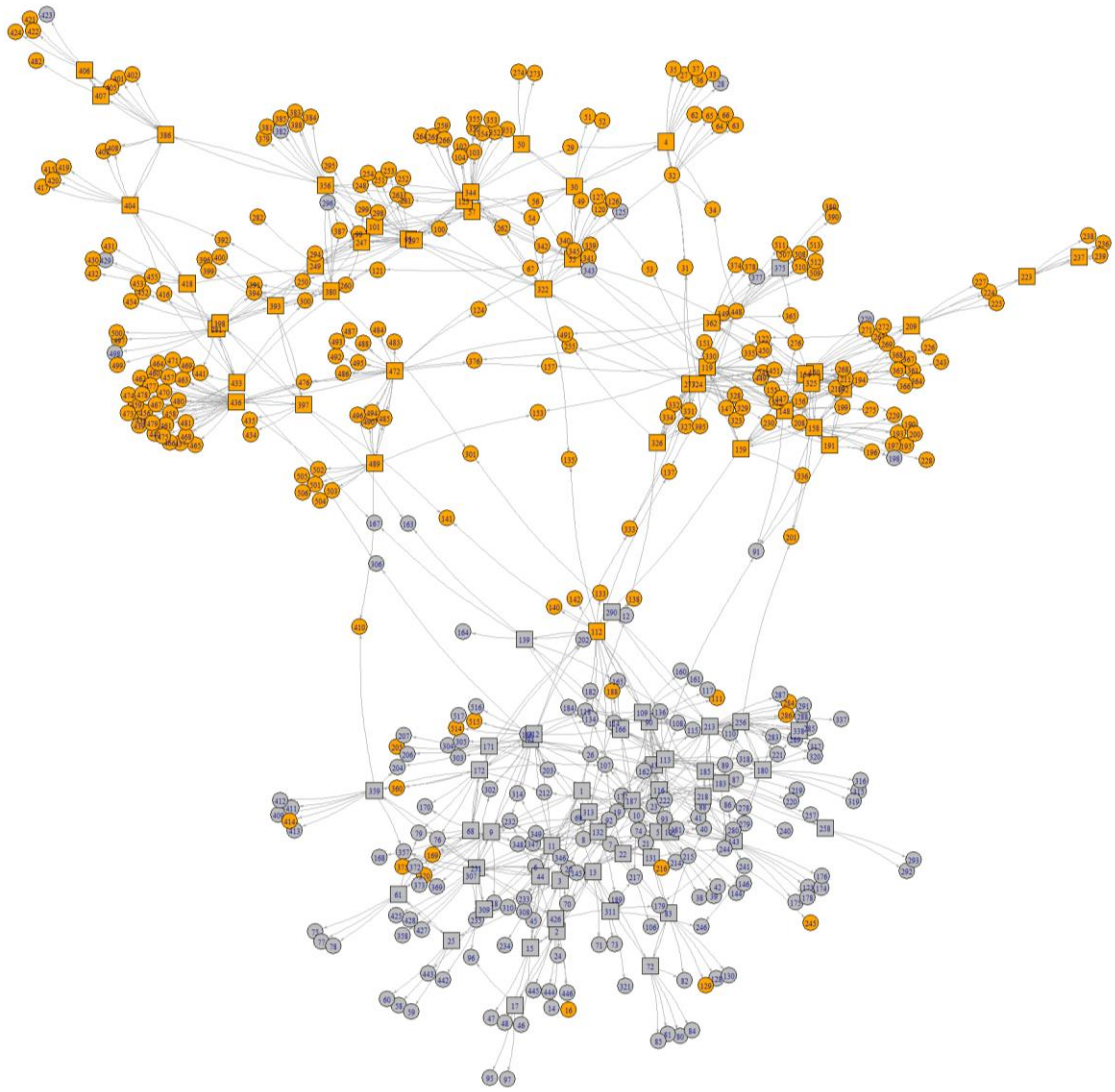
Network Sampling with Memory (NSM) Handbook and Training Manual



Ted Mouw, Giovanna Merli, Ashton Verdery.

