data, Convert to RDS (Coupon Format, Recruiter ID Format)	Edit Factor: Adds or subtracts values of a categorical variable. Recode Variables: Recodes variables into existing or new variables. Transform: Makes complicated changes to variables (see transformation panel). Edit Meta Data: Specifies characteristics of RDS data (see Figure 1.11). Convert to RDS: Direct and manual way to (re)convert a data set into an RDS data set. It is rarely used (as the program does this automatically), but is useful if the automatic method missed something.
Sample	**Frequencies, Descriptives, Contingency Tables, Recruitment Homophily**	Unadjusted exploratory analysis of continuous, categorical and descriptive data. Frequencies: Tables of one or more variables, possibly stratified by others. Descriptives: Unadjusted summary statistics. Contingency Tables or Crosstabs: Includes test statistics. Recruitment Homophily: Statistic of differential recruitment in the sample.

TAB NAME	FEATURES UNDER TAB	MAIN PURPOSE OF TAB
Population	**Frequency Estimates, Descriptive Estimates, Population Crosstabs, Test Difference in Proportions, Population Homophily,** Differential Activity	Adjusted statistics of continuous, categorical and descriptive data to describe the population sampled. Frequency Estimates: Tables of one or more variables, possibly stratified by others. Descriptive Estimates: Adjusted summary statistics. Population Crosstabs: Cross tabulations of categorical variables. Population Test Differences in Proportions: Tests hypothesis that two population proportions are equal. Population Homophily: Statistic of differential recruitment in the population. Differential Activity: Ratio of the mean network size for those with the outcome to the mean network size of those without it.
Plots	**Plot Recruitment Tree, Diagnostics,** Plot Builder, Import Template, Open Plot, Quick, Interactive	Plots to visually explore RDS data. Plot Recruitment Tree: Graphics plot of the recruitment tree. Recruitment Diagnostics: Bar plot of number of recruits by wave, scatter plot of network size verses wave, bar chart of number of recruits from each seed, histogram of number of recruits for each respondent, boxplots by wave, seed, etc. Plot Builder: Interface to create simple and sophisticated plots, including pie charts, histograms, bar plots, scatter plots, bubble plots, etc. The Plot Builder, allows the import of an existing template (Import Template), opening an existing plot (Open Plot), or working interactively with existing plot primary forms (Quick and Interactive).
Packages & Data	Object Browser, Data Viewer, GUI Add-ons, Package Manager, Package Installer, Example: Faux, Example: FauxMadrona, Example: Fauxsycamore	To re-open the Console and Data Viewer, install additional packages for underlying R engine and look at packages currently loaded, edit and view any "objects" in workspace, such as RDS data sets (spreadsheets with the prefix: rds), spreadsheets (data frames), functions, etc. There are three example data sets.
Window	Close Window, Close Same Type, Close All, Console, Data Viewer	Use this to go between the Console and Data Viewer windows, to choose a graphics window. Also, lists currently open windows to choose from and to choose a window to bring to the front to work on.
Help	R Help, About, Deducer Help, RDS Analyst Introduction Manual, RDS Analyst Reference Manual, Citation Information	Help for R, Deducer and RDS-A, and suggested citation for publications.

Data viewer

The other window is titled "Data Viewer". This is where data are displayed and can be edited. The Data Viewer provides an easy-to-use, spreadsheet-like environment to view and edit data. Copy and pasting is supported, and is compatible with Microsoft Excel 2003/2007, so data can be moved from Microsoft Excel by simply copying it to the Data Viewer. Contextual menus can also be used to insert, delete and copy rows and columns.

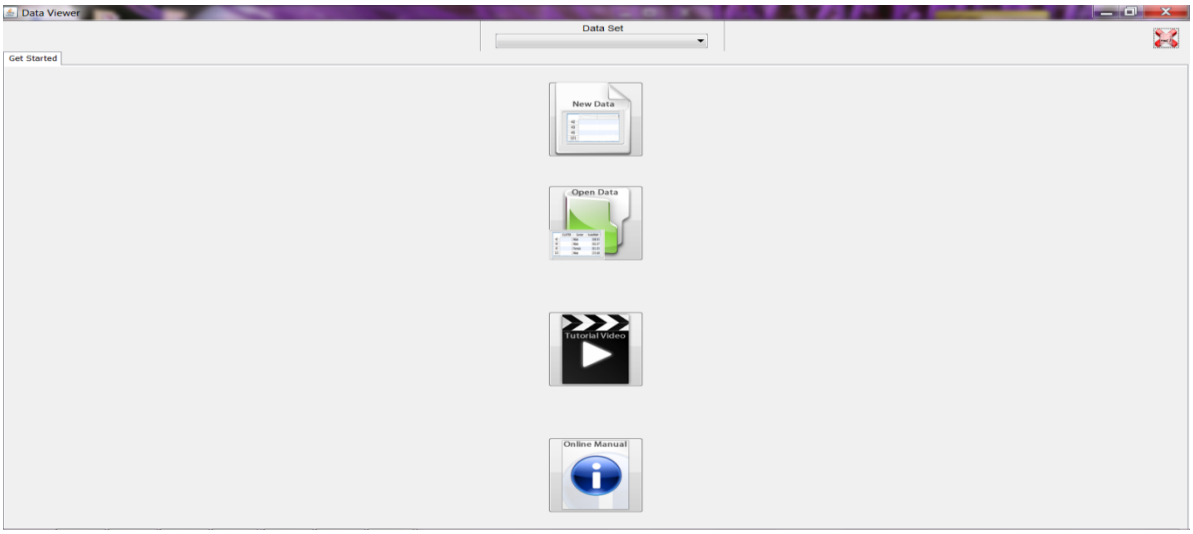


Figure 1.6 RDS-A Data Viewer

Open data

There are two ways to open data:

1) Go to the Data Viewer and select the "Open Data" icon (see Figure 1.7)

Figure 1.7 Open Data icon

2) Go to the top left-hand corner of the Console, go to File |Open Data| (Figure 1.8) and select the file you want to load.

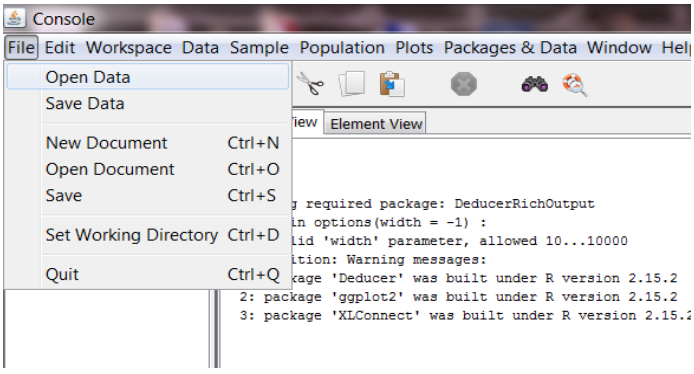


Figure 1.8 Open Data from Console

Once you select the file you want to use, click on it. The example below (Figure 1.9) is loading a SPSS Statistics file titled FSW Demo RDS Analyst.

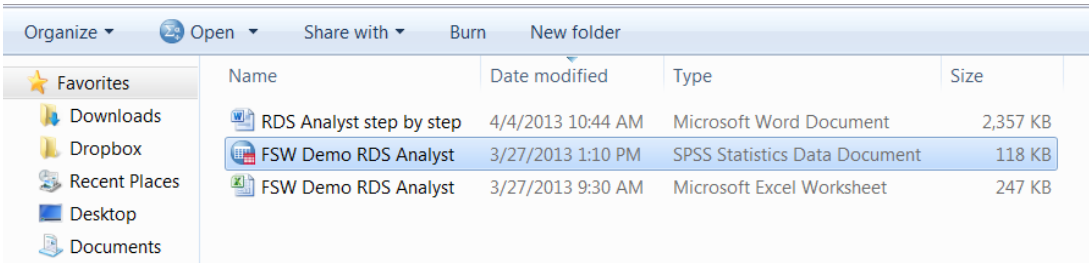


Figure 1.9 Selecting a file to open

Loading data

Once you click on your file, the "Load RDS Data" window will open (Figure 1.10). In this window the variables are located in the panel on the left-hand side. Your data should represent valid RDS data which includes a row for each respondent (i.e. case), and columns for each survey response variable. Most data will be in Data Format|Coupon. Required columns of data include unique IDs and network size data for each participant. In addition, the file should have coupon numbers formatted such that there is one column for the participant's coupon number and additional columns for each of the coupon numbers connected to that participant.

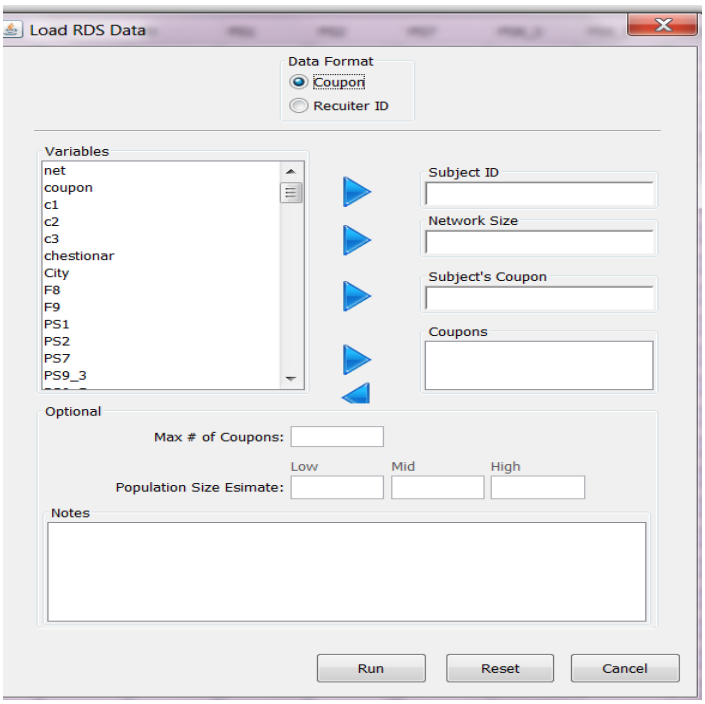


Figure 1.10 Load RDS Data window

The following pieces of information[1] are entered so that RDS-A can properly read the data file:

1. "Subject ID": These are unique ID numbers which can be serial numbers starting from 1 to your final sample size or some other unique ID number (do not use the coupon numbers as your unique ID number). In the example below (Figure 1.11), the unique ID number is named "chestionar", which is each participant's unique questionnaire number and is the 6th variable in the data set (see Figure 1.10).
2. "Network Size": The number of people each participant reported as being part of their personal social network. This variable is named "net" (Figure 1.11.) and was the first variable in the data set (see Figure 1.10).
3. "Subject's Coupon": This is the coupon that each participant used to enrol in the survey. This variable is named "coupon" in the data set (Figure 1.11).
4. "Coupons": These are the coupons given to each participant for them to use in recruiting their peers into the survey. These variables are named C1, C2, and C3 (there were a maximum of three coupons used in this study) in the data set (Figure 1.11).
5. "Max # of Coupons" (optional): This is the maximum number of coupons used in the survey. For instance, if during the beginning of the survey four coupons were provided to each participant to use in recruiting others and this was reduced to three coupons in the middle of the survey and then two near the end of the survey, the maximum number of coupons would be four. There were a maximum of three coupons used in this survey (Figure 1.11)
6. "Population Size Estimate" (optional if you use the default): The mid-range population size estimate for the survey below is 1500 (Figure 1.11). You only need to put in the mid range, but can also put in the low and high range if you know that. This is used in some of the computations. An accurate population size estimation is often unknown, so it is all right to make an approximate guess which should not be close to the sample size. The defaulted low estimate of the population size is 500, the mid estimate is 1000 and the high estimate is 1500.

Once you completed these steps, click "Run".

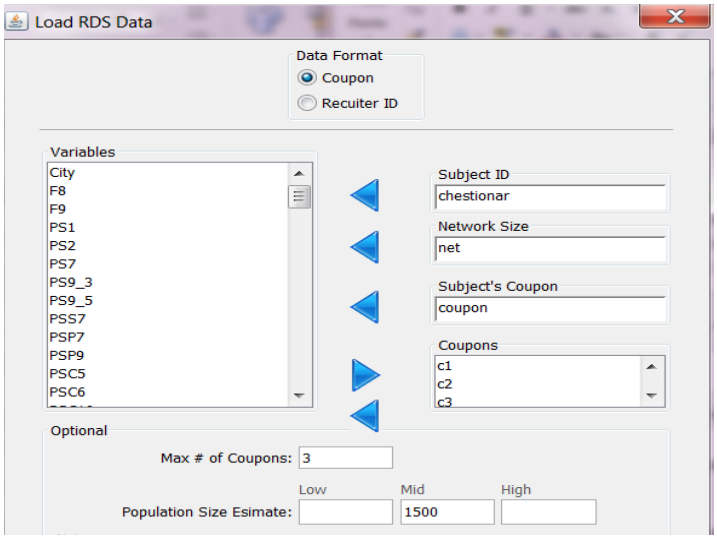
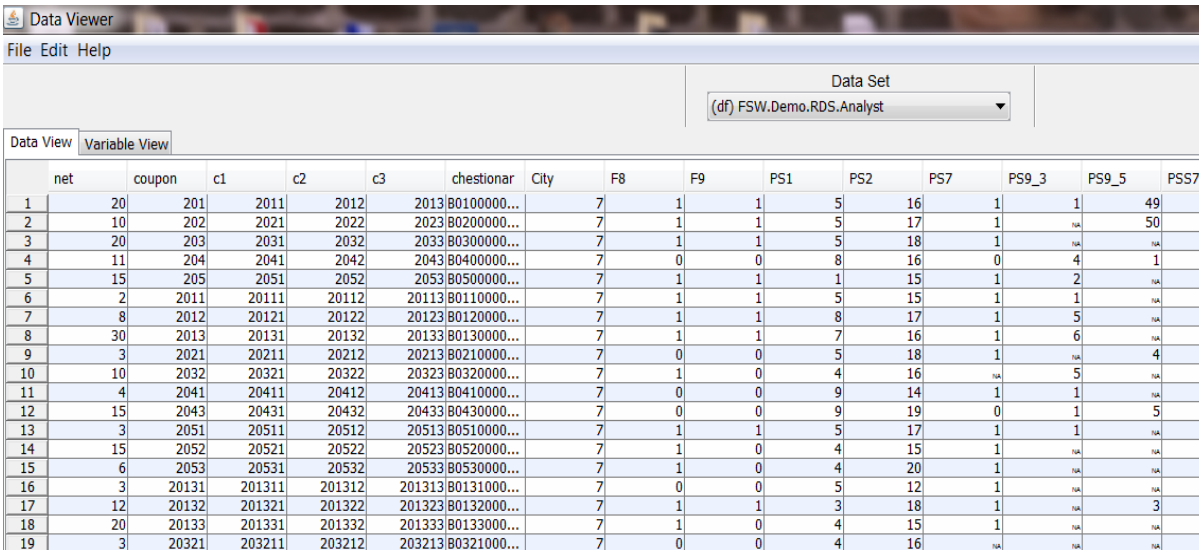


Figure 1.11 Loading the RDS data window: partial view

[1] Note that the pieces of information entered do not need to be in any special order in your data set (for instance, the variable "net" could have been at the end of the data set rather than the first variable of the data set).

Data Viewer

When data are loaded the "Data Viewer" window will open. There are two tabs in the "Data Viewer" window: "Data View" and "Variable View". The "Data View" looks similar to a Microsoft Excel worksheet (Figure 1.12). At the top, centre, is the name of the data set: FSW.Demo.RDS.Analyst.



	net	coupon	c1	c2	c3	chestionar	City	F8	F9	PS1	PS2	PS7	PS9_3	PS9_5	PSS7
1	20	201	2011	2012	2013	B0100000...	7	1	1	5	16	1	1	49	
2	10	202	2021	2022	2023	B0200000...	7	1	1	5	17	1	NA	50	
3	20	203	2031	2032	2033	B0300000...	7	1	1	5	18	1	NA	NA	
4	11	204	2041	2042	2043	B0400000...	7	0	0	8	16	0	4	1	
5	15	205	2051	2052	2053	B0500000...	7	1	1	1	15	1	2	NA	
6	2	2011	20111	20112	20113	B0110000...	7	1	1	5	15	1	1	NA	
7	8	2012	20121	20122	20123	B0120000...	7	1	1	8	17	1	5	NA	
8	30	2013	20131	20132	20133	B0130000...	7	1	1	7	16	1	6	NA	
9	3	2021	20211	20212	20213	B0210000...	7	0	0	5	18	1	NA	4	
10	10	2032	20321	20322	20323	B0320000...	7	1	0	4	16	NA	5	NA	
11	4	2041	20411	20412	20413	B0410000...	7	0	0	9	14	1	1	NA	
12	15	2043	20431	20432	20433	B0430000...	7	0	0	9	19	0	1	5	
13	3	2051	20511	20512	20513	B0510000...	7	1	1	5	17	1	1	NA	
14	15	2052	20521	20522	20523	B0520000...	7	1	0	4	15	1	NA	NA	
15	6	2053	20531	20532	20533	B0530000...	7	1	0	4	20	1	NA	NA	
16	3	20131	201311	201312	201313	B0131000...	7	0	0	5	12	1	NA	NA	
17	12	20132	201321	201322	201323	B0132000...	7	1	1	3	18	1	NA	3	
18	20	20133	201331	201332	201333	B0133000...	7	1	0	4	15	1	NA	NA	
19	3	20321	203211	203212	203213	B0321000...	7	0	0	4	16	NA	NA	NA	

Figure 1.12 Data View

More than one data set can be loaded into the "Data Viewer" and accessed during an RDS-A session. To identify the data set you want to work with, use the down arrow button to the right of the file name and a drop down box of all the files loaded during the same work session will be displayed. You can select a specific file and it will appear in the "Data Viewer". For instance, in Figure 1.13, there are three files from which to choose during this session.

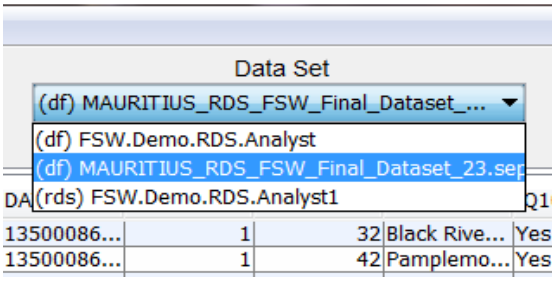
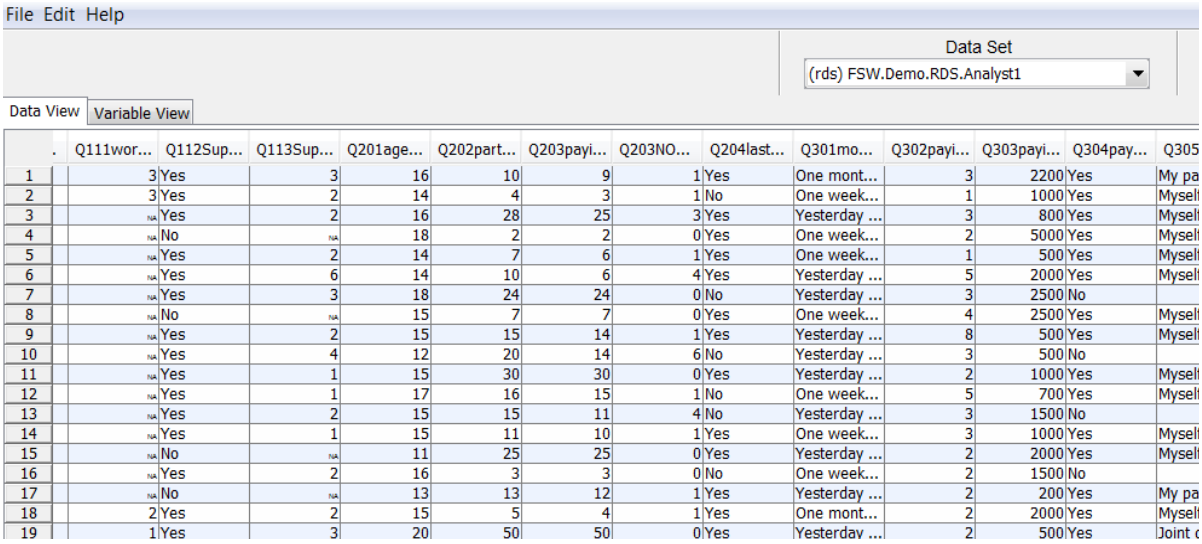


Figure 1.13 Selecting from multiple files

Take a minute to look at the data under the "Data View" tab in Figure 1.12 and Figure 1.14. Some things to note are the following:

- Missing data are automatically transformed into "NA"
- Text can remain in the file (only shown in Figure 1.14)
- The seeds do not need to be identified as missing (RDS-A can recognize the seeds since they were not recruited by anyone).

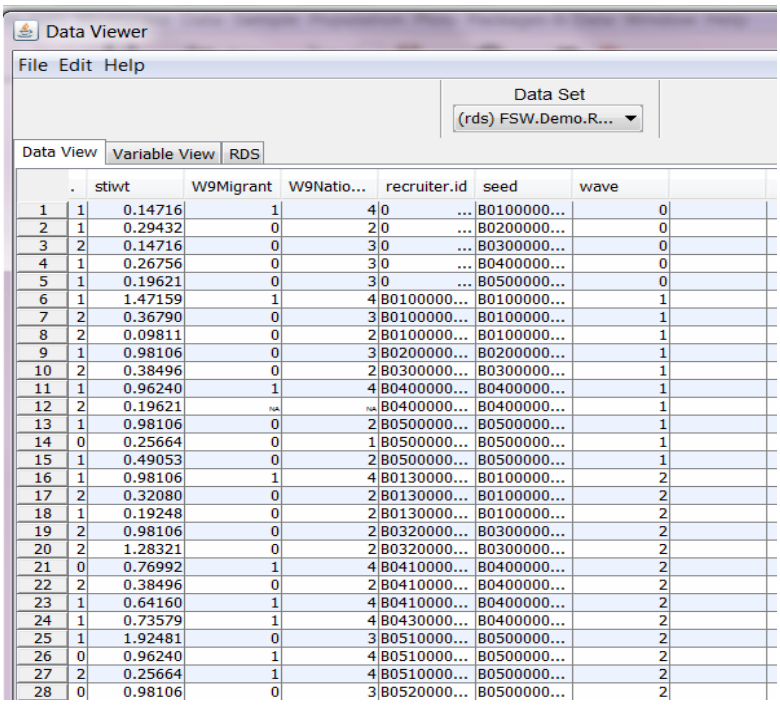


	Q111wor...	Q112Sup...	Q113Sup...	Q201age...	Q202part...	Q203payi...	Q203NO...	Q204last...	Q301mo...	Q302payi...	Q303payi...	Q304pay...	Q305
1	3	Yes	3	16	10	9	1	Yes	One mont...	3	2200	Yes	My pa
2	3	Yes	2	14	4	3	1	No	One week...	1	1000	Yes	Myself
3	NA	Yes	2	16	28	25	3	Yes	Yesterday ...	3	800	Yes	Myself
4	NA	No	NA	18	2	2	0	Yes	One week...	2	5000	Yes	Myself
5	NA	Yes	2	14	7	6	1	Yes	One week...	1	500	Yes	Myself
6	NA	Yes	6	14	10	6	4	Yes	Yesterday ...	5	2000	Yes	Myself
7	NA	Yes	3	18	24	24	0	No	Yesterday ...	3	2500	No	
8	NA	No	NA	15	7	7	0	Yes	One week...	4	2500	Yes	Myself
9	NA	Yes	2	15	15	14	1	Yes	Yesterday ...	8	500	Yes	Myself
10	NA	Yes	4	12	20	14	6	No	Yesterday ...	3	500	No	
11	NA	Yes	1	15	30	30	0	Yes	Yesterday ...	2	1000	Yes	Myself
12	NA	Yes	1	17	16	15	1	No	One week...	5	700	Yes	Myself
13	NA	Yes	2	15	15	11	4	No	Yesterday ...	3	1500	No	
14	NA	Yes	1	15	11	10	1	Yes	One week...	3	1000	Yes	Myself
15	NA	No	NA	11	25	25	0	Yes	Yesterday ...	2	2000	Yes	Myself
16	NA	Yes	2	16	3	3	0	No	One week...	2	1500	No	
17	NA	No	NA	13	13	12	1	Yes	Yesterday ...	2	200	Yes	My pa
18	2	Yes	2	15	5	4	1	Yes	One mont...	2	2000	Yes	Myself
19	1	Yes	3	20	50	50	0	Yes	Yesterday ...	2	500	Yes	Joint c

Figure 1.14 Data view of another file during the same work session

Identification of seeds and waves in Data Viewer

Under the "Data View" tab, at the far right, are three additional columns: The recruiter ID, seed and wave. The "recruiter.ID" column identifies the seeds and the "wave" column identifies the number of waves in the sample. In Figure 1.15, the last column shows five seeds (wave 0), 10 participants who were in wave one, 13 participants who were in wave two and so on.



	.	stiwt	W9Migrant	W9Natio...	recruiter.id	seed	wave
1	1	0.14716	1	4	0 ...	B0100000...	0
2	1	0.29432	0	2	0 ...	B0200000...	0
3	2	0.14716	0	3	0 ...	B0300000...	0
4	1	0.26756	0	3	0 ...	B0400000...	0
5	1	0.19621	0	3	0 ...	B0500000...	0
6	1	1.47159	1	4	B0100000...	B0100000...	1
7	2	0.36790	0	3	B0100000...	B0100000...	1
8	2	0.09811	0	2	B0100000...	B0100000...	1
9	1	0.98106	0	3	B0200000...	B0200000...	1
10	2	0.38496	0	2	B0300000...	B0300000...	1
11	1	0.96240	1	4	B0400000...	B0400000...	1
12	2	0.19621	NA	NA	B0400000...	B0400000...	1
13	1	0.98106	0	2	B0500000...	B0500000...	1
14	0	0.25664	0	1	B0500000...	B0500000...	1
15	1	0.49053	0	2	B0500000...	B0500000...	1
16	1	0.98106	1	4	B0130000...	B0100000...	2
17	2	0.32080	0	2	B0130000...	B0100000...	2
18	1	0.19248	0	2	B0130000...	B0100000...	2
19	2	0.98106	0	2	B0320000...	B0300000...	2
20	2	1.28321	0	2	B0320000...	B0300000...	2
21	0	0.76992	1	4	B0410000...	B0400000...	2
22	2	0.38496	0	2	B0410000...	B0400000...	2
23	1	0.64160	1	4	B0410000...	B0400000...	2
24	1	0.73579	1	4	B0430000...	B0400000...	2
25	1	1.92481	0	3	B0510000...	B0500000...	2
26	0	0.96240	1	4	B0510000...	B0500000...	2
27	2	0.25664	1	4	B0510000...	B0500000...	2
28	0	0.98106	0	3	B0520000...	B0500000...	2

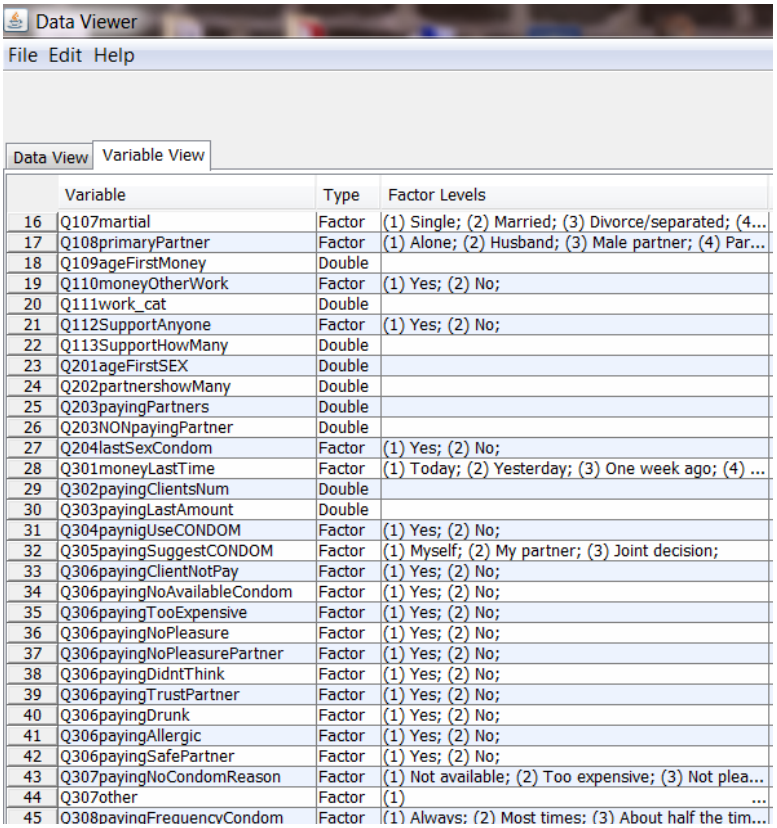
Figure 1.15 Identification of seeds and waves in Data Viewer

Variable View

Under the "Variable View" tab (Figure 1.16), normally three columns of data are displayed in addition to the far left column of serial numbers. The first column is "Variable", which displays the variable name.

The second column is "Type". There are eight possible types of data under the "Type" column (see Figure 1.17): "Character" or string, "Factor" (the term "factor" in the underlying R engine designates a categorical variable; factors are nominal but they are used to represent both ordinal and nominal variables), "Double" (non-integers), "Integer", "Logical" (true, false), "Date", "Time", and "Other". Sometimes if the program is not reading your data properly, you will need to change the type of data assigned to your variables.

The third column is "Factor Levels". In the data set in Figure 1.16, all of the factor (or category) labels from the original SPSS Statistics file were retained in RDS-A.



	Variable	Type	Factor Levels
16	Q107martial	Factor	(1) Single; (2) Married; (3) Divorce/separated; (4...
17	Q108primaryPartner	Factor	(1) Alone; (2) Husband; (3) Male partner; (4) Par...
18	Q109ageFirstMoney	Double	
19	Q110moneyOtherWork	Factor	(1) Yes; (2) No;
20	Q111work_cat	Double	
21	Q112SupportAnyone	Factor	(1) Yes; (2) No;
22	Q113SupportHowMany	Double	
23	Q201ageFirstSEX	Double	
24	Q202partnershowMany	Double	
25	Q203payingPartners	Double	
26	Q203NONpayingPartner	Double	
27	Q204lastSexCondom	Factor	(1) Yes; (2) No;
28	Q301moneyLastTime	Factor	(1) Today; (2) Yesterday; (3) One week ago; (4) ...
29	Q302payingClientsNum	Double	
30	Q303payingLastAmount	Double	
31	Q304paynigUseCONDOM	Factor	(1) Yes; (2) No;
32	Q305payingSuggestCONDOM	Factor	(1) Myself; (2) My partner; (3) Joint decision;
33	Q306payingClientNotPay	Factor	(1) Yes; (2) No;
34	Q306payingNoAvailableCondom	Factor	(1) Yes; (2) No;
35	Q306payingTooExpensive	Factor	(1) Yes; (2) No;
36	Q306payingNoPleasure	Factor	(1) Yes; (2) No;
37	Q306payingNoPleasurePartner	Factor	(1) Yes; (2) No;
38	Q306payingDidntThink	Factor	(1) Yes; (2) No;
39	Q306payingTrustPartner	Factor	(1) Yes; (2) No;
40	Q306payingDrunk	Factor	(1) Yes; (2) No;
41	Q306payingAllergic	Factor	(1) Yes; (2) No;
42	Q306payingSafePartner	Factor	(1) Yes; (2) No;
43	Q307payingNoCondomReason	Factor	(1) Not available; (2) Too expensive; (3) Not plea...
44	Q307other	Factor	(1) ...
45	Q308payingFrequencyCondom	Factor	(1) Always; (2) Most times; (3) About half the tim...

Figure 1.16 Variable View

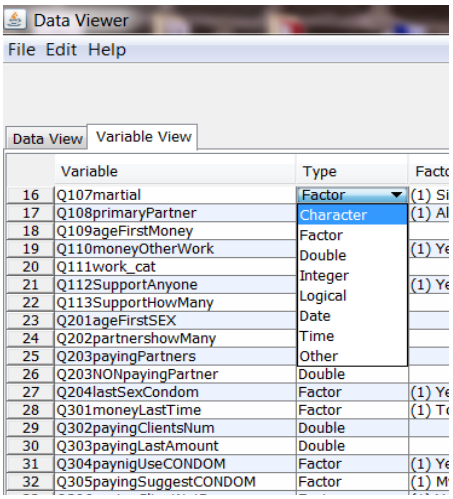


Figure 1.17 Data types

Save data

To avoid having to re-enter the data parameters again, you can save your file as a RDSOBJ file which is easily recognizable by RDS-A as an RDS file. To do this, go to "Console" and select File|Save Data. Once you do this the following window will open and you will select "OK" to save the file (Figure 1.18). Once you click "OK", you will need to tell it where to save the file and give it a name. Once you select a file name (the name selected below is the same name used for the original SPSS file), click "Save".

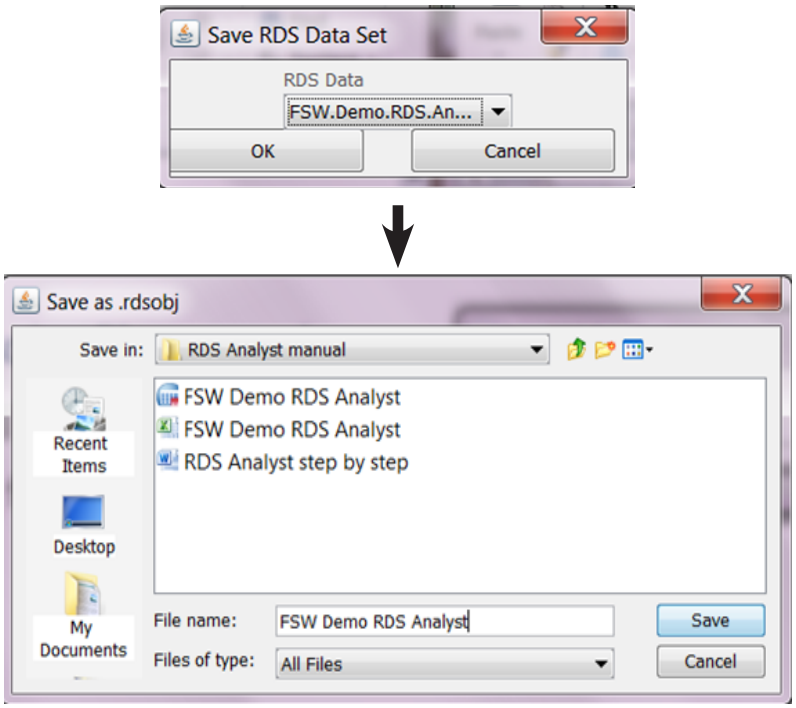


Figure 1.18 Saving RDS dataset from RDS-A

In Figure 1.19, you can see the saved RDSOBJ file, highlighted. This action will also automatically convert your file to a DL file and a vna file to be used in recruitment graphics (see the step-by-step guide to NetDraw in Section 2 of this Supplement).

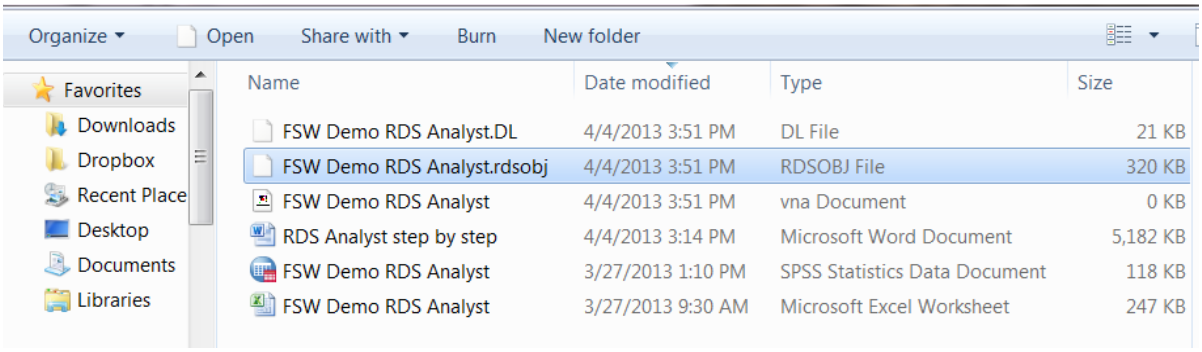


Figure 1.19 Saved RDS-A datasets from RDS-A

Working with data

Recode variables

To recode a variable in the data set, go to Data|Recode Variables. Select a variable from the left-hand side panel of the "Recode Variables" window and move it into the panel labeled "Variables to Recode" to the right by using the blue arrow keys. If you want to recode your variable into another variable name, you can change the new variable name by clicking "Target" and entering the new name. Select "Define Code". In the example in Figure 1.20, the variable selected is "lab_HIV", which are results from the HIV test (see window on left, "Recode Variables").

After selecting "Define Code", the "Set Variable Codings" window will open (Figure 1.20, window on right). "Lab_HIV" is currently coded as 1 for yes and 0 for no. For example, if you want to change 0 to 2 so that 2 is the code for no, type the code you want to change (0) into the "Value" panel, under "Code". The options for values are "=", "≥" and "≤" (the later two options are for ranges) or "Else". Under the 0, put 2 into the panel to the right of "into" and then select "Add". The new coding will appear in the bottom panel, "Recodings". Click "OK".

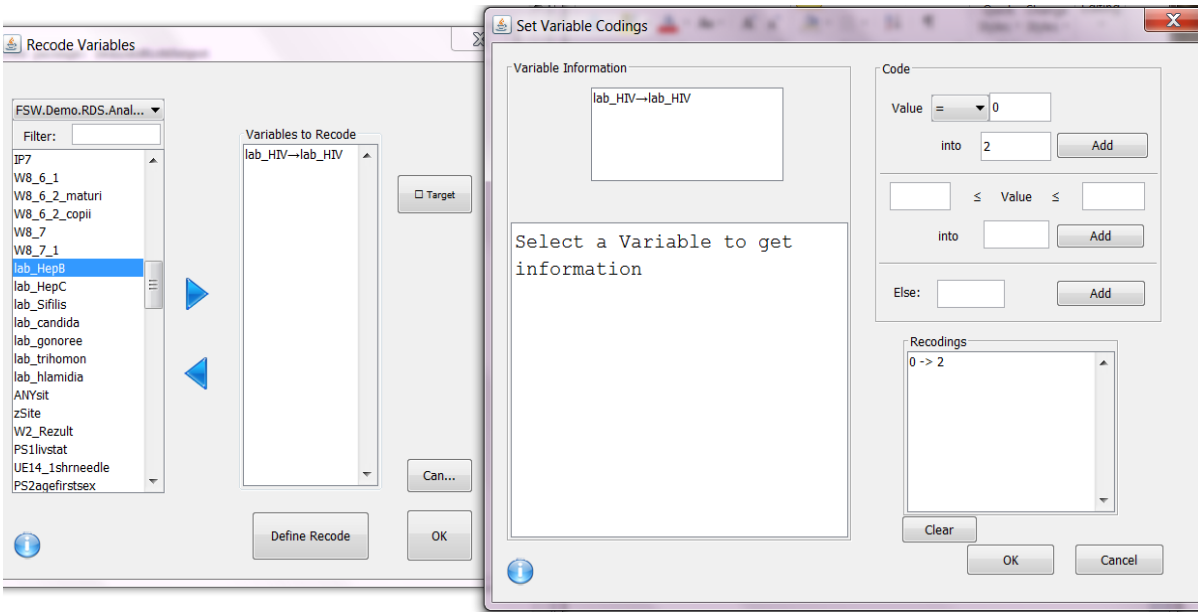


Figure 1.20 Recoding variables and setting variable codes

Compute weights[1]

To compute weights used in analysis of RDS data select Data|Compute Weights|Run. The "Compute Weights" window will open (Figure 1.21). Here you can compute weights for a specific variable using a specific "Type" (estimator method) by typing the variable name into the "Group Variable" panel. The Gile's Successive Sampling (Gile's SS) method and RDS-II do not need a group variable, however for RDS-I a group variable is needed.

[1] Weights are often computed and exported for running regression analyses in other software packages. However, regression analyses for network data are complicated and the use of exported weights will not ensure that your regression output is representative of anything and should be noted as a limitation of your analysis. Furthermore, exporting weights to another program to run descriptive analyses will not eliminate biases in the sample.

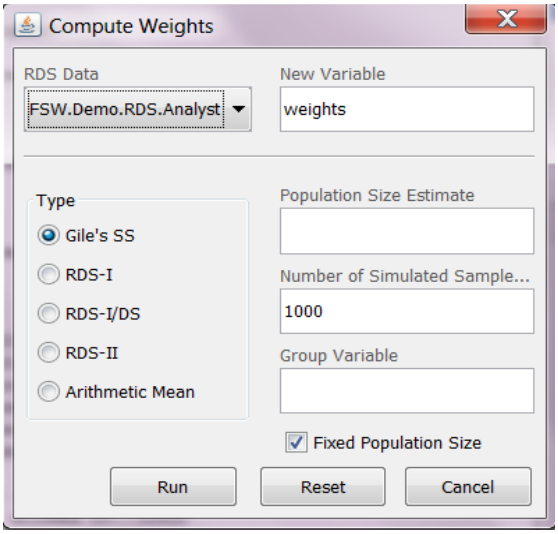


Figure 1.21 Compute Weights

Analysis of the sample

To analyse the sample (unweighted analysis), use the "Sample" tab in the "Console" window.

Sample frequencies

To conduct frequencies of the sample click Sample|Frequencies. A window named "Run Frequencies" will open. From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it/them in the right-side panel using the blue arrows. In the example in Figure 1.22, "lab_HIV", "lab_HepB" and "lab_HepC" are selected. Click "OK".

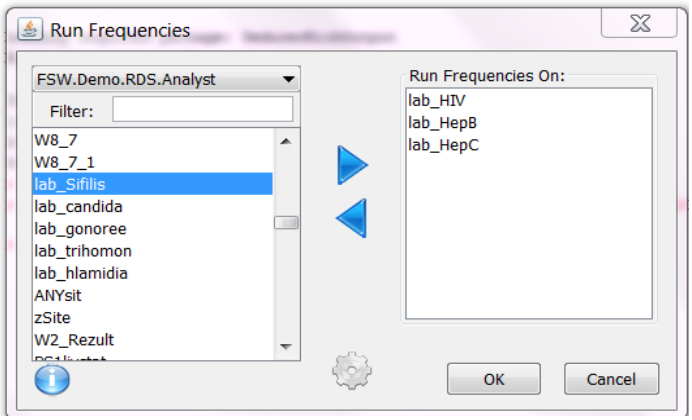


Figure 1.22 Selecting variables: Sample frequency

Note: Most windows have an information icon which you can click to learn more about the feature you are using. Upon clicking this icon, you will be taken to the manual at http://www.deducer.org/. You will need to be connected to Internet to access this. This online manual is being updated frequently so any updates to the software can be found here.

The output appears in the "Console View" panel. In Figure 1.23, The "Frequencies" for lab_HIV show the values 0 (HIV negatives) and 1 (HIV positives). The "Console View" displays the number of cases (0 = 551, 1 = 107), the percentage (0 = 83.7%, 1 = 16.3%) and the cumulative percentage. If your data have labels, the labels will show up in the far left-hand column. Below the "Frequencies" table is a "Case Summary" table with information about the number of valid responses, missing data, total and percentage missing.

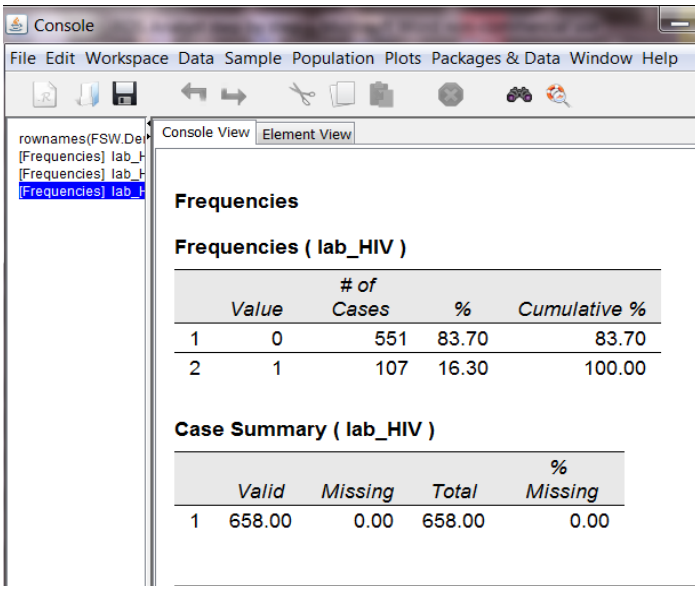


Figure 1.23 Sample frequency output of HIV

Descriptives

Click Sample|Descriptives from the "Console" window. The "Descriptives" window will open (Figure 1.24). From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting it/them and putting it/them in the right-side panel using the blue arrows. To stratify variables, put them into the "Stratify By" panel. The variable selected for "Descriptives of" is PS1 (number of sexual partners in the past three months [continuous variable]). Click "Run".

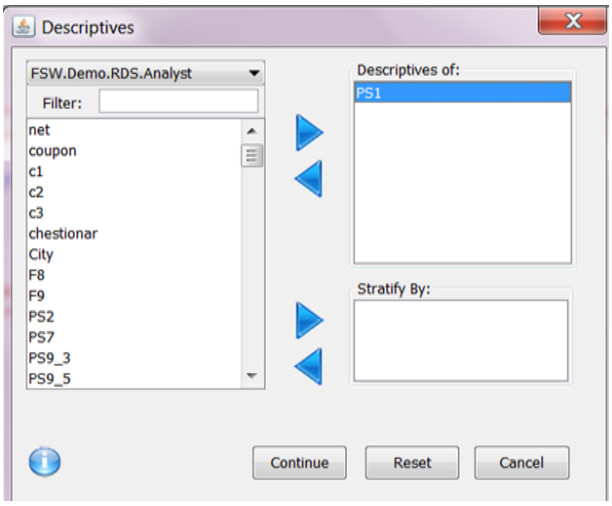


Figure 1.24 Selecting a variable: Descriptive analysis

Once you select the variable(s) you want to analyse, another window, also named "Descriptives", will open (Figure 1.25). This window will allow you to select from the "Functions" panel the descriptive analyses you want by moving the listed functions to the "Run Descriptives" panel. The functions selected for this analysis are "Mean", "St. Deviation", "Valid N", "Maximum", "Minimum" and "Median". Click "Run".

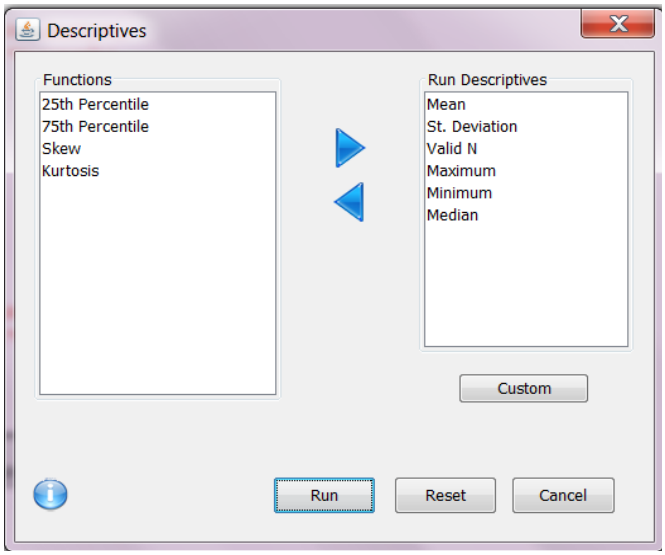


Figure 1.25 Selecting functions: Descriptive analysis

Figure 1.26 shows the output of the descriptive analysis of sexual partners in the past three months. In this output, there is an average (mean) of 5.48 sexual partners in the past three months with a standard deviation of 2. The valid N is 658, the maximum is nine partners and the minimum is one partner. The median is five partners.

Descriptive Statistics

	Mean	*St. Deviation*	*Valid N*	*Maximum*	*Minimum*	*Median*
PS1	5.48	2.00	658	9.00	1.00	5.00

Figure 1.26 Output: Descriptive analysis

Contingency tables (crosstabs)

Click Sample|Contingency Tables from the "Console" window. A window named "Contingency Tables" will open (Figure 1.27). From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it in the right-side panel using the blue arrows. Put at least one variable in the "Row" panel and one variable in the "Column" panel. In Figure 1.27, "lab_HIV" is selected for the row and "lab_HepC" is selected for the column. Click "Run".

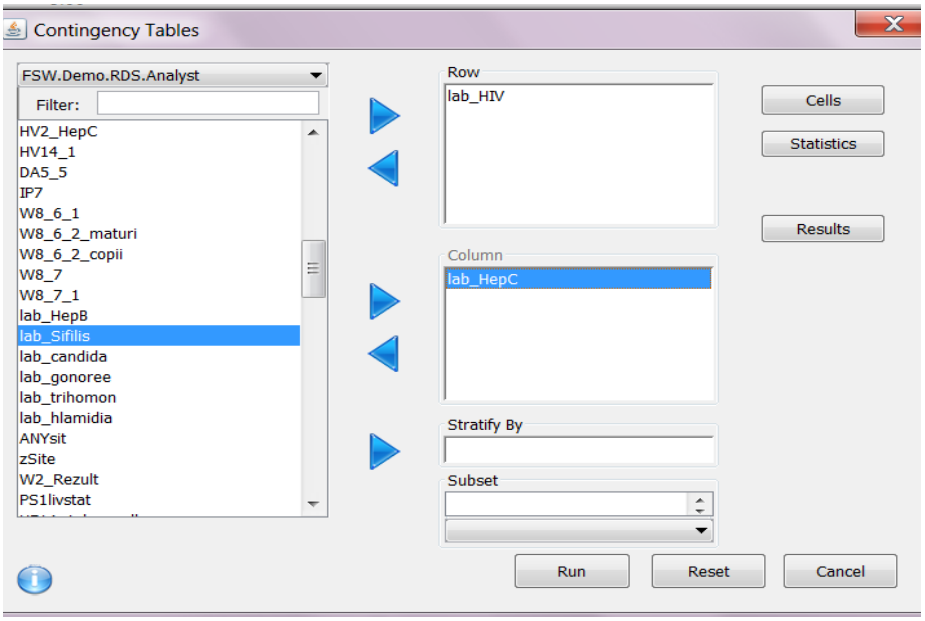


Figure 1.27 Contingency Tables: Selecting variables

On the right-hand site of the "Contingency Tables" window is a tab named "Cells". By selecting this tab, a window named "Table Cell Contents" will open (Figure 1.28). In addition to observation counts, this window allows you to select from the following:

- "Percentages"
  - "Row" (within each row percentage)
  - "Column" (within each column percentage)
  - "Total" (percentage in cell).
- "Chi-squared"
  - "Expected" (the expected count of the cell if there were no relationship between the two variables)
  - "Residuals" (the observed count minus the expected count)
  - "Standardized Residuals" (the residuals standardized such that, if the two variables were independent, they have mean 0 and standard deviation 1; these residuals are useful in determining which cells of a contingency table contribute most to a significant chi-squared test)
  - "Adjusted Residuals" (these adjust the residuals by the row and column totals).

Click "OK".

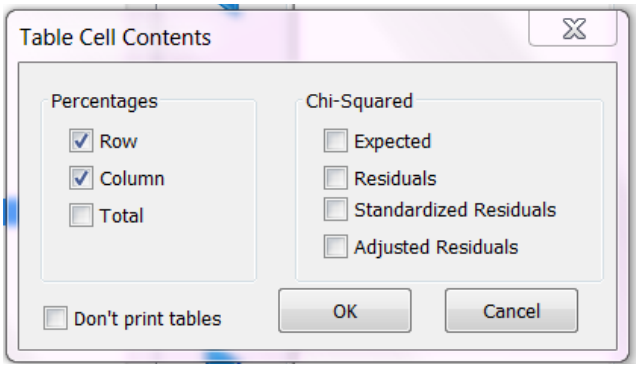


Figure 1.28 Contingency Tables: Selecting table cell contents

On the right-hand site of the Contingency Tables window is a tab named "Statistics". By selecting this tab, a window named "Table Statistics" will open (Figure 1.29). For this analysis, the "Chi-Squared" statistic is selected. Click "OK".

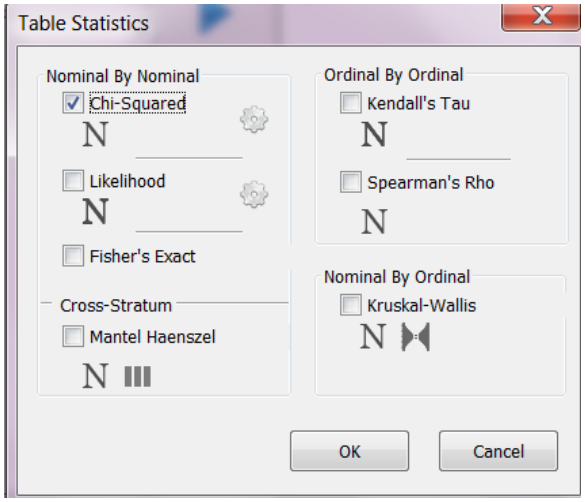


Figure 1.29 Contingency Tables: Selecting table statistics

Figure 1.30 shows output from the contingency analysis. For instance, in the Console View in Figure 1.30, 58.9% of those infected with HIV, were also infected with hepatitis C.

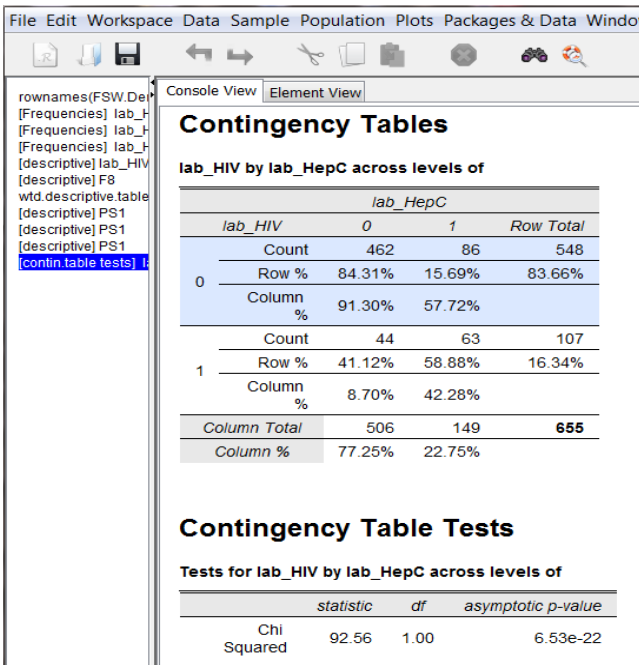


Figure 1.30 Contingency Tables: Output

Recruitment homophily

Click Sample|Recruitment Homophily from the "Console" window. A window named "Homophily" will open (Figure 1.31). From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it in the right-side panel using the blue arrows. Click "Run".

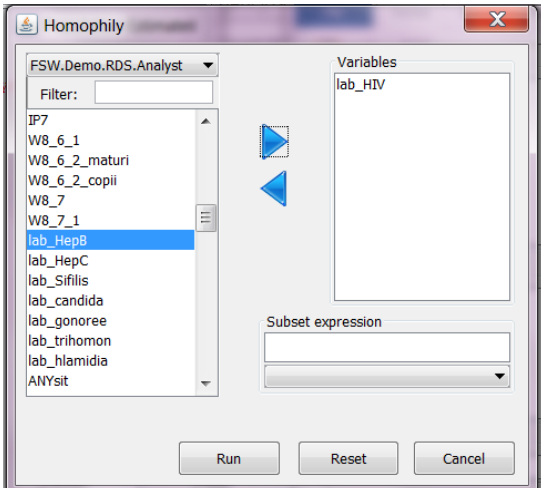


Figure 1.31 Recruitment homophily: Selecting a variable

Figure 1.32 shows an example of recruitment homophily using HIV serostatus. In this case, it is the ratio of number of recruits that have the same HIV serostatus as their recruiter to the number we would expect if there was no homophily on HIV serostatus. This homophily is calculated from the recruitment chain (rather than the population of social ties). For example, if the recruitment homophily on HIV status is about 1 (specifically 1.069922), there is little effect of recruitment homophily on HIV serostatus (as the numbers of homophilous pairs are close to what we would expect by chance).

```
Recruitment Homophily for lab_HIV

Homophily = 1.069922

                      lab_HIV of recruit
lab_HIV of respondent   1   0
                    1  34  79
                    0  67 465

Number of cases in table: 645
Number of factors: 2
Test for independence of all factors:
        Chisq = 21.599, df = 1, p-value = 3.36e-06
```

Figure 1.32 Recruitment homophily: Output

Analysis of the population

Analysing the population (weighted analysis) is done using the "Population" tab on the Console window. RDS-A offers several estimators, all of which have variations on the methods used to derive estimates and confidence intervals.

RDS estimators[1]

Table 1.2 provides information about the estimators available in RDS-A. For more information on these estimators, review the source provided in the first column and listed in the reference section at the end of this guide.

[1] Not including Heckathorn, 1997.

Table 1.2 Estimators available in RDS-A[1]

RDS estimator and type	Data needed	Limitations	Variance estimation	Analysis feature	Software available
Heckathorn, 2002 Reciprocity model-based estimator (RDS I); Linear least squares used for >2 categories and data smoothing used to estimate >2 categories	Recruitment matrix; self-reported network sizes	Limited to categorical data and by RDS assumptions	Bootstrap	Controls for differences in network sizes, homophily across groups; uses data smoothing for narrower confidence intervals	Yes: STATA RDS estimator (Schonlau and Liebau, 2010)
Salganik and Heckathorn, 2004 Reciprocity model-based estimator (RDS I)	Recruitment matrix; self-reported network sizes	Limited to categorical variables and by RDS assumptions	Bootstrap	Proof that estimate is asymptotical-ly unbiased; estimate of average group network size	Yes: RDSAT, RDS Analyst
Heckathorn, 2007 Dual-component estimator (RDS I)	Recruitment matrix; self-reported network sizes	As other estimators, limited by RDS assumptions	Bootstrap	Allows analysis of continuous variables; controls for differential recruitment	Yes: RDSAT, RDS Analyst
Volz and Heckathorn, 2008 Probability-based estimator; data-smoothing (RDS II)	Recruitment matrix; self-reported network sizes	Limited to nominal variables	Analytic	Allows analysis of continuous variables; shows convergence between reciprocity- and probability-based RDS estimators; uses data smoothing to control for differential recruitment	Yes: RDS Analyst
Gile's successive sampling, 2011 Probability-based estimator	Working estimate of population size, recruitment matrix; self-reported network sizes	Currently limited to categorical variables (may be extended)	Bootstrap	Corrects for finite population effects	Yes: RDS Analyst
Gile and Handcock's model-assisted estimator, 2012 Probability-based estimator	Working estimate of population size, recruitment matrix; self-reported network sizes, self-reported composition of contacts helpful but not necessary	Currently limited to binary variables (may be extended)	Bootstrap	Corrects for finite population effects and for some forms of seed bias	To be made available in RDS Analyst

[1] Parts of this table were adapted from Wenjert C, Heckathorn DD. Respondent-driven sampling: Operational procedures, evolution of estimators, and topics for future research. In: Williams M, Vogt PW, eds. The SAGE handbook of innovation in social research methods. London, Sage Publications, 2011.

Frequency estimates

To analyse frequency estimates click Population|Frequency Estimates. A window named "Population Frequency Estimates" will open. From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it into the right-side panel using the blue arrows. In the example in Figure 1.33, "lab_HIV", "lab_HepB" and "lab_HepC" have been selected. Also note the following items which are set as defaults in the program.

- "Weights": Gile's SS.
- "Confidence": 95%. The confidence interval is computed using Gile's bootstrap method. This is a computationally-intensive procedure and can take a minute or longer to complete.
- "Population Size"[1]: 1500 (this was added to the data viewer when loading the data set).
- Number of simulations/iterations: 1000.

Click "Run".

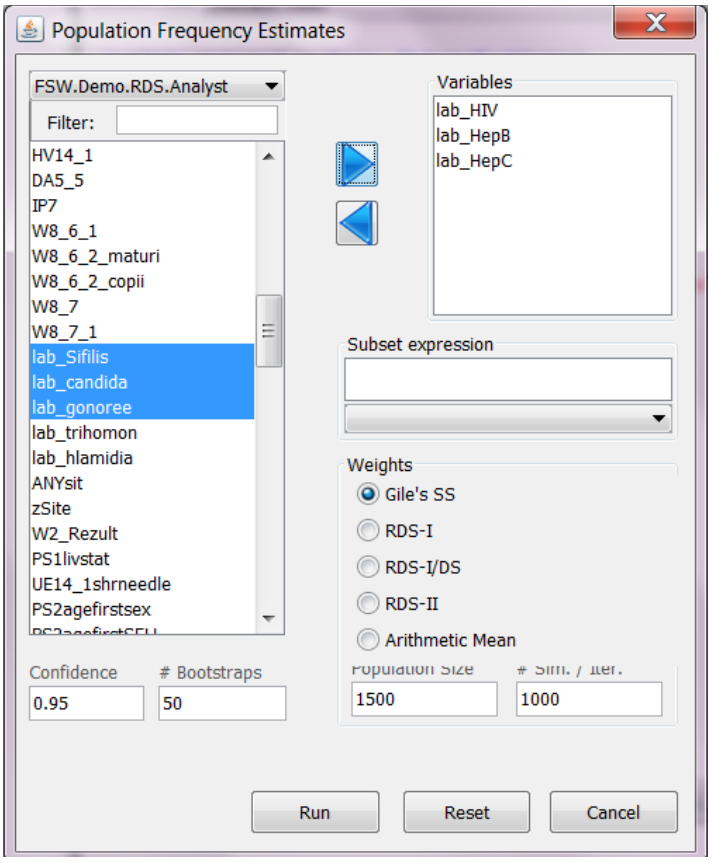


Figure 1.33 Population Frequency Estimates: Selecting variables

The output appears in the output panel of the "Console" window (Figure 1.34). In Figure 1.34, the frequencies for "lab_HIV" show the values 0, which are HIV negatives and 1 which are HIV positives. The window displays the estimates (0 = 83.8%, 1 = 16.2%), confidence bounds (80.8 to 86.9 for group 0 and 13.1 to 19.3 for group 1), the estimated design effect (2.15), standard error (0.016) and number of cases for each category (0 = 551, 1 = 107). If your data have labels, the labels will show up in the left-hand column.

[1] Needed for the Gile's successive sampling and Gile and Handcock's model-assisted estimator.

Gile's SS Estimate for lab_HIV

	Point Estimate	95% Lower Bound	95% Upper Bound	Estimated Design Effect	Standard Error	Sample Size
0	0.8384	0.8075	0.8693	2.15	0.0158	551
1	0.1616	0.1307	0.1925	2.15	0.0158	107
					Total	658

Figure 1.34 Frequency estimate of HIV: Output

Parsing data

In the "Population Frequency Estimates" window, it is also possible to parse data using the "Subset expression" panel (Figure 1.35). In the window in Figure 1.35, "lab_HepC" is added to the variables panel and "lab_HIV" (along with "==1", the code for HIV positive) is written into the "Subset expression" panel. Click "Run".

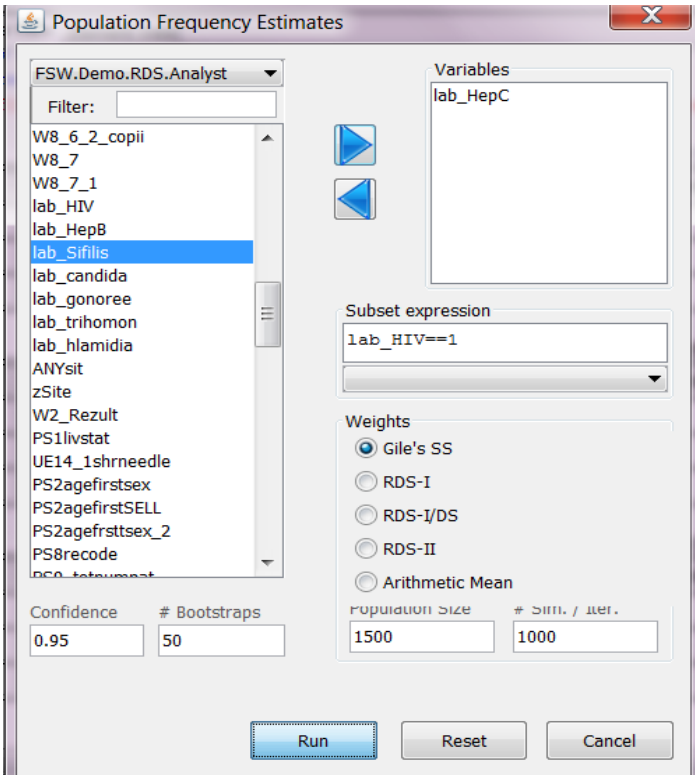


Figure 1.35 Parsing data

In the example in Figure 1.36, the sample sizes for each category are equal to 107 (the sample size for all those who were HIV positive in Figure 1.30). Figure 1.36 shows that among those who were HIV positive, 47.7% (95% CI: 37.1, 58.4) were also infected with hepatitis C and 52.3% (95% CI: 41.6, 62.9) were not infected with hepatitis C. The figure also displays design effects and standard errors.

Gile's SS Estimate for lab_HepC

	Point Estimate	95% Lower Bound	95% Upper Bound	Estimated Design Effect	Standard Error	Sample Size
0	0.5225	0.4157	0.6292	1.37	0.0545	44
1	0.4775	0.3708	0.5843	1.37	0.0545	63
					Total	107

Figure 1.36 Output: Parsing data

Descriptive estimates

To produce descriptive estimates, click Population|Descriptive Estimates. A window named "Population Estimates" will open. From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it into the right-side panel using the blue arrows. In Figure 1.37, PS2 (age at first sexual intercourse) is selected. From the "Statistics" panel, "Minimum", "Maximum" and "Median" are selected. Click "Run".

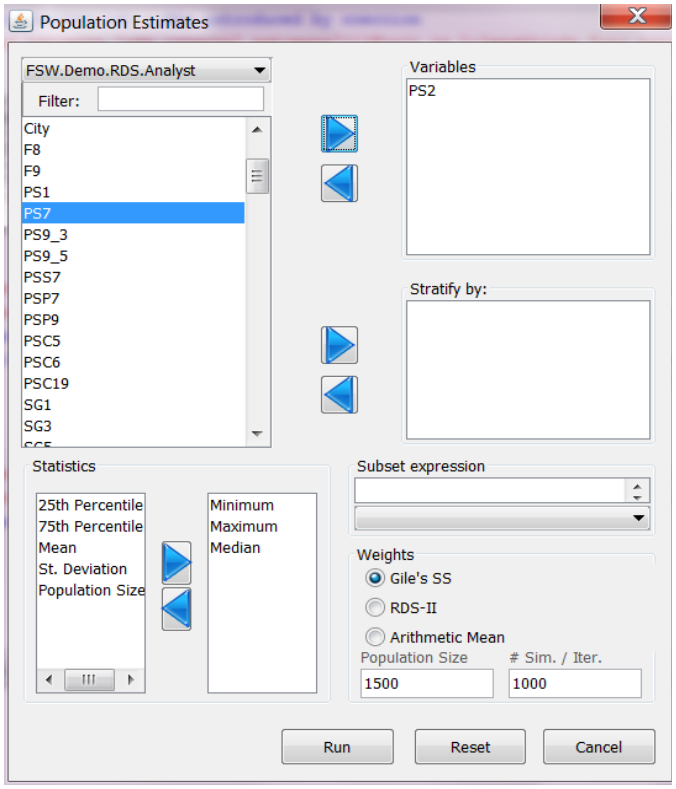


Figure 1.37 Descriptive estimates

In the output in Figure 1.38, the minimum age at first sexual intercourse was 6 years old, the maximum was 24 years old and the median was 16 years old.

```
> wtd.descriptive.table(vars=d(PS2), data=FSW.Demo.RDS.Analyst, func.names=c("Minimum","Maximum","Median"),
+ weights=compute_weights(FSW.Demo.RDS.Analyst, weight.type="Gile's SS", N=1500, number.ss.samples.per.iteratio
$`strata: all cases `
   Minimum.PS2    Maximum.PS2 Median.PS2.50%
             6             24             16
```

Figure 1.38 Descriptive estimates: Output

Population crosstabs

To produce population crosstabs, click Population|Population Crosstabs. A window named "Population Crosstabs" will open. From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting it into the right-side panel using the blue arrows. In the example in Figure 1.39, "lab_HIV" is selected for the "Rows" and "EverHIVtest" (persons who reported ever having had an HIV test) is selected for the "Columns". Click "Run".

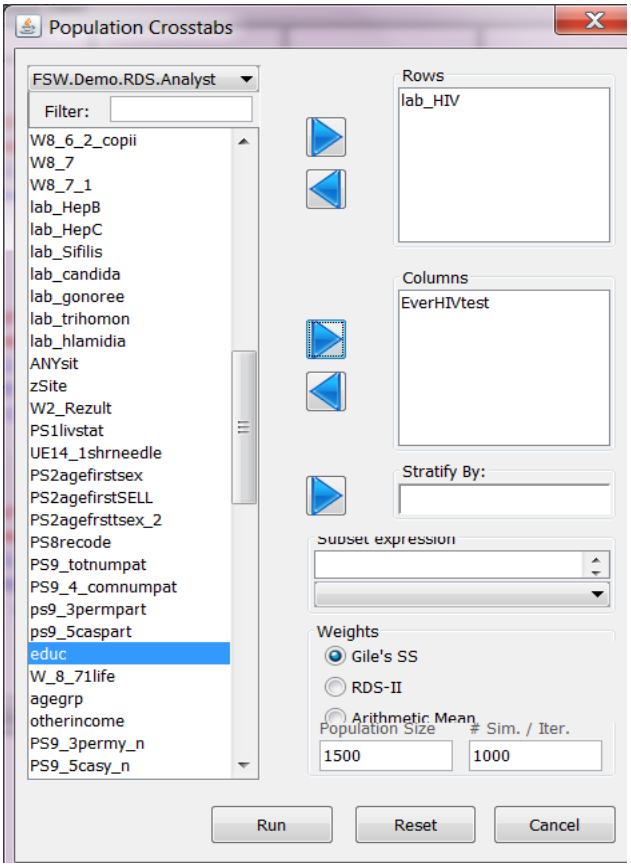
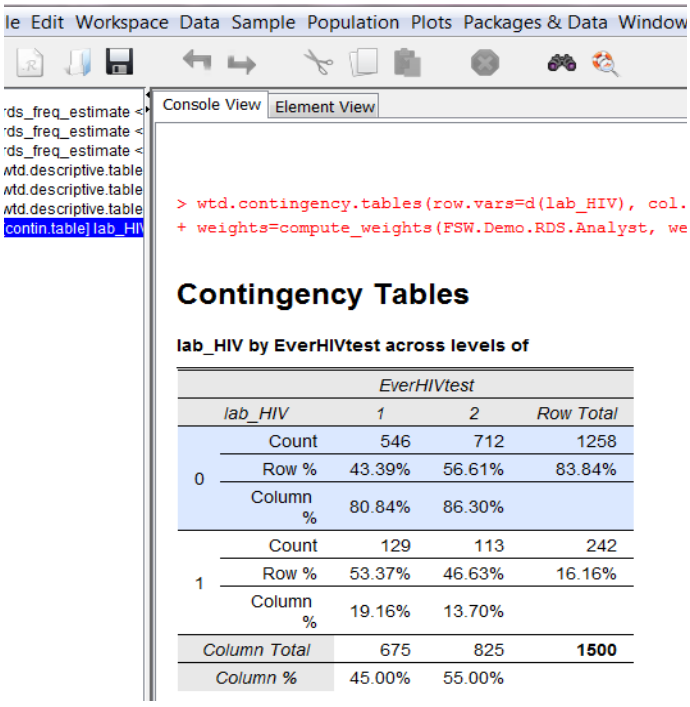


Figure 1.39 Population Crosstabs

The "Console View" (Figure 1.40) displays "EverHIVtest" as columns: 1 is yes, 2 is no. "Lab_HIV" (HIV serostatus) is presented as the rows: 1 is yes, 0 is no. This figure shows that, among those who were HIV positive, 53.37% ever had an HIV test.



		EverHIVtest		
lab_HIV		1	2	Row Total
0	Count	546	712	1258
	Row %	43.39%	56.61%	83.84%
	Column %	80.84%	86.30%	
1	Count	129	113	242
	Row %	53.37%	46.63%	16.16%
	Column %	19.16%	13.70%	
Column Total		675	825	**1500**
Column %		45.00%	55.00%	

Figure 1.40 Population Crosstabs: Output

Population homophily

To calculate population homophily, click Population|Population Homophily. A window named "Homophily" will open (Figure 1.41). From the left-side panel, select the variable(s) for analysis (you can select more than one variable at a time) by highlighting the variable and putting in the right-side panel using the blue arrows. In this example, "lab_HIV" is selected. Click "Run".

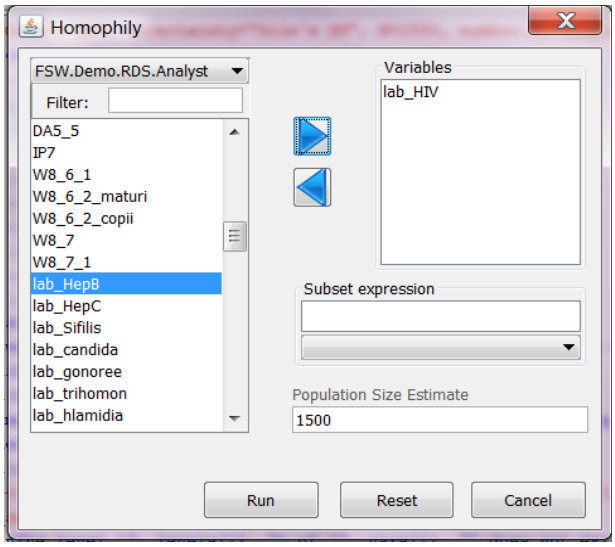


Figure 1.41 Population Homophily

The output is displayed in the "Console View" (Figure 1.42) as 1.39. This is the ratio of the number of recruits that have the same HIV serostatus as their recruiter to the number we would expect if there were no homophily on HIV serostatus and is calculated based on the population of social ties (rather than from the recruitment chain); also see recruitment homophily section above. For example, the homophily of 1.39 on HIV status demonstrates a moderate effect of homophily on HIV status (The numbers of homophilous pairs are more than we would expect by chance).

```
> print(homophily.estimates(rds.data=FSW.Demo.RDS.Analyst, outcome.va
Population Homophily Estimate for lab_HIV
1.389375
```

Figure 1.42 Population Homophily: Output

Plots to visually explore RDS data

Plot recruitment tree

To create a recruitment tree, go to Plots|Plot Recruitment Tree. A window named "Plot Recruitment Tree" will open. Use the blue arrow keys to select a variable from the variables panel and add it to "Node Color" or "Node Size" of you want to highlight the categories in the variable. Otherwise, there is no need to select a variable and you can just click "Run". In Figure 1.43, "lab_Sifilis" is selected and added to the "Node Color" panel. Output can be either in a "Graphics Window" or "PDF Report".

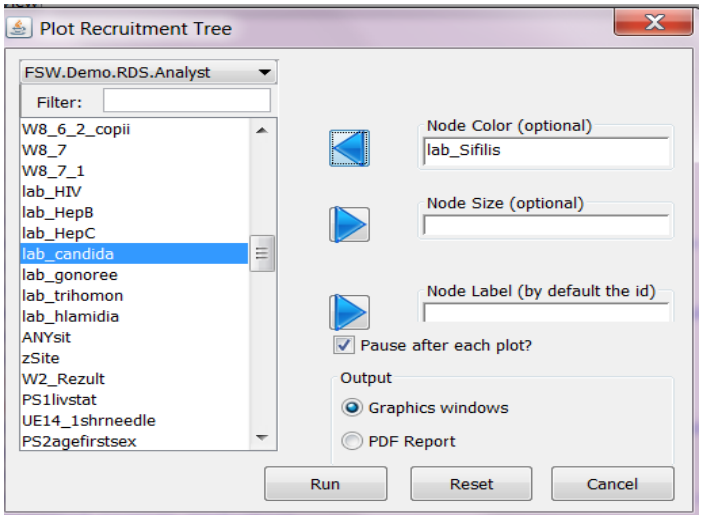


Figure 1.43 Plot Recruitment Tree window

Figure 1.44, displays a graphic of the recruitment trees. Red are those who had reactive tests for syphilis and black are for those who had non-reactive tests. Blue, which is not visible in this example, are for those who had indeterminate results. In this portion there are three recruitment trees. For more aesthetically pleasing recruitment trees there are other software programs such as NetDraw and Gephi.

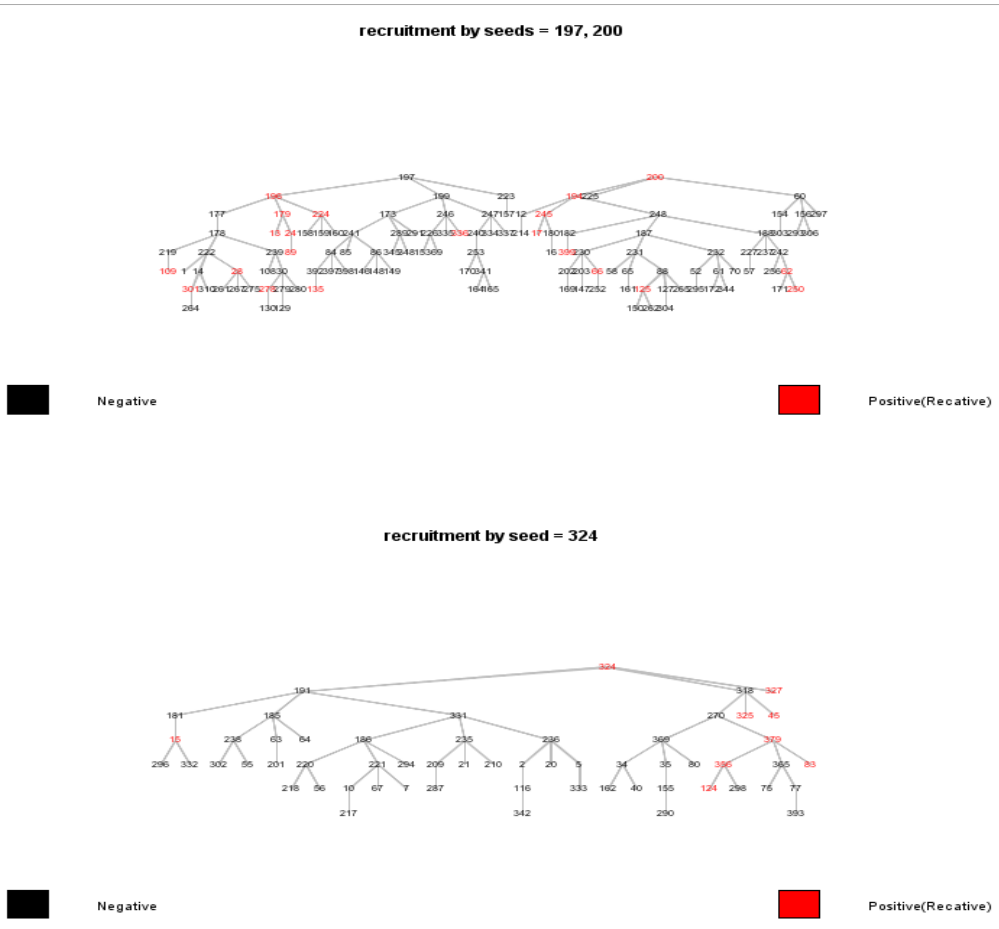


Figure 1.44 Plot recruitment tree with colour nodes: Output

Recruitment diagnostics

For other types of diagnostic graphics, go to Plots|Recruitment Diagnostics. This will open up a window called "Diagnostics Plots" (Figure 1.45). Under "Plots" (on the right-side of the window) you can click on the different types of plots you want to produce, including the "Recruitment tree". You can choose to stratify by a variable or not. No variable is selected in Figure 1.45. Each of the plot types are described in more detail below.

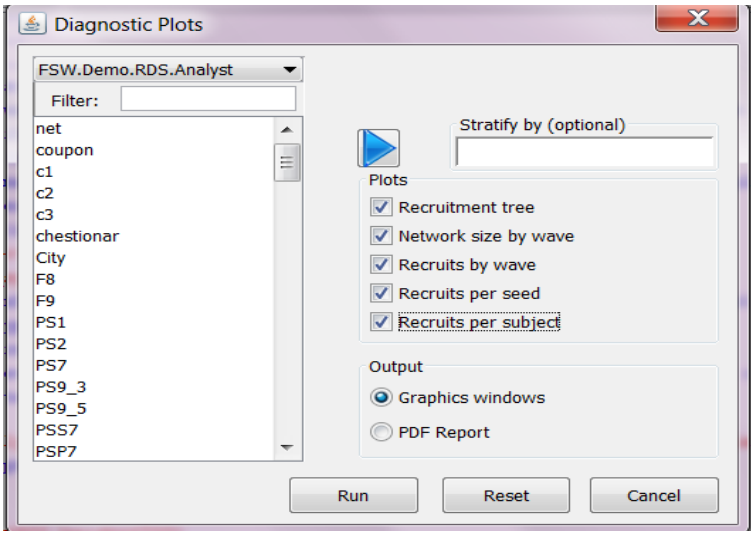


Figure 1.45 Window for diagnostic plots

Network size by wave

Figure 1.46 shows a scatter plot of network size by wave. The network sizes for this sample range from 1 to just under 100. The number of waves is six. The legend to the right of the graph shows the number of subjects with identical network sizes. The largest black filled circle depicts 60 subjects. Most of the network sizes fall above zero (there should never be a network size of zero) to 25. The red lines through the network sizes are the mean of all network sizes for each wave. In the upper part of the graph are outliers having network sizes of between 75 and 100. Overall, the mean network sizes across the waves are similar.

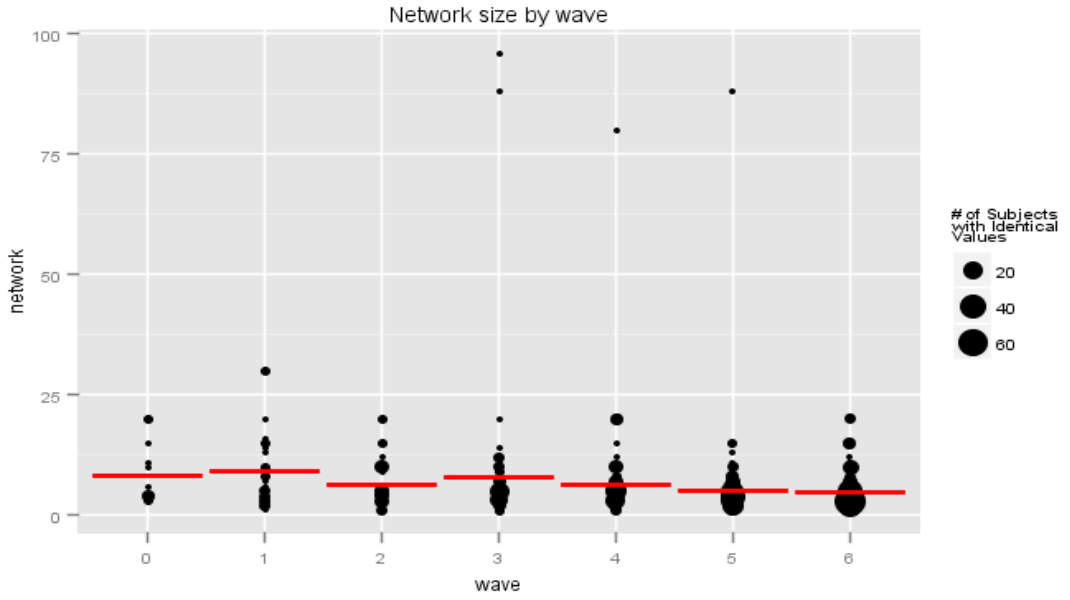


Figure 1.46 Network size by wave

Recruits by wave

Figure 1.47 shows a bar chart of recruitment by wave. The recruitment count ranges from zero to 200. The number of waves is six. There is a steady increase of recruitment by wave, starting with 0 (seeds) to wave 6, shown in the chart.

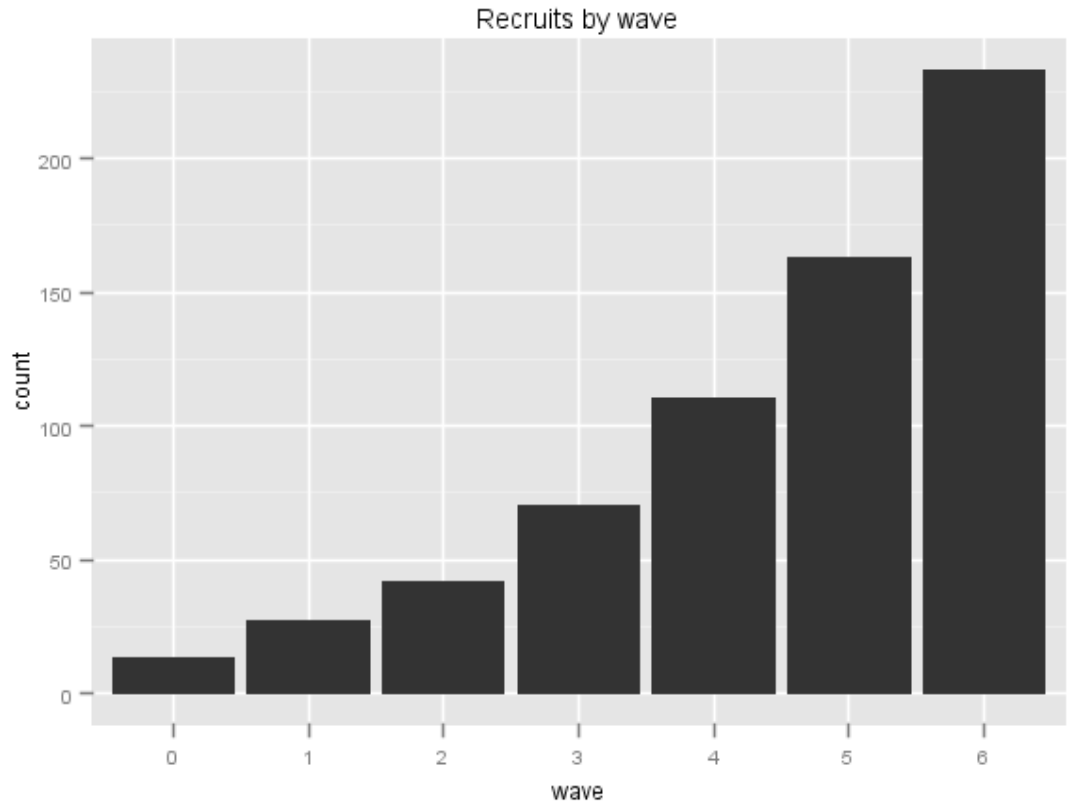


Figure 1.47 Recruits by wave

Recruits per seed

Figure 1.48 displays a bar chart of the number of recruits from each seed stratified by HIV serostatus. To the right of the graph is a legend showing that HIV positive persons are pink and HIV negative persons are blue. This example shows six seeds with different counts for recruits. The fourth bar from the left (seed ID number 322) appears to have the largest number of recruits in the sample and the largest number of HIV positive recruits.

Figure 1.48 Recruits per seed

Recruits per subject

Figure 1.49 shows a bar chart of the number of recruits per subject stratified by HIV serostatus. To the right of the graph is a legend showing HIV negative persons to be in pink, HIV positive persons to be in blue and data that are not available (i.e. recruitment coupons coded as missing data or no longer provided to participants because of the ending of the survey) to be in green. The bar to the far left shows that the majority of HIV positive recruits (in blue) did not recruit any others. In this particular survey, the recruitment process was fast so coupons were reduced from three to two and then from two to three early in the survey.

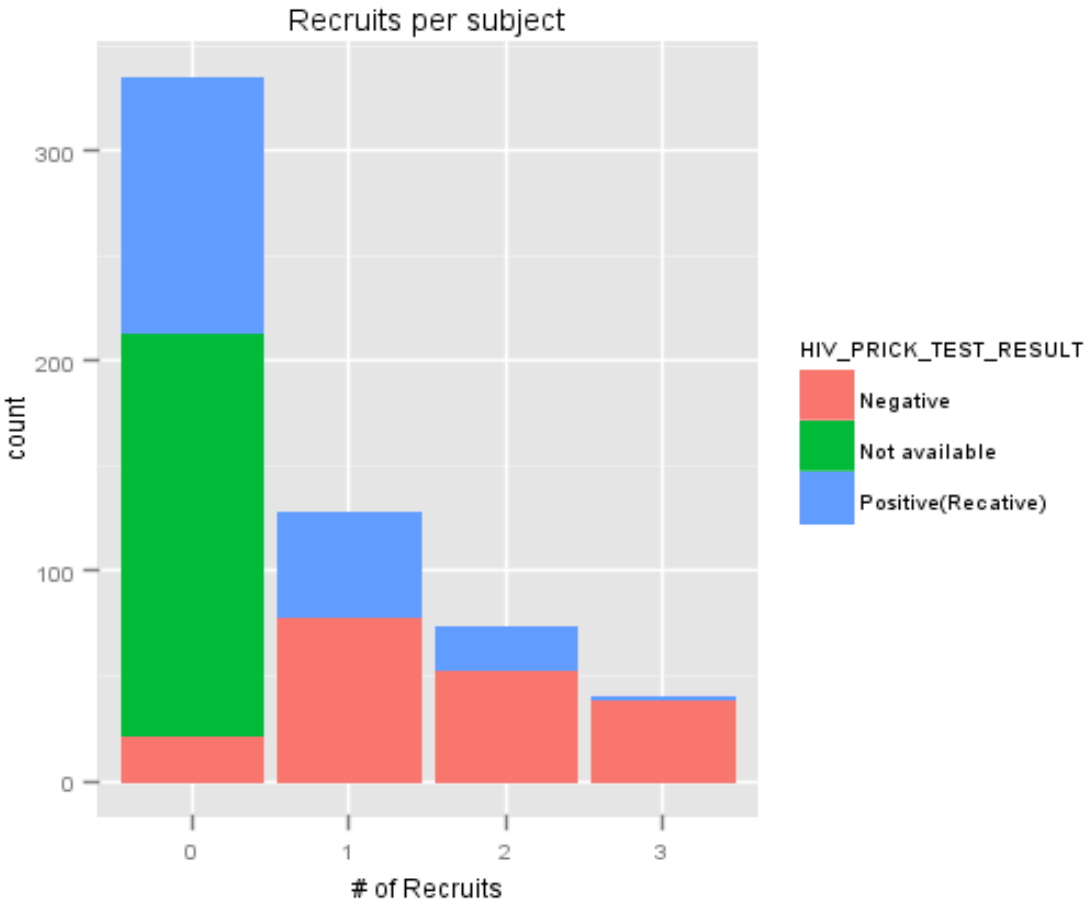


Figure 1.49 Recruits per subject

Conclusion

This step-by-step guide was developed to provide an overview of how to use some of the most useful features of RDS-A. There are a number of other features in RDS-A which can be learnt through practice and by using the manual features available at the HPMRG website. Please be sure to sign up for the RDS-A users group on the HPMRG website in order to provide feedback, offer suggestions and ask for help.

References for the estimator table

Gile K. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, 2011, 106:498:135–146.

Gile K, Handcock MS. Network model-assisted inference from respondent-driven sampling data. Under revision. Available at: http://arxiv.org/pdf/1108.0298v1.pdf, accessed 27 January 2014.

Heckathorn DD. Respondent driven sampling: a new approach to the study of hidden populations. *Social Problems*, 1997, 44(2):174–199.

Heckathorn DD. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 2007, 37:151–207.

Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent driven sampling. *Sociological Methodology*, 2004, 34:193–239.

Schonlau M, Liebau E. Respondent driven sampling (August 2010). DIW Berlin Discussion Paper No.1048. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1679748, accessed 27 January 2014.

Volz E, Heckathorn DD. Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, 2008, 24(Suppl. 1):79–97.

2

Using NetDraw for interpreting RDS data and for problem solving

2. Using NetDraw for interpreting RDS data and for problem solving

Getting started with NetDraw

Introduction

NetDraw[1] is free software for graphing network data. It is commonly used with RDS data to graph recruitment chains. Because RDS data includes information about who recruited whom, we can produce graphs of recruitment chains.

To download NetDraw software go to the Analytic Technologies web site at:

https://sites.google.com/site/netdrawsoftware/home

- Click on "Download"
- Select either "Run/Open" or "Save"
- Double click on the installation file. The installation procedure will then begin. Just follow the prompts (you can just accept all the default choices).

Prepare data for NetDraw

Preparation of data for NetDraw application use can be done in RDS Analyst (www.hpmrg.org) or RDSAT (www.respondentdrivensampling.org).

In RDS Analyst

1. Run data analysis in RDS Analyst (see RDS Analyst manual at www.hpmrg.org on how to format a dataset).
2. To save your data in a DL format, go to the RDS Analyst Console and click File|Save Data| (Figure 2.1).

[1] Borgatti, SP. *NetDraw: Graph visualization software*. Harvard, Analytic Technologies, 2002.

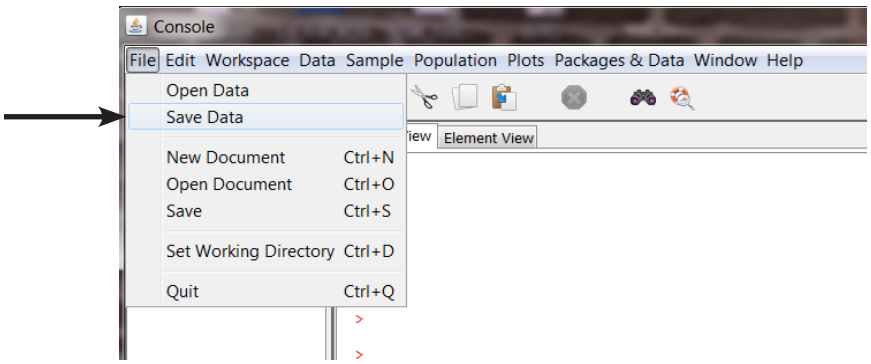


Figure 2.1 RDS Analyst: Console

3. A box will open asking you which dataset you want to save (if you are using multiple datasets at once, select the dataset you want) (Figure 2.2).

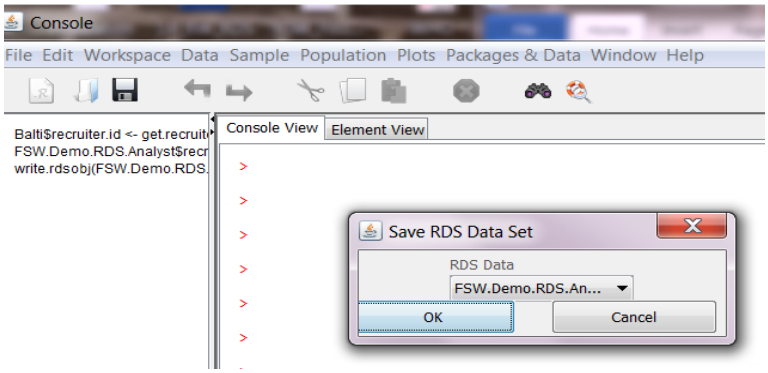


Figure 2.2 RDS Analyst: Saving a file

4. Select the folder in which to place the file(s) and add the file name (Figure 2.3)

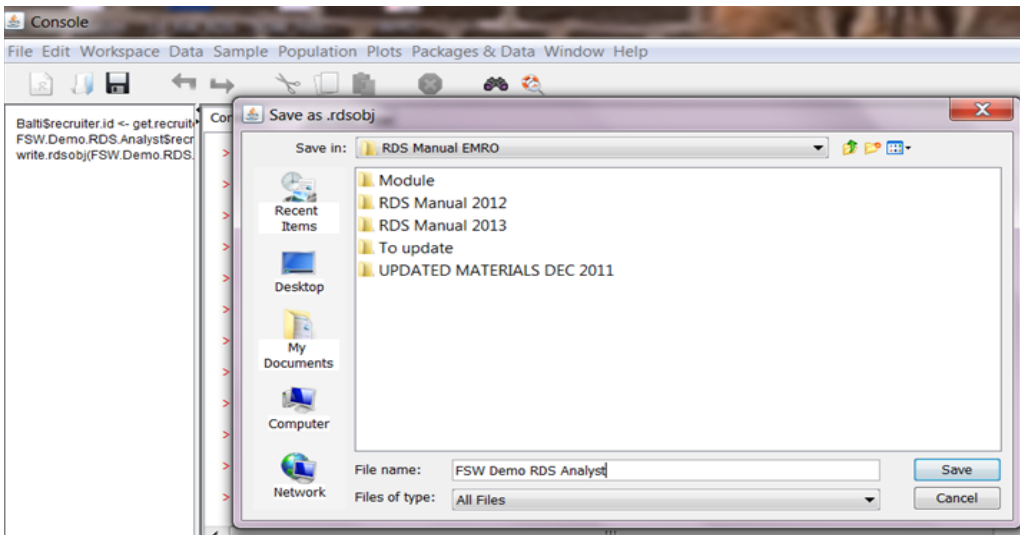


Figure 2.3 RDS Analyst: Naming the file and adding to folder

5. The program will automatically create three files: the DL file, the VNA file (see section on setting up the attribute file on page 52 below) and the RDSOBJ (RDS Analyst modified file) (Figure 2.4).

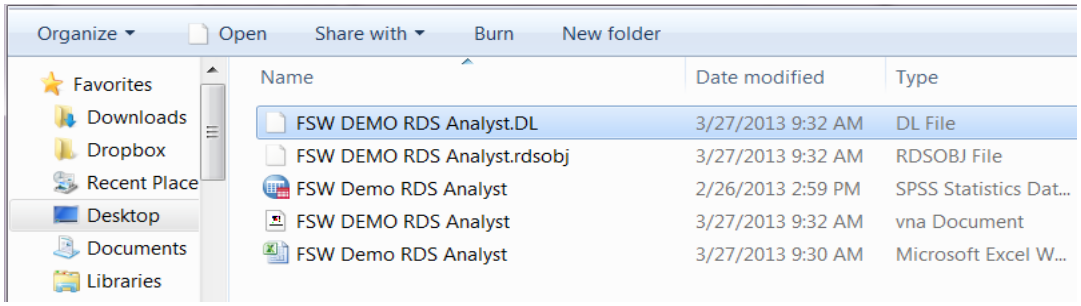


Figure 2.4 RDS Analyst: DL, VNA and RDSOBJ files saved

In RDSAT

1. Run data analysis in RDSAT (see RDSAT step-by-step manual on how to format a dataset at www.respondentdrivensampling.org).
2. To save your data in a DL format, go to File|Export DL Network File| (see Figure 2.5).
3. Change the name of your file by adding DL to the end of the file name.
4. Save in the folder of your choice.

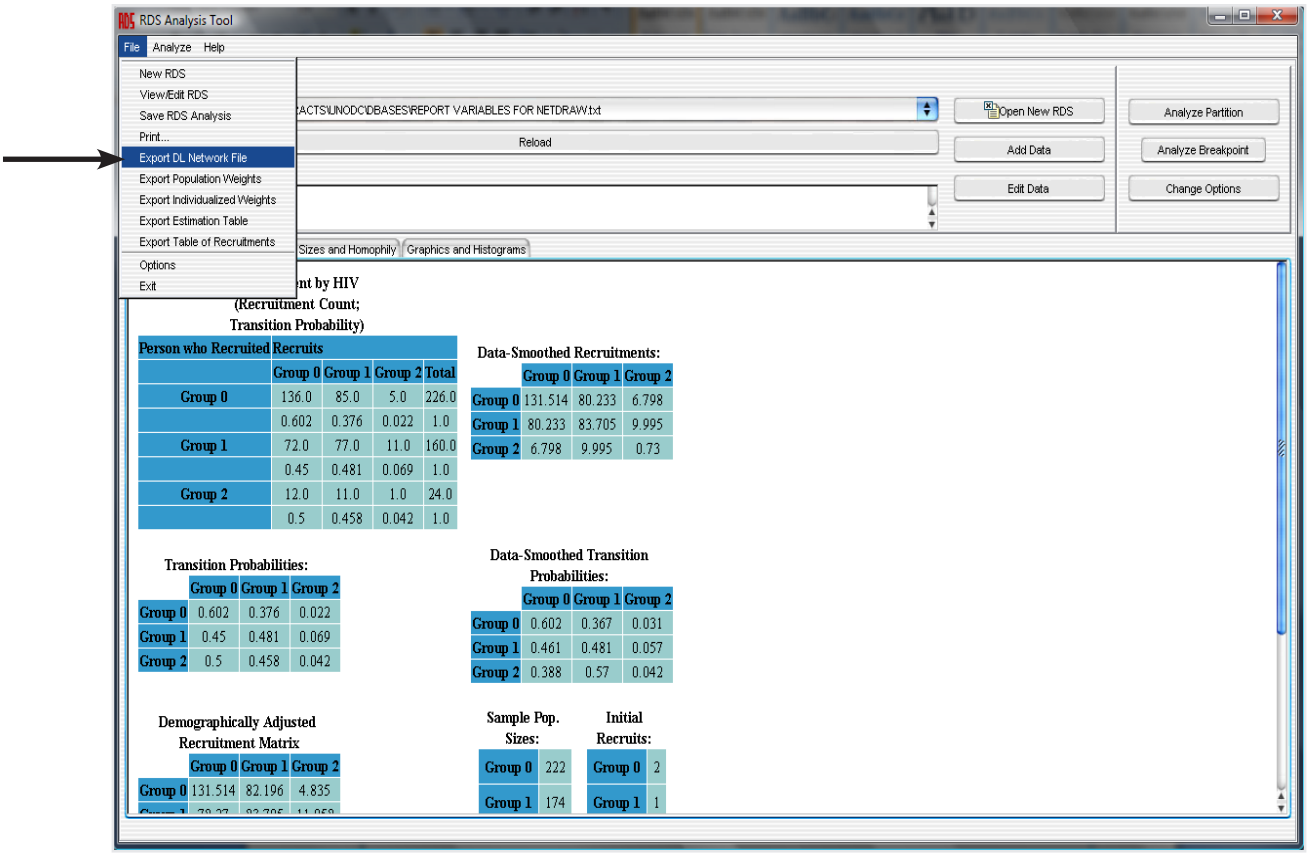


Figure 2.5 RDSAT analysis page

DL files

The DL protocol is a flexible language for describing data. Your dataset in a DL file will look something like Figure 2.6.

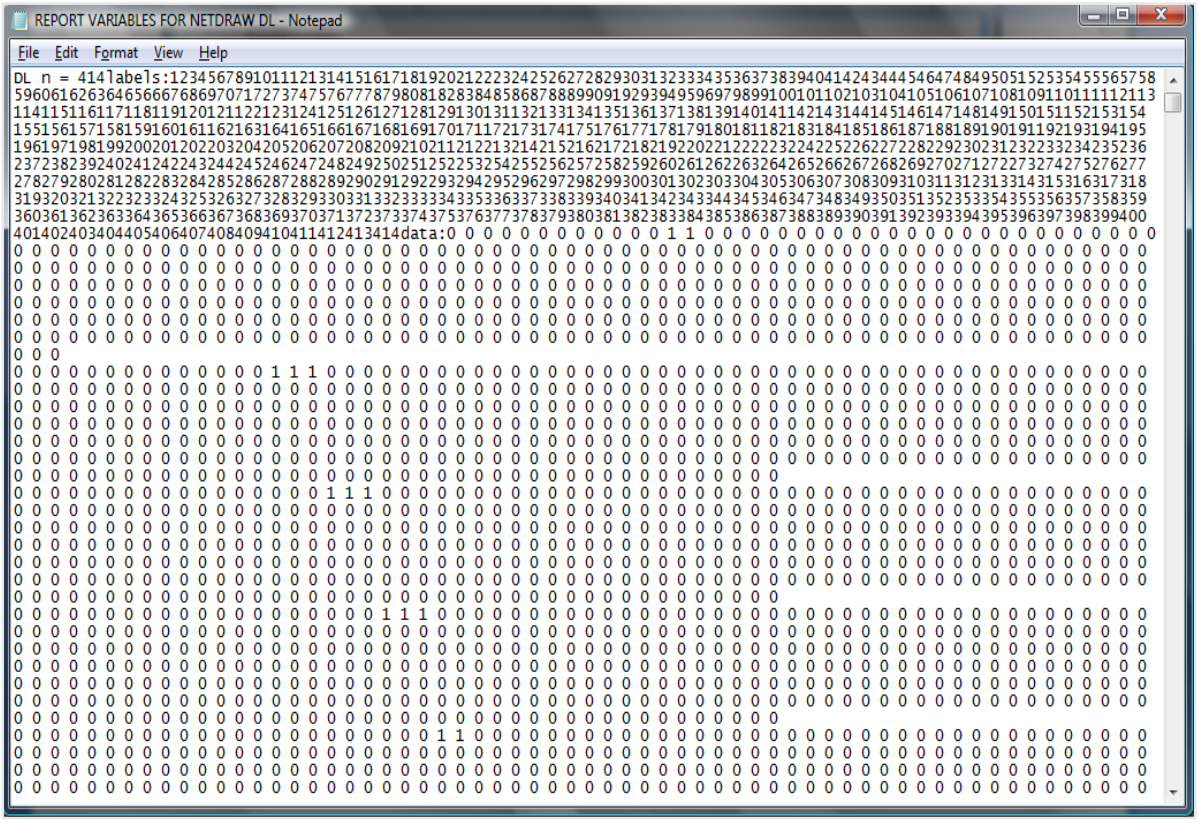


Figure 2.6 Example of a DL file saved as REPORT VARIABLES FOR NETDRAW DL

Where:

DL is the format of the file

N = sample size (414)

Labels: if not modified the NetDraw software will assign a number as label to every node

Data: actual data

Your data are now ready for use in NetDraw.

Generating recruitment chains

Double click on the NetDraw icon to launch application. The NetDraw screen is shown in Figure 2.7.

Figure 2.7 NetDraw screen

Loading the file into Netdraw

To load your file, select the newly created DL file using the browser. Click on File|Open |Ucinet DL text file| (see Figure 2.8). Select appropriate network type ("Network 1-mode"). The screen in Figure 2.9 will appear. Click "OK".

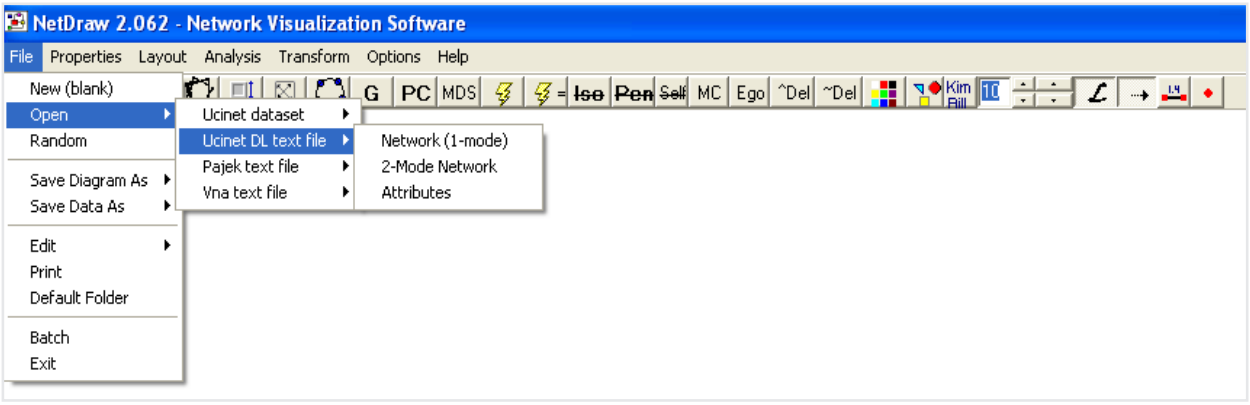


Figure 2.8 NetDraw screen file drop-down box

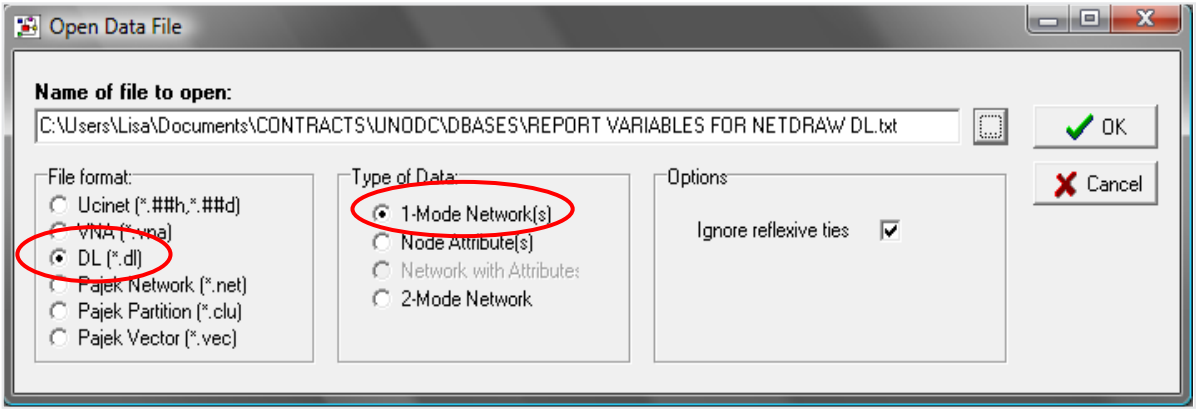


Figure 2.9 Open data file in NetDraw

Upon clicking "OK", a graphic looking something like that displayed in Figure 2.10 or Figure 2.11 will appear. This is a chain made up of 414 nodes. Sometimes the graphic will appear as several isolated nodes. The numbers next to nodes indicate ID numbers generated by NetDraw. In the right lower corner of the screen are check boxes for nodes attributes.

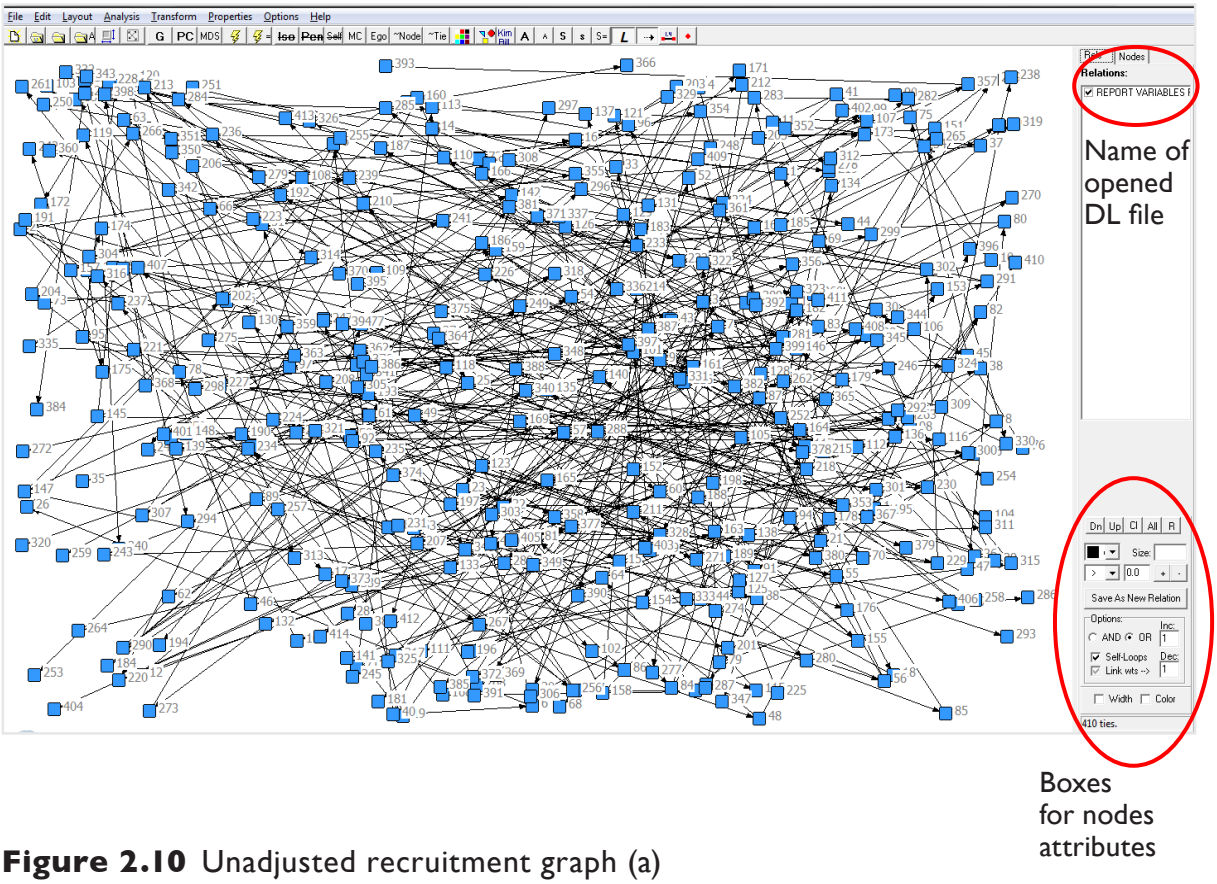


Figure 2.10 Unadjusted recruitment graph (a)

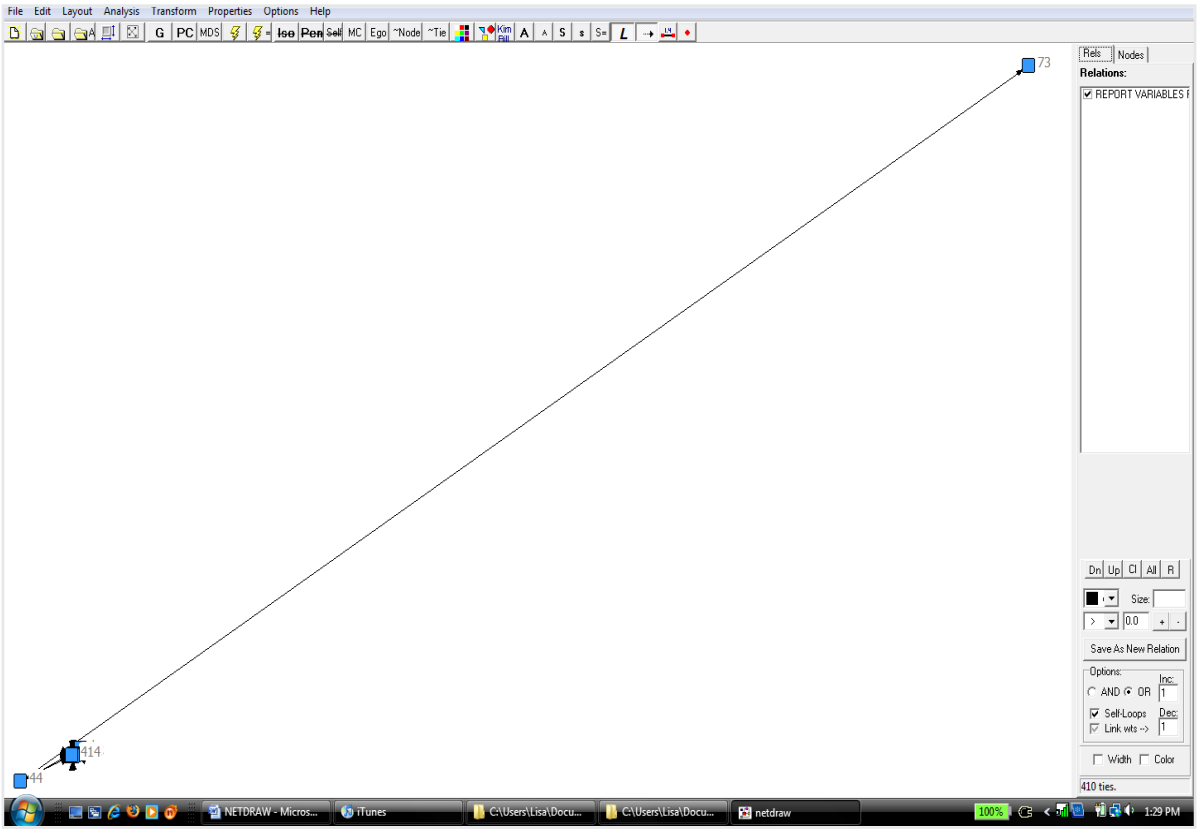


Figure 2.11 Unadjusted recruitment graph (b)

Graphing the recruitment chain

To graph the recruitment chains, Click on Layout|Graph-Theoretic layout|Spring embedding| as shown in Figure 2.12.

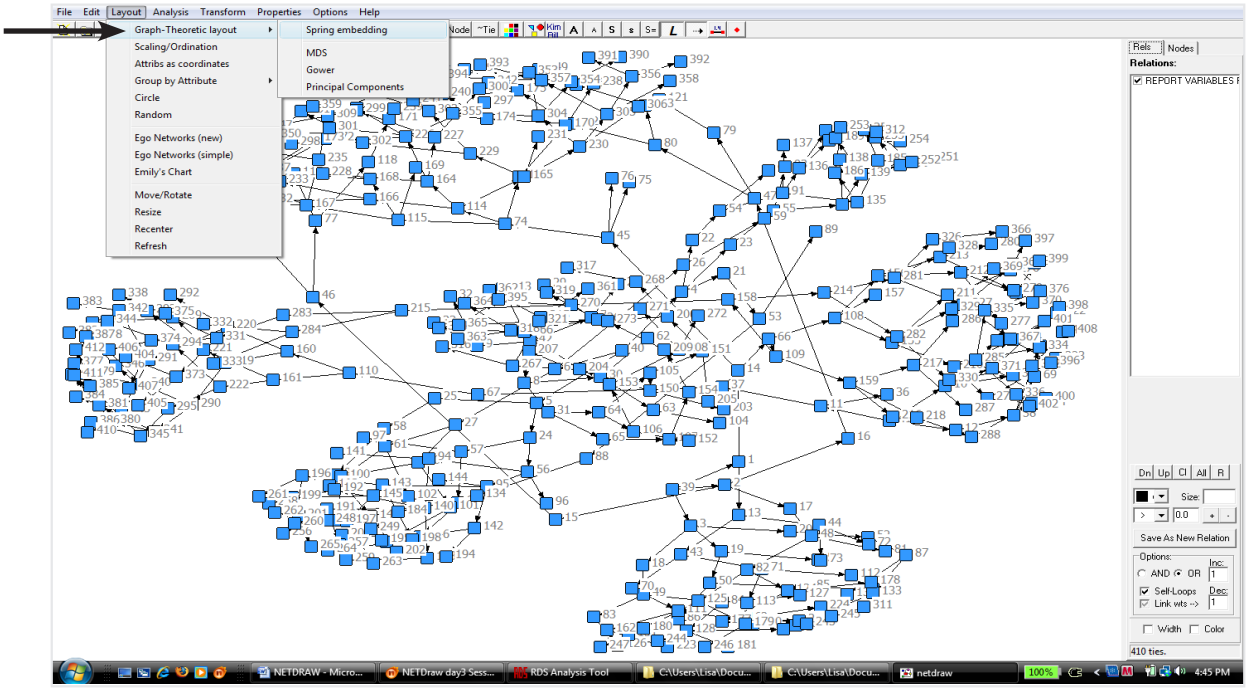


Figure 2.12 Layout

The "Spring Embedding" pop-up box appears (see Figure 2.13). In the menu the number of iterations needs to be increased to at least 1000 and the distance between components needs to be increased to at least 10, but you can play around with it and increase it to 13, 14 or higher. These values will affect how the graph looks. You may have to experiment with the options to see which produce the best display for your data.

You can also use the icon along the top of the display page to make more iterations of the recruitment chains.

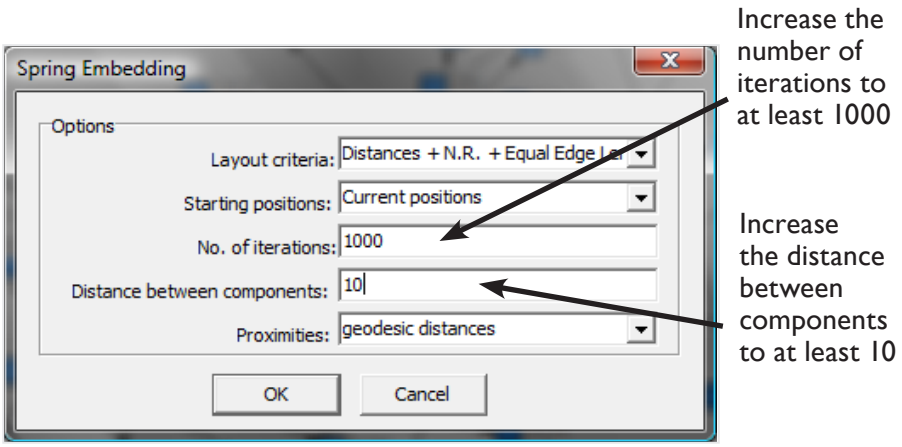


Figure 2.13 Spring Embedding pop-up box

What to look for in the graph

Once the graph has been organized, it should portray the following information (see Figure 2.14):

- Node numbers should match up to the respondent's coupon number in the original dataset.
- There should be the same number of nodes as the sample size. This graph has 414 nodes which is the same as the sample size.
- On the bottom right, there is a display showing the number of ties. This graph has 410 ties (sample size minus the four seeds).
- The number of recruitment clusters should be the same as the number of seeds. This graph has four recruitment clusters, which is equivalent to four seeds.
- Isolated nodes in the upper right hand corner display seeds that have not recruited anyone or are nodes that have been incorrectly numbered. All of the seeds for this sample recruited others and there were no incorrectly numbered participants. If an independent node appears that is not a seed, it will be important to verify why that node is not connected. This may happen if the recording of the identification of who recruited whom is incorrect. To correct this, use the NetDraw recruitment graph to determine which number node is not connected and compare that to the sequential identification numbers (column 2) in the original RDSAT Excel file to find and correct the recruiter's coupon number.
- Correct file name appears on the relations box.

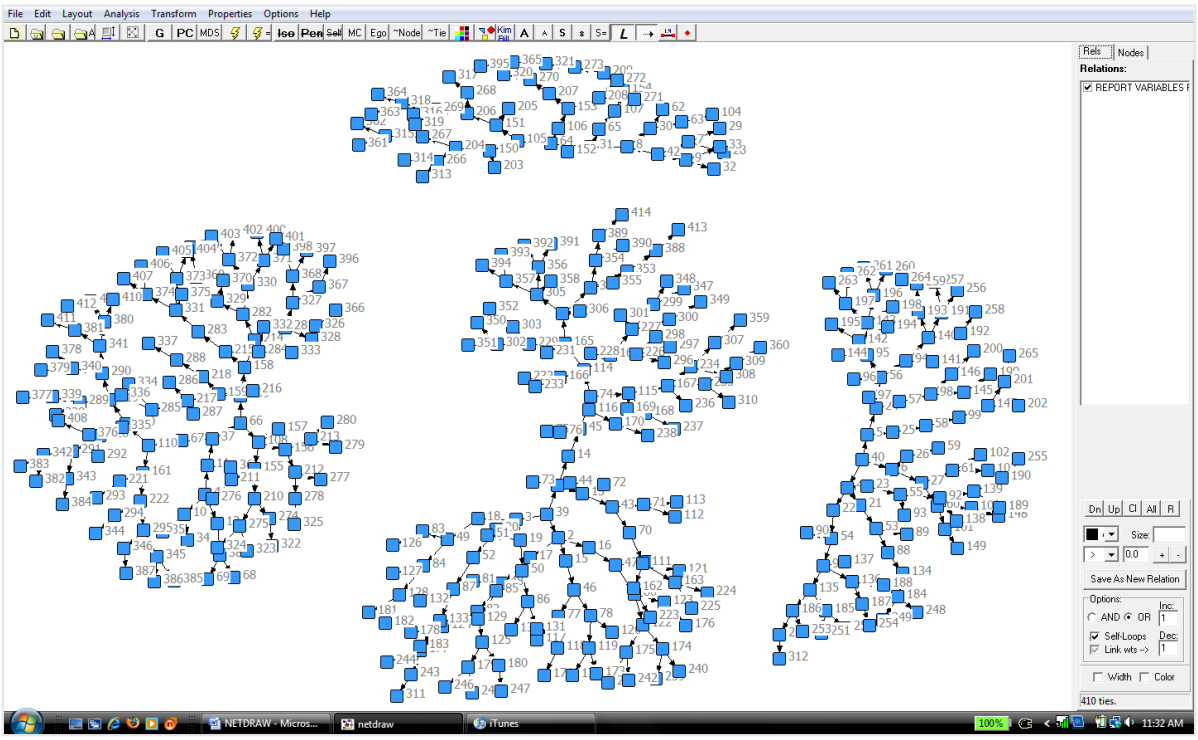


Figure 2.14 A recruitment graph after spring embedding[1]

Colour code components

It is also helpful to be able to see the different recruitment chains. This can be done by using different colours for each network (called components in NetDraw). Click on Analysis menu|Components|Select color|OK on the components dialog box. Once you do this each cluster will have a different colour (see Figure 2.15).

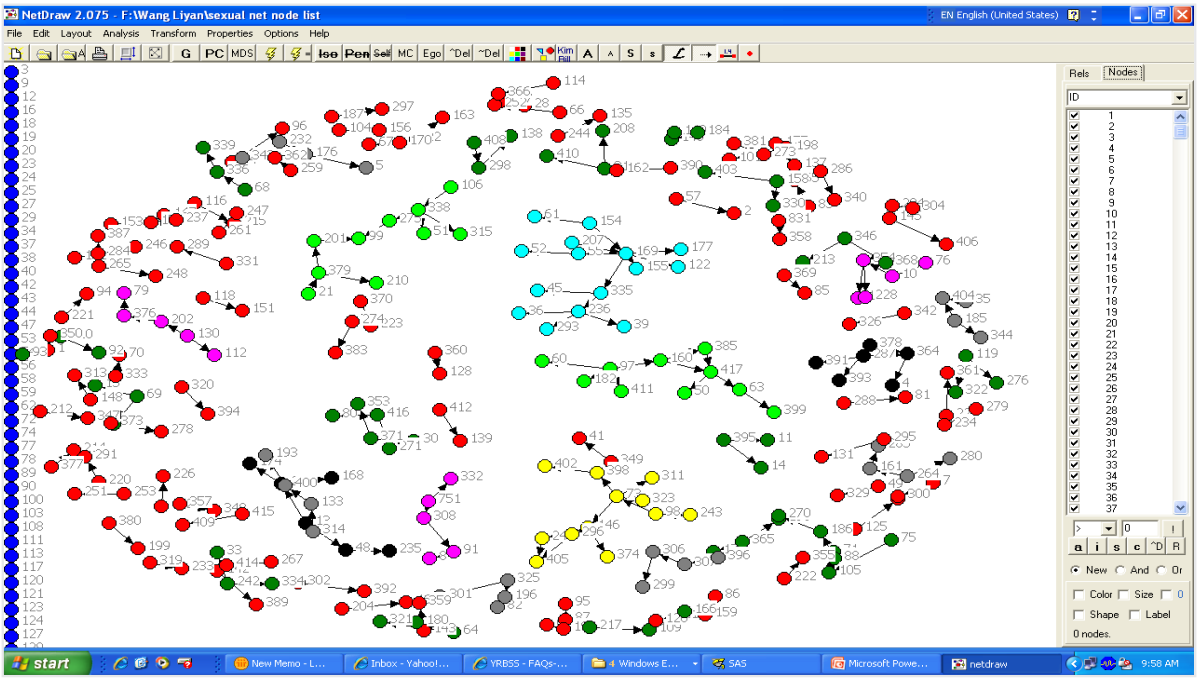


Figure 2.15 Colour coded components

[1] For this file the number of iterations was set to 1000 and the distance between components was set to 16.

The attribute file contains information on the attributes of each case in the data and will allow you to highlight differences (by colour, shape, size, etc.) in the recruitment chains based on the attributes or variables (e.g. HIV positive can be a green colour and shaped as a triangle and HIV negative can be a blue circle).

Setting up the attribute file in RDS Analyst

As described already, prepare data for NetDraw. RDS Analyst automatically creates three files when you save your file from the program. You will need the VNA file for making graphics using attributes (Figure 2.16).

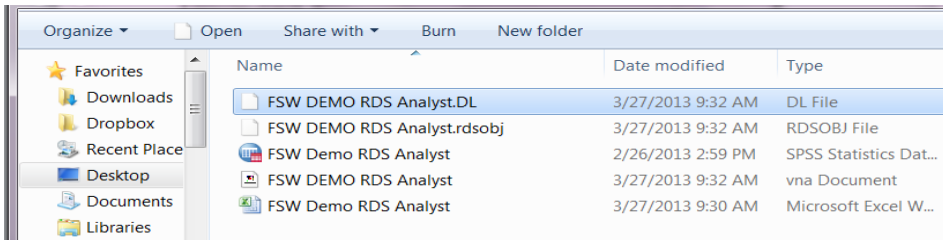


Figure 2.16 RDS Analyst: DL, VNA and RDSOBJ files saved

Setting up the attribute file in RDSAT

Setting up the attribute file in RDSAT requires it to be modified manually from the original RDS data file (see Figure 2.17a and 2.17b). To create the attribute file in RDSAT:

1. Open the original RDS data file.
2. Replace the cell for RDS (cell A1 in Excel) with "*node data".
3. Replace the sample size (cell A2 in Excel) with "ID".
4. Replace the number of coupons cell (B2 in Excel) with "Degree". Degree should now be above the column of participants' social network size.
5. Delete the columns of coupon numbers (this will include the participants' coupons and the coupons given to participants to use in recruiting peers; for instance in this survey a maximum of three coupons were used so that columns C, D, E and F are removed).
6. Save the file as a "Tab delimited text file". Be careful not to overwrite your original RDS files.
7. Close the text file.

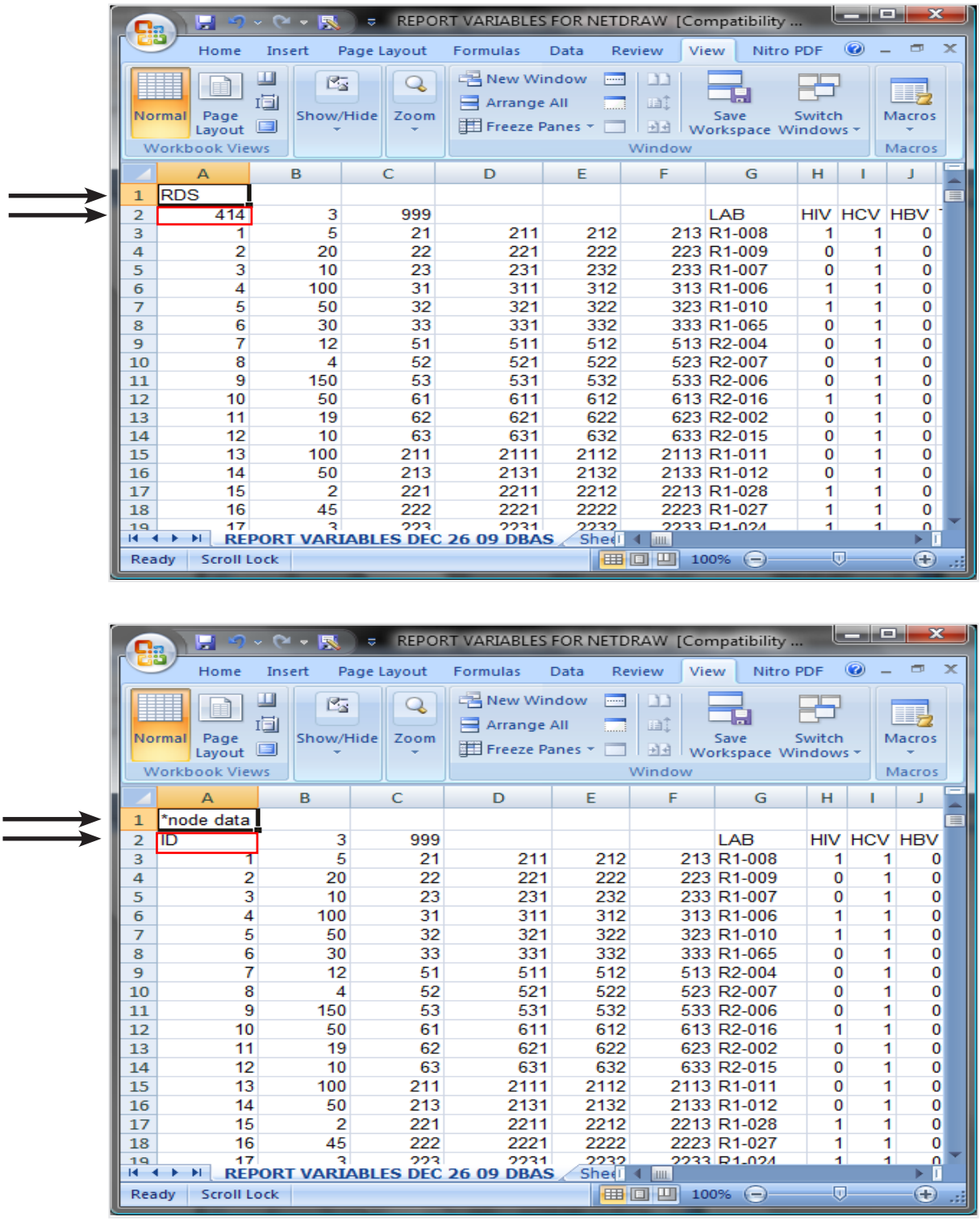


Figure 2.17a Attribute file set-up I

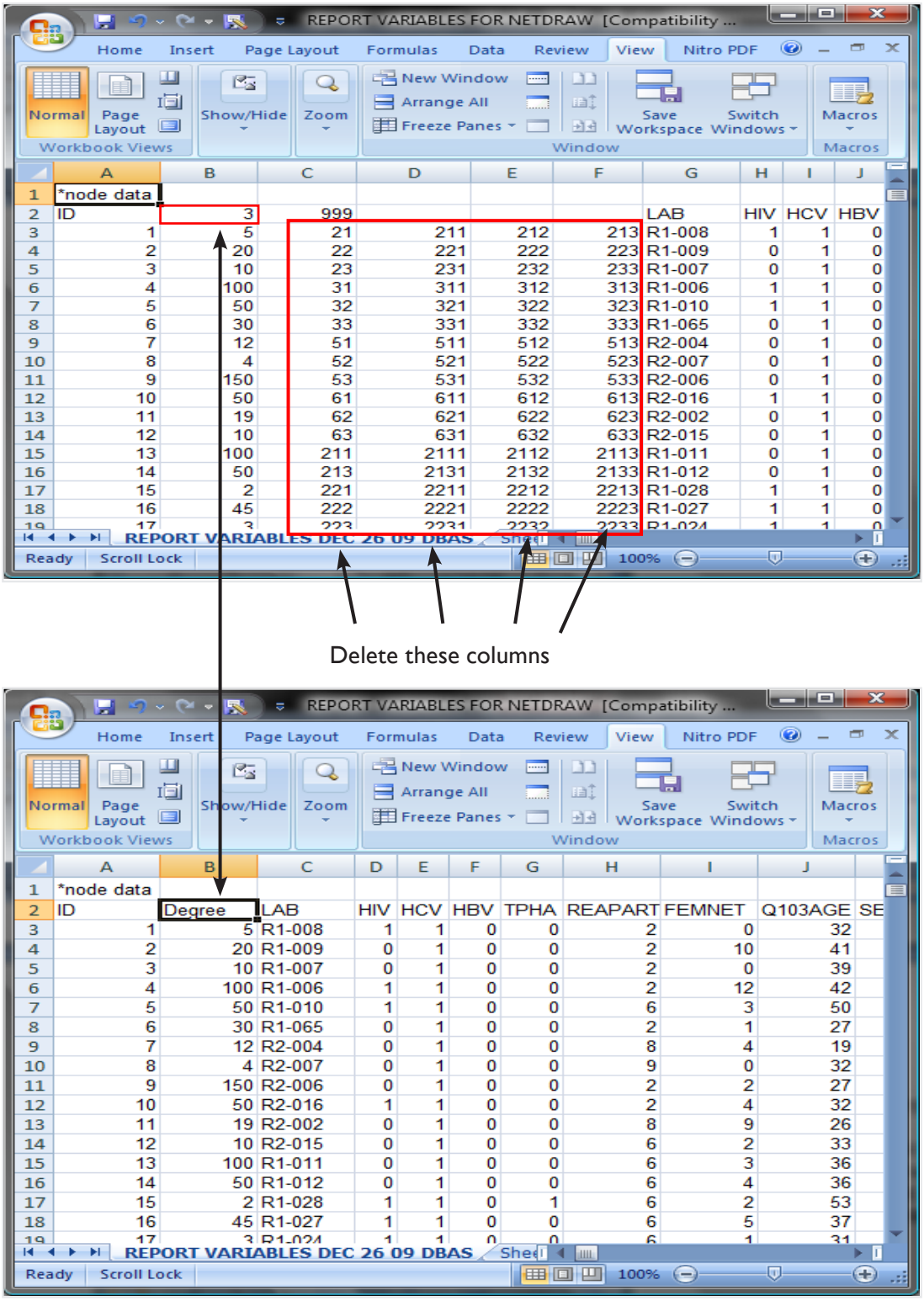


Figure 2.17b Attribute file set-up 2

Importing the attribute file

Once your attribute file is created, import it into NetDraw by doing the following: click File|Open |Vna text file|Attributes| (See Figures 2.18 and 2.19). Use the pop-up browser to find and open the attribute file. Before clicking "OK", make sure the file format is set to "VNA(*.vna)" and type of data to "Node Attributes".

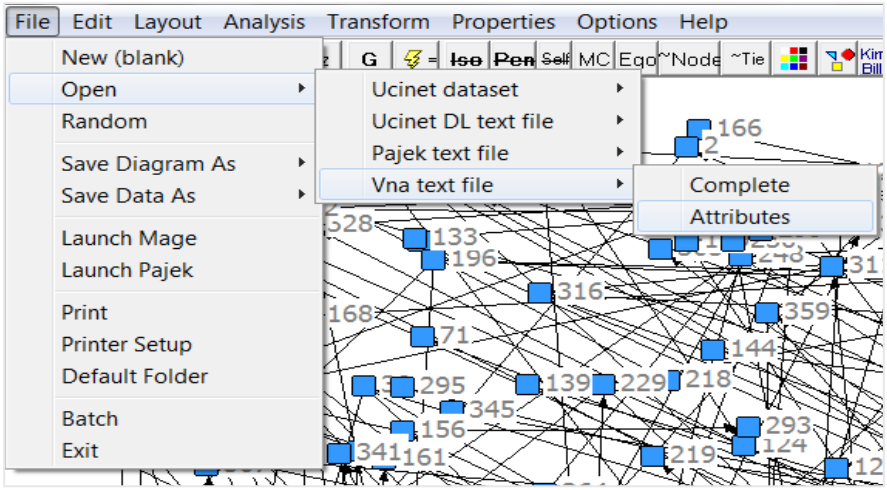


Figure 2.18 Opening an attribute file in NetDraw (a)

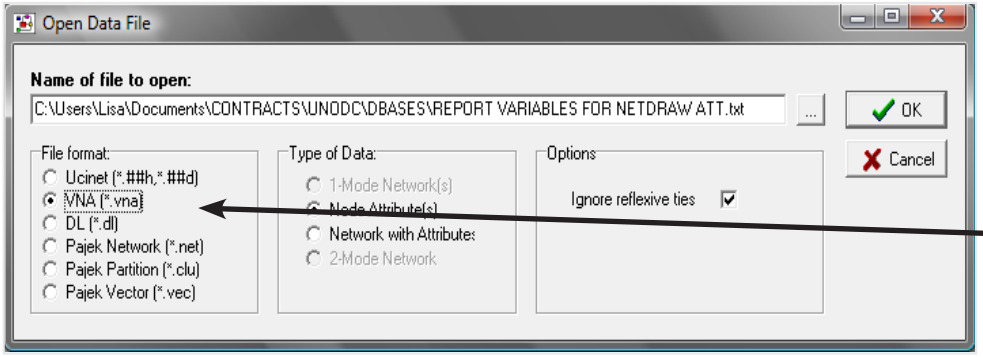


Figure 2.19 Opening an attribute file in NetDraw (b)

If you get the error message (see Figure 2.20), you will need to import the attribute file again. Make sure the VNA (*.vna) file format is selected. This should be the last thing you do before clicking "OK". Another reason you might get this error box is if RDSAT is still open or if your original attribute file database (in Excel) is still open. Make sure you close these before opening the attribute file in NetDraw.

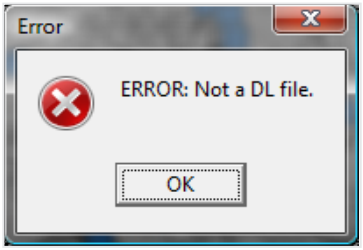


Figure 2.20 Error message

If the attribute file is successfully downloaded into NetDraw, the box in Figure 2.21 ("Node Attribute Editor") may appear, depending on your version of NetDraw. The "InternalID" column (column 1) should match the "ID" column (column 2). If the "InternalID" and the "ID" columns match, then close this box and the recruitment chains will appear. If you do not get this box, do not worry, just continue to follow the instructions below. You may have to repeat the spring embedding settings or click the icon in Figure 2.22 to get distinct clusters of recruitment chains.

Node Attribute Editor

File Edit

InternalID	ID	Degree	LAB	HIV	HCV
1	1	5	R1-008	1	1
2	2	20	R1-009	0	1
3	3	10	R1-007	0	1
4	4	100	R1-006	1	1
5	5	50	R1-010	1	1
6	6	30	R1-065	0	1
7	7	12	R2-004	0	1
8	8	4	R2-007	0	1
9	9	150	R2-006	0	1
10	10	50	R2-016	1	1
11	11	19	R2-002	0	1
12	12	10	R2-015	0	1
13	13	100	R1-011	0	1
14	14	50	R1-012	0	1

Figure 2.21 Node Attribute Editor box

Figure 2.22 Recruitment train icon

Moving nodes

Selected nodes can be dragged to other parts of the graph by clicking on a particular node and then dragging it (see Figure 2.23). You will know if the node has been properly selected when it turns from a solid colour to a hashed colour. If multiple nodes are selected and one is moved, the rest will follow, maintaining their structure or you can create a highlight box (Note: highlight box only works starting at the upper left corner) and dragging the nodes into the box.

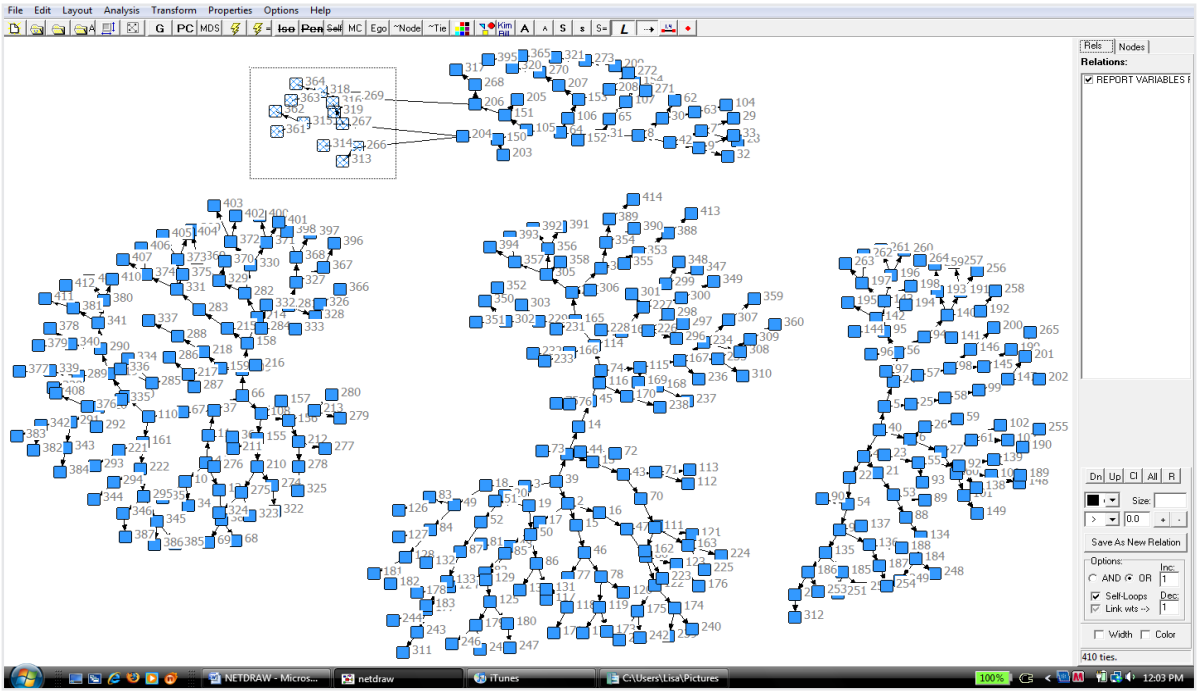


Figure 2.23 Selecting nodes and dragging them

Creating graphs with the attribute file

To create graphs with the attribute file, you will have to use the "Properties" menu. The Properties menu will allow you to change nodes, lines and background sizes, shapes and colours. There are two types of options:

1. "General" option: this assigns the same sizes, shapes and colours to all visible object in a category.
2. "Attribute-based" option: this assigns properties by attribute.

To select the General or Attribute-based options for colour click Properties|Nodes|Symbols |Color|General-all active nodes| or |Attribute-based| (see Figure 2.24).

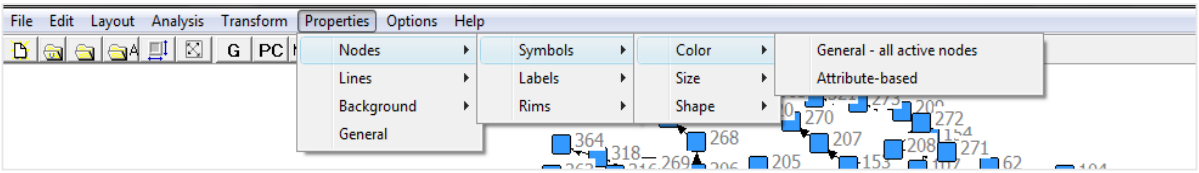


Figure 2.24 Selecting general or attribute-based options

Attribute-based options

This section will provide some examples of using the attribute-based node properties. To select a shape to distinguish nodes by attributes, click Properties|Nodes|Symbols|Shape| Attribute-based. A box entitled "Shapes of Nodes" will appear. Use the drop down menu to select a variable (see Figure 2.25). The variable selected for this example is "Sex" (box 1) and the shapes selected are "Circle" for 1 (males) and a "Square" for 2 (females). There are several shapes to choose from. After selecting the shapes, click the box with the green check symbol.

The same procedure can be used to set node colour and size.

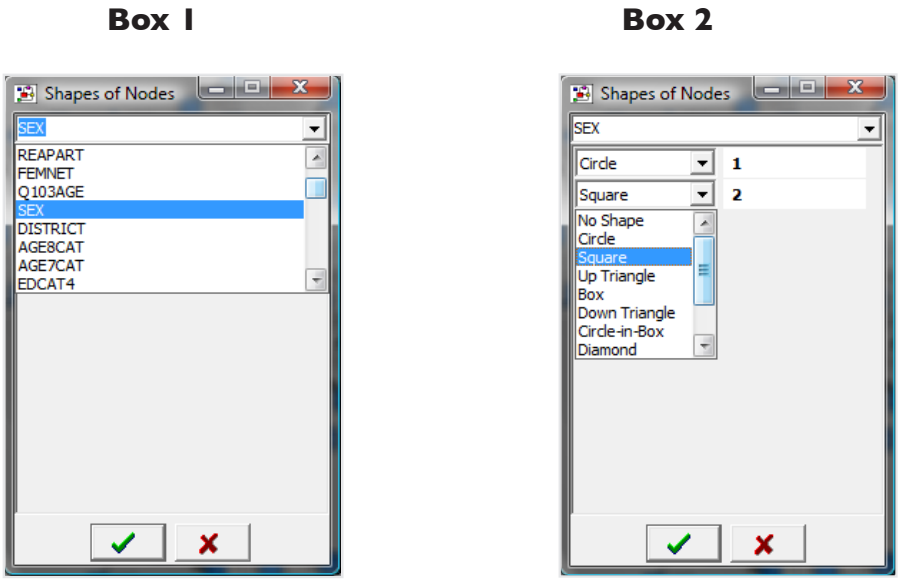


Figure 2.25 Shapes of nodes boxes

Figure 2.26 shows one of the four recruitment chains with squares and circles. Nodes are shape coded for sex.

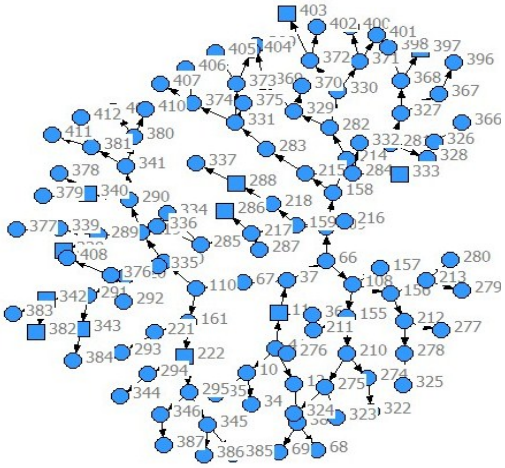


Figure 2.26 Shape coded recruitment chain

Nodes tab

The right side of the NetDraw display window has a series of useful tabs.

- The "Rels" and "Ties" tabs are not used for RDS data.
- The "Nodes" tab allows the selection of visible nodes by variable value (upper part of box). To locate a variable in the dataset, use the down arrow (see Figure 2.27). In this example the variable "SEX" is selected.
- By unchecking the box all nodes matching that value are no longer visible (Figure 2.28). In this example, value 1 for the variable SEX is being selected. Only value 2 remains in the recruitment chains.

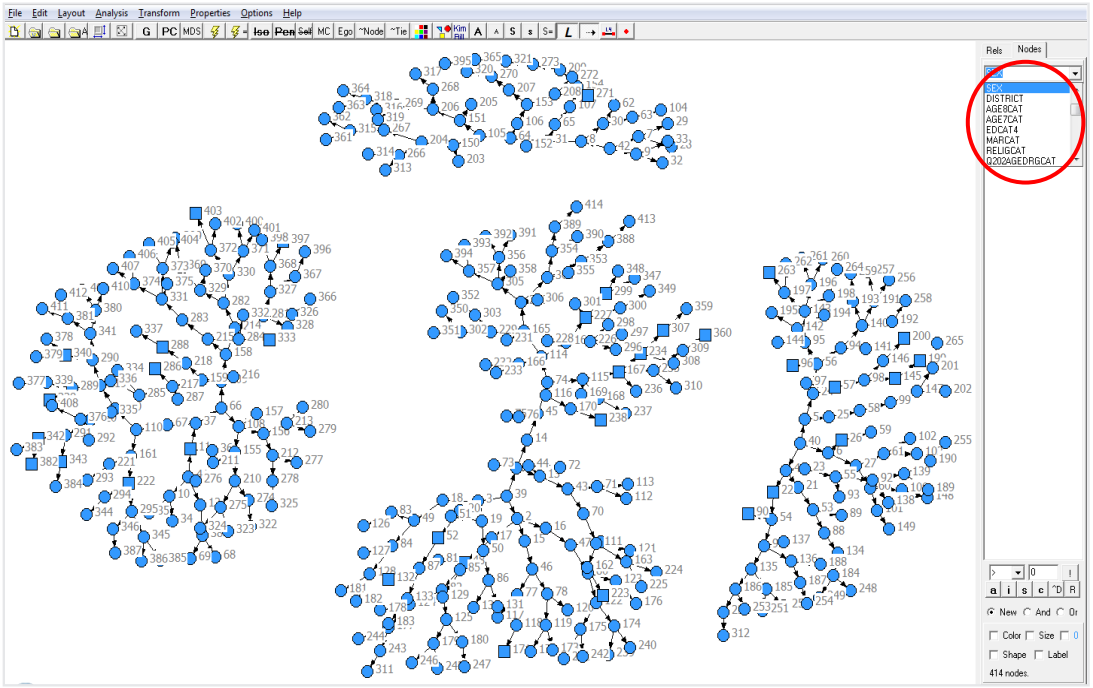


Figure 2.27 Displaying nodes by variable values (a)

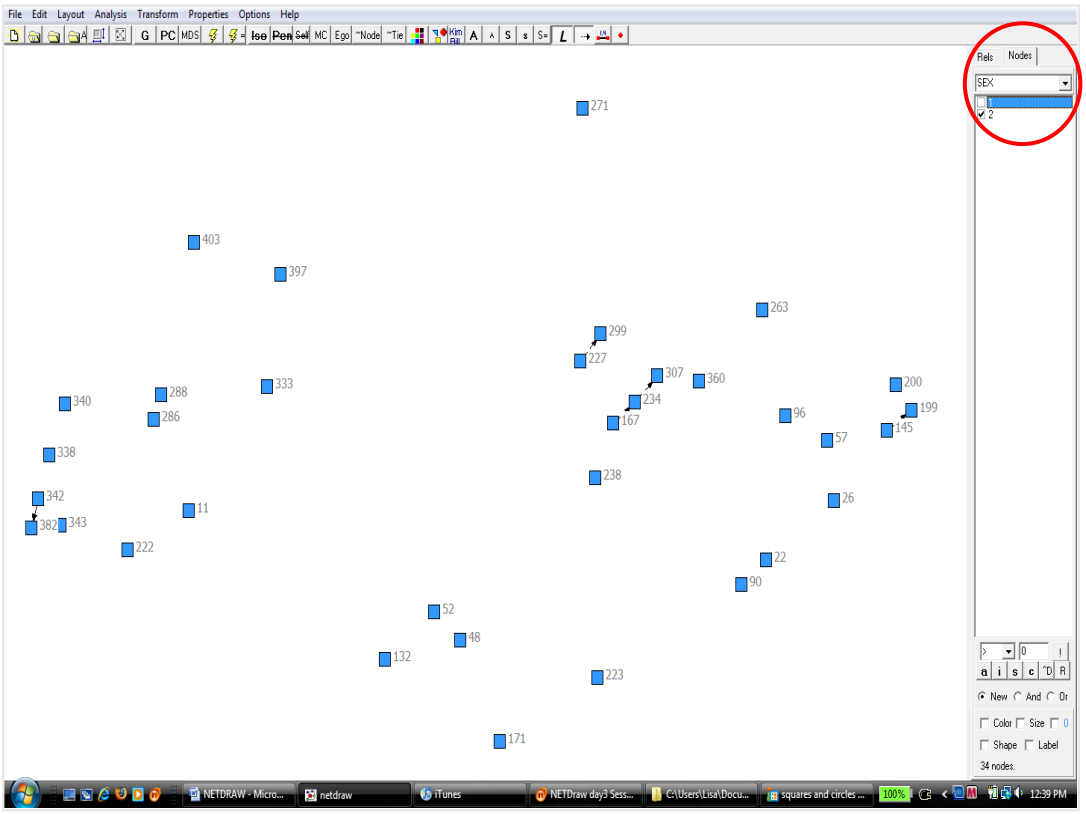


Figure 2.28 Displaying nodes by variable values (b)

- On the lower part of the right hand side of the NetDraw display window are six "short-cut" buttons that are useful for manipulating visible and invisible nodes (see Figure 2.29). These buttons are:
  - **a** All categories—nodes visible
  - **i** Switch—checked categories are visible and unchecked categories are hidden
  - **s** Next category—selects the next category and unselects the current category
  - **c** Add next—adds the next category to the selection
  - ^D Permanently delete—permanently deletes all hidden nodes
  - R Return—returns to the original respondent ID variable.
- The bottom of the right hand side of the NewDraw display window has boxes that can be checked for defaulted colour, size, shape and label.

Figure 2.29 Short-cut buttons

Displaying labels

To get rid of node labels click Properties|Nodes|Labels|Visible and the "Node Labels" box in Figure 2.30 should appear. To get rid of labels (e.g. ID numbers), click "Off" under "Turn labels..." (left side of box). You can apply this to all nodes or only active nodes (right side of box). An easier way to do this is to use the icons at the top of the NetDraw display window. The icon "**L**" allows labels to be turned on and off.

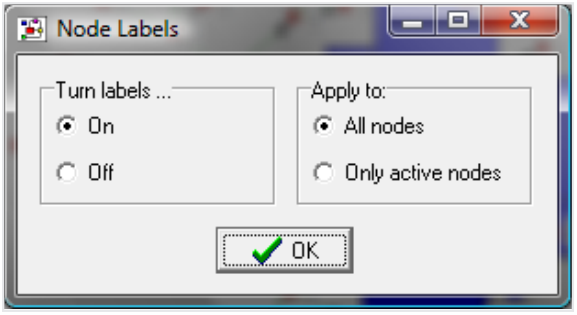


Figure 2.30 Node Labels box

To label nodes by variable value, use the box for "Label" at the bottom of right-hand side of the NetDraw display window. By selecting the variable in the "Nodes" box (upper right hand side of the NetDraw display window) and then checking the "Label" box at the bottom, the nodes will be labelled according to the values given to each category.

To assign text labels, it is possible to click Properties|Nodes|Labels|Text| which will display the "Label Text" box (see Figure 2.31). Under the "Label" column it is possible to put in a text label such as "Male" or "Female" for the variable "Sex", corresponding to the appropriate value (males for 1 and females for 2).

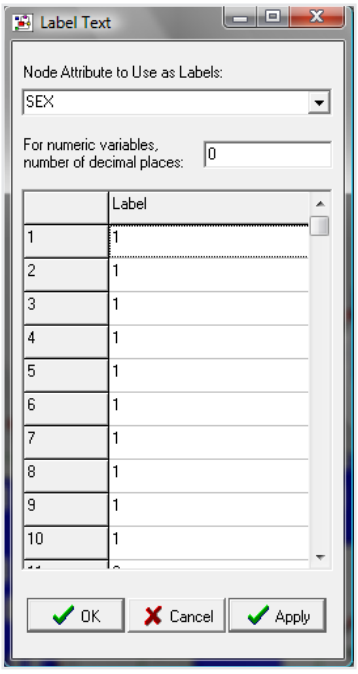
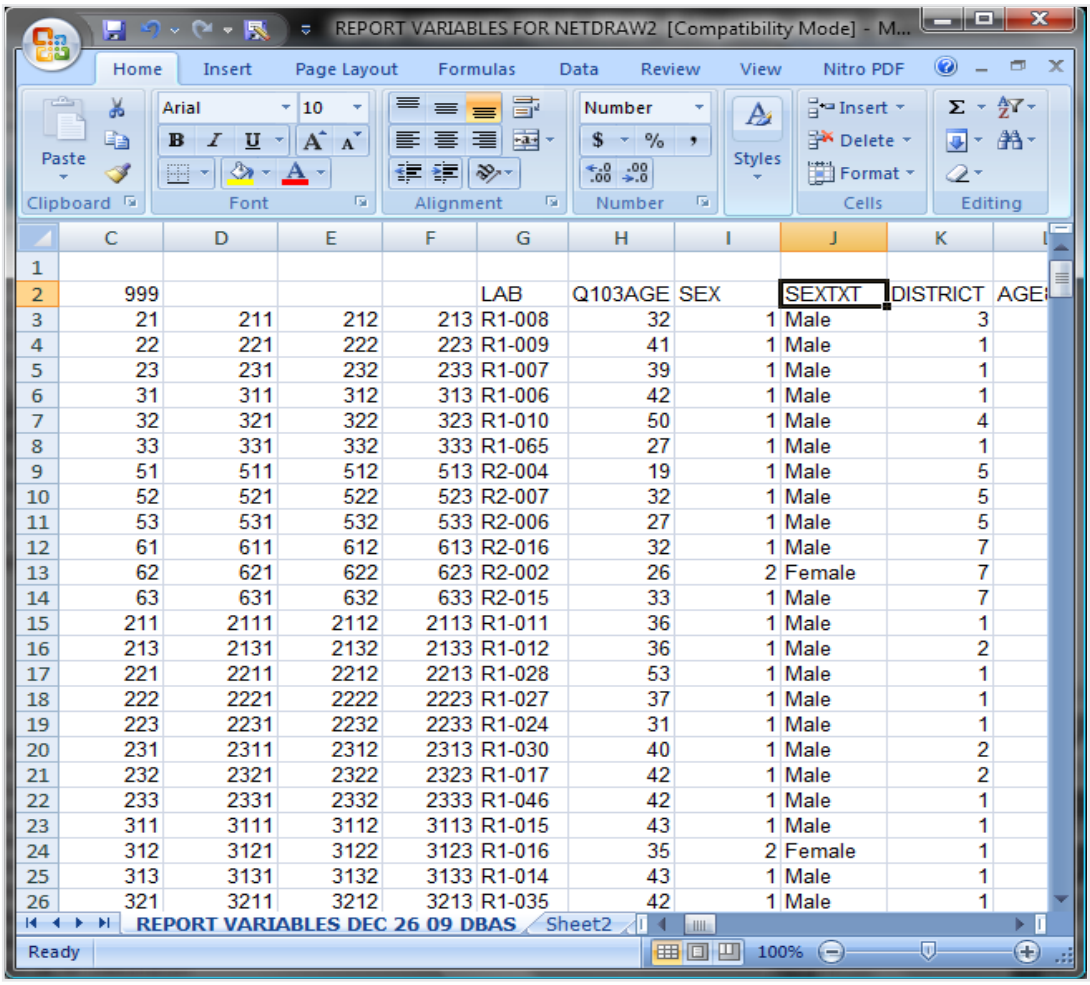


Figure 2.31 Label Text box

An easier way to do this may be to go back to your original database and copy and paste the sex column into a new column (see "SEXTXT", in Figure 2.32) and "find" the values (e.g. 1) and "replace" them with the corresponding text (e.g. "Male ") (Figure 2.32).



	C	D	E	F	G	H	I	J	K
1									
2	999				LAB	Q103AGE	SEX	SEXTXT	DISTRICT
3	21	211	212	213	R1-008	32	1	Male	3
4	22	221	222	223	R1-009	41	1	Male	1
5	23	231	232	233	R1-007	39	1	Male	1
6	31	311	312	313	R1-006	42	1	Male	1
7	32	321	322	323	R1-010	50	1	Male	4
8	33	331	332	333	R1-065	27	1	Male	1
9	51	511	512	513	R2-004	19	1	Male	5
10	52	521	522	523	R2-007	32	1	Male	5
11	53	531	532	533	R2-006	27	1	Male	5
12	61	611	612	613	R2-016	32	1	Male	7
13	62	621	622	623	R2-002	26	2	Female	7
14	63	631	632	633	R2-015	33	1	Male	7
15	211	2111	2112	2113	R1-011	36	1	Male	1
16	213	2131	2132	2133	R1-012	36	1	Male	2
17	221	2211	2212	2213	R1-028	53	1	Male	1
18	222	2221	2222	2223	R1-027	37	1	Male	1
19	223	2231	2232	2233	R1-024	31	1	Male	1
20	231	2311	2312	2313	R1-030	40	1	Male	2
21	232	2321	2322	2323	R1-017	42	1	Male	2
22	233	2331	2332	2333	R1-046	42	1	Male	1
23	311	3111	3112	3113	R1-015	43	1	Male	1
24	312	3121	3122	3123	R1-016	35	2	Female	1
25	313	3131	3132	3133	R1-014	43	1	Male	1
26	321	3211	3212	3213	R1-035	42	1	Male	1

Figure 2.32 Creating text labels in Excel

If text is created after already downloading your DL and Attribute files into NetDraw, it will be necessary to create and reload the new files with the text coding into NetDraw. Once the DL and then the attribute files (NetDraw must read the DL file before it can read the attribute file) with the text is loaded into NetDraw:

- go to the upper right-hand side of the display window and select the variable (in this example the variable is "SEXTXT")
- go to the lower right-hand side of the display window and select the label box.

The labels should appear next to the appropriate nodes as seen in Figure 2.33.

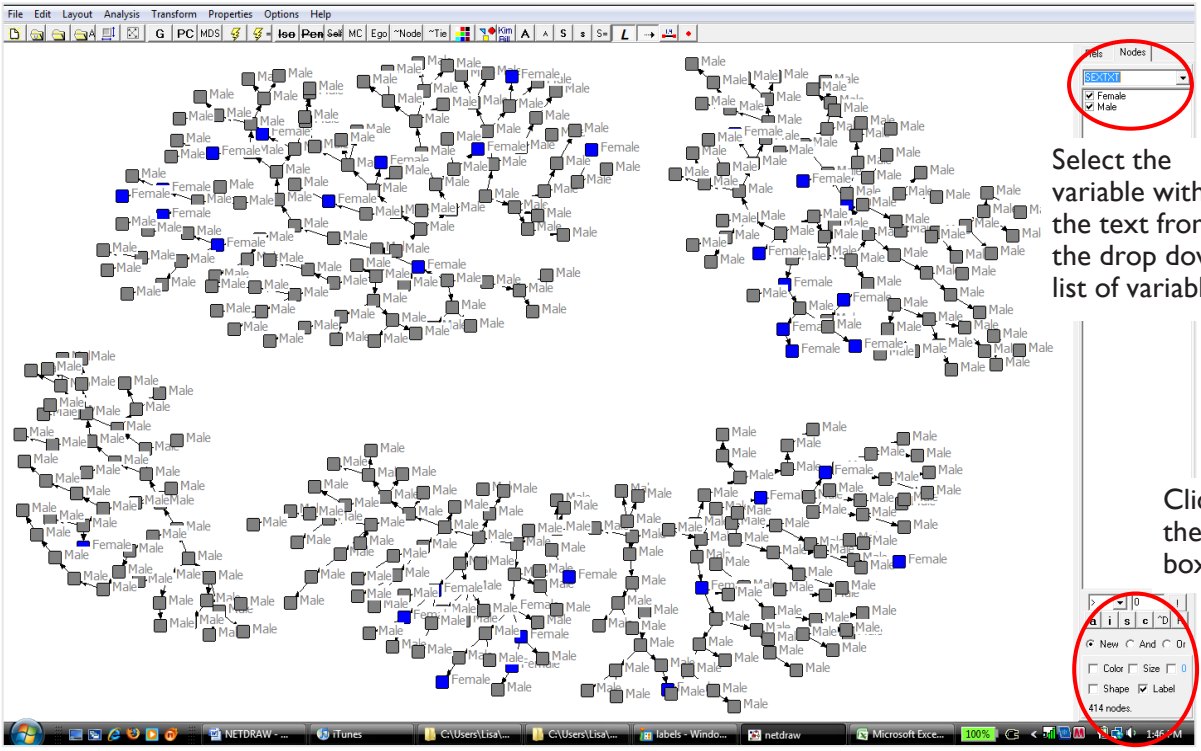
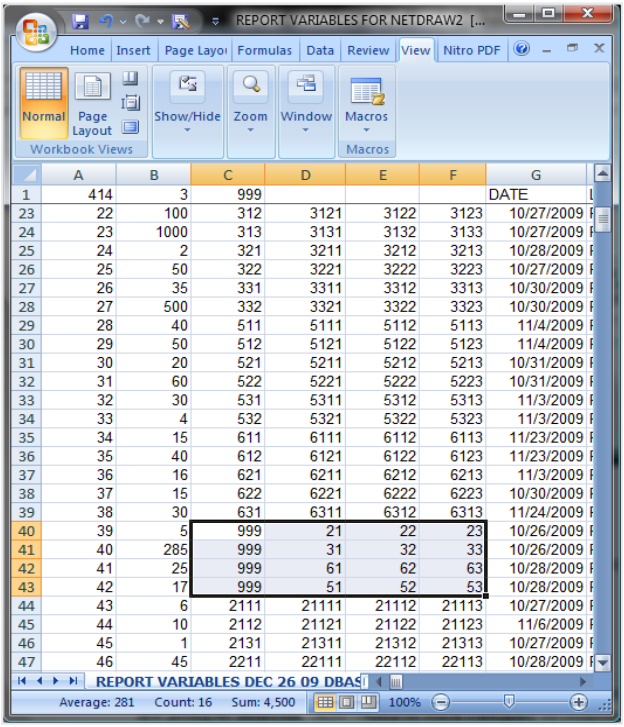


Figure 2.33 Recruitment chains with text labels

Finding seeds

To find seeds in the recruitment graphs it is necessary to look in the original database to find the IDs connected with the seeds. In the database below (see Figure 2.34) the four seeds have ID numbers 40, 41, 42 and 43 (use the ID numbers assigned by Excel in the far left column [not column A] as these will be the same ID numbers in the attribute file).



	A	B	C	D	E	F	G
1	414	3	999				DATE
23	22	100	312	3121	3122	3123	10/27/2009
24	23	1000	313	3131	3132	3133	10/27/2009
25	24	2	321	3211	3212	3213	10/28/2009
26	25	50	322	3221	3222	3223	10/27/2009
27	26	35	331	3311	3312	3313	10/30/2009
28	27	500	332	3321	3322	3323	10/30/2009
29	28	40	511	5111	5112	5113	11/4/2009
30	29	50	512	5121	5122	5123	11/4/2009
31	30	20	521	5211	5212	5213	10/31/2009
32	31	60	522	5221	5222	5223	10/31/2009
33	32	30	531	5311	5312	5313	11/3/2009
34	33	4	532	5321	5322	5323	11/3/2009
35	34	15	611	6111	6112	6113	11/23/2009
36	35	40	612	6121	6122	6123	11/23/2009
37	36	16	621	6211	6212	6213	11/3/2009
38	37	15	622	6221	6222	6223	10/30/2009
39	38	30	631	6311	6312	6313	11/24/2009
40	39	5	999	21	22	23	10/26/2009
41	40	285	999	31	32	33	10/26/2009
42	41	25	999	61	62	63	10/28/2009
43	42	17	999	51	52	53	10/28/2009
44	43	6	2111	21111	21112	21113	10/27/2009
45	44	10	2112	21121	21122	21123	11/6/2009
46	45	1	2131	21311	21312	21313	10/27/2009
47	46	45	2211	22111	22112	22113	10/28/2009

Figure 2.34 Seeds in the Excel file

Find the seeds by looking at ID numbers under the "Nodes" tab. Upon selecting the corresponding ID number, the seed will disappear. In Figure 2.35, the red circle indicates where the seed for ID 42 was located. The identified seed can be distinguished by right clicking the mouse and selecting one of the options. In Figure 2.35, seeds 40, 41 and 43 are distinguished by large size and different colour.

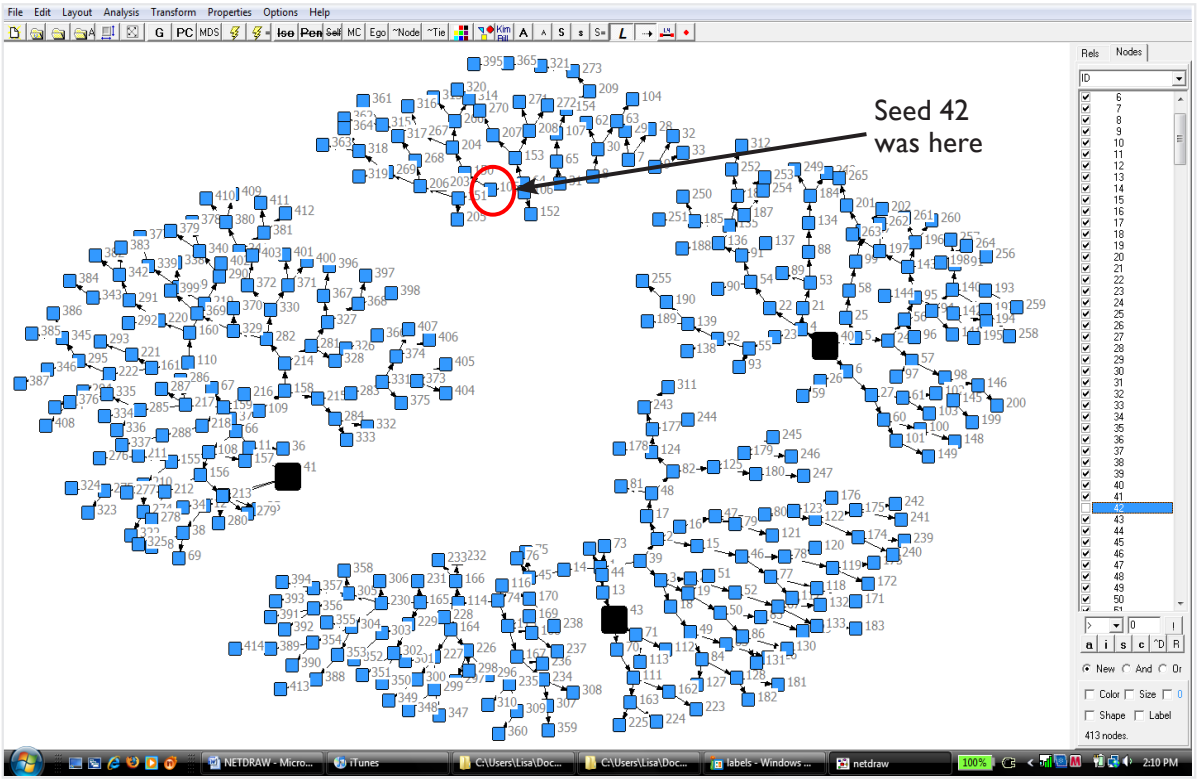


Figure 2.35 Finding and distinguishing seeds

Making seeds with rims

To change the node rim size and colour of the seeds, click "i" (one of the six buttons at the bottom of the right hand side of display screen) to uncheck all nodes (all nodes will disappear from the screen). Then check just the seed nodes (i.e. 40, 41, 42 and 43) and they will appear (see Figure 2.36). Then click Properties|Nodes|Rims|Color|General-all active nodes| and the "Rim Color" box will appear. Click a rim colour. In the example in Figure 2.37, the rim colour selected is "dLime".

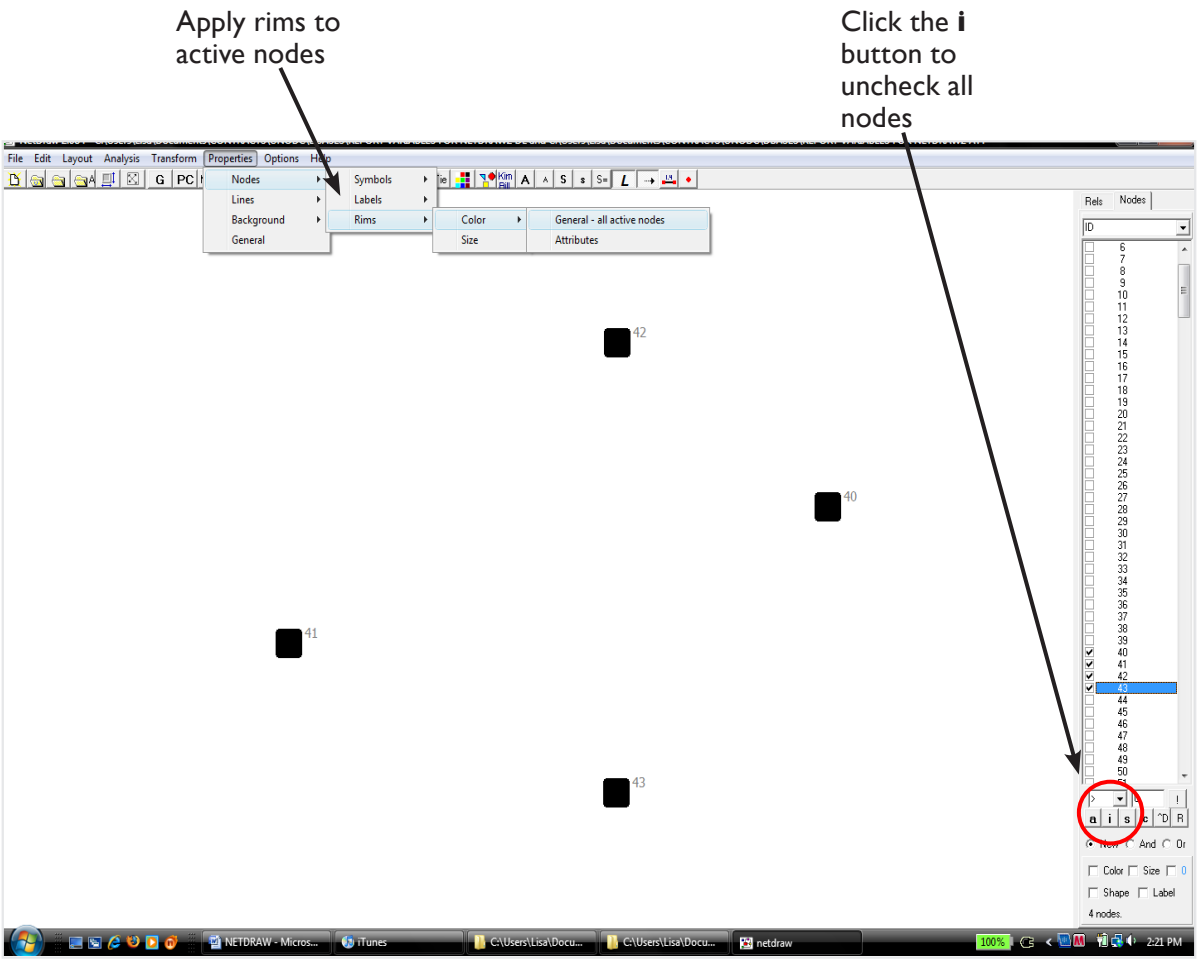


Figure 2.36 Isolating the seeds

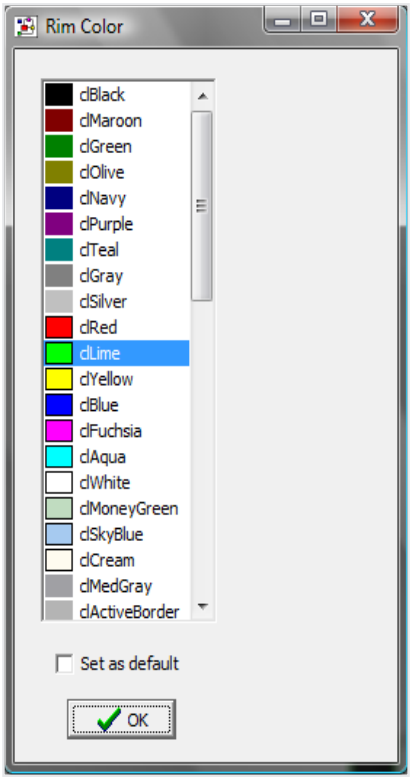


Figure 2.37 Selecting a rim colour to apply to seeds

Exporting a graph

To export a graph, click File|Save diagram as| then select the preferred picture format (Metafile, Bitmap, JPEG), all of which will produce the same image.

Saving your workspace

To save your workspace (the network and attribute data) so you can return to work with it at a later time without again importing the network and attribute data, select "VNA complete *". Note that if you have deleted nodes they will not be saved.

Surveillance is the systematic, regular collection of information on the occurrence, distribution and trends of a specific infection, disease or other health-related event. HIV surveillance is designed to collect and integrate data reported from a variety of sources, including behavioural surveillance, case reporting, seroprevalence surveillance, and sexually transmitted infections surveillance. The goals of second generation HIV surveillance are to help countries better understand the HIV epidemic trends over time, to better understand the behaviours driving the epidemic, to focus on subpopulations at highest risk for infection and to better use surveillance data for planning the response to the epidemic.

HIV surveillance in the Eastern Mediterranean Region needs to be strengthened in order to fill the gaps in our understanding of the dynamics of the epidemic and to be in a better situation to plan appropriately for an effective response. This training course is part of a series of 4 training modules and has been adapted to the regional context from a module originally developed by the Centers for Disease Control and Prevention, USA. It describes how to plan and implement a respondent-driven sampling (RDS) survey. The supplement provides guidance on the use of the software RDS Analyst and NetDraw for analysis of RDS surveys. The course is intended for university and ministry of health staff, public and private public health researchers, and surveillance officers, who will be involved in the planning, organizing, monitoring or implementing of biological and/or behavioural surveillance surveys on HIV/AIDS and associated risk factors in key populations at higher risk of HIV exposure. Countries are free to further adapt these modules or to translate them into local use.