Version 1
This version: 6-14-2017

Note: to download the current version of this handbook and to download the programs and data discussed here, go to http://www.tedmouw.info/nsm/nsm.htm

Table of Contents

1. Introduction and Overview

Introduction

Network Sampling with Memory (NSM) is a method for sampling and making inferences about a study population that uses a link-tracing approach where researchers identify and recruit new cases into the sample based on referrals made by current respondents to population members among their social network affiliates. As such, NSM is similar to other better known link-tracing sampling methods like snowball sampling or Respondent-Driven Sampling (RDS), but it has some important differences that make it perform more effectively, yielding sample statistics that tend to be less biased and more precise. The key difference between NSM and other link-tracing sampling designs is that researchers using NSM ask respondents to enumerate aspects of their social networks and then use this network data to reveal the network and to make the sampling process over the network more efficient. For example, the picture on the cover page of this handbook is an example of a sample drawn using NSM (using simulated sampling on a known network, as described below). The squares represent people who have been sampled and interviewed, and the circles are members of the network who have been nominated by respondents but have not actually been interviewed themselves. The lines between the members of the network represent nominations of friends and contacts from the network roster on the survey. Other link-tracing approaches, by contrast, are "flying blind" and do not typically collect information on the circles in this picture. More importantly, other link tracing approaches do not use this information to adapt the sample and improve its efficiency.

A paper by Mouw and Verdery (2012)[1] proposes NSM and provides details on how the collection of network data can be used to improve the precision of samples that are collected from hidden or rare populations. In the paper, Mouw and Verdery show that in simulated sampling on known networks the design effect (the ratio of the sampling variance to the sampling variance of simple random sampling) is considerably lower for NSM compared to other link-tracing methods such as RDS, a popular sampling approach for hidden and rare populations. Another way of thinking about this finding is that the effective sample size of a given NSM study would be considerably higher than the effective sample size of an RDS study with the same number of participants. The advantage of NSM over these alternative designs for hidden and rare populations is that, by collecting network data as part of the survey (in *addition* to getting referrals to network members), it provides a connection between link-tracing sampling approaches and more conventional probability sampling approaches. In fact, the idea of using the accumulation of information about the underlying population social network to improve the process of sampling is a pretty general concept that falls within a long statistical literature on adaptive sampling, and it is likely that there are multiple ways to do this in practice, in addition to the NSM approach itself.

However, although the idea of using network data to sample from hidden or rare populations is fairly intuitive, the practical challenge of implementing NSM as proposed by Mouw and Verdery (2012) is the bookkeeping aspect of

[1] Ted Mouw and Ashton Verdery. 2012. "Network Sampling with Memory: A Proposal for More Efficient Sampling from Social Networks" *Sociological Methodology*. 42(1):206-256.

keeping track of the sampling process. The goal of this handbook (and the programs and data that come along with it) is threefold:

1. To provide a step by step tutorial on how NSM can be implemented in the field.
2. To show that the details of keeping track of the network data can be done automatically in the background by using the Stata programs discussed here.
3. To allow NSM (and modifications of it) to be tested using simulated sampling, as illustrated by the examples in Section 6.

Overview

The basic concept is NSM is that if you collect network data as part of your survey, then you can use the information about the revealed social network to aid in the sampling process. By "network data" we mean asking whether the respondent knows other people who are members of the study population, and then asking for some minimally identifying information about those people that will make it possible to reconstruct the network of ties between people in the population but maintain the anonymity of those described. For example, if the survey asks for information about the respondent's friends or acquaintances (who are members of the study population) and collects information on the first two letters of their last name and the last four digits of their cell-phone numbers, then this information can be used to see if a person has been nominated by multiple respondents. At the same time, such information is unlikely to be useful for identifying individuals who have not consented to participate in the study.

For example, if respondent A nominates a friend who has the first two letters of the last name and the last four digits of a phone number "WH 5444", and respondent B also nominates a friend with "WH 5444," then we can conclude that these nominations are likely to refer to the same person (acknowledging the remote possibility of multiple people with the same name and phone combination, and the possibility of reporting errors[2]). As a result, when we create the network based on the current interviews that have been collected, we would draw an indirect connection between A and B through person "WH 5444".

This information about people such as "WH 5444" who have been nominated, but not yet interviewed, is useful because it allows us to reconstruct a sample of the population network as it has been revealed by sample members up until that point, and it provides us with a list of people who are members of the study population and eligible to be interviewed.

NSM selects new cases to sample by using information from the revealed social network to select the next person to contact for an interview. It does this by using two different sampling modes:

The Search Mode uses information about the number of times that people in the network have been nominated (listed as a contact of current sample participants) to look for areas of the network that appear to be unexplored and are likely to lead to new parts of the network. The logic behind this is that if there are areas or clusters of the network that have been heavily sampled (for instance, a group of people who all know each other), then there will be lots of ties in the survey among people in this part of the network, which is a sign that the survey should prioritize a different, as of yet unexplored, part of the network.

The search mode is the "exploration" part of the NSM sampling algorithm. It is designed to push the sampling towards the frontiers of the currently revealed network. By using the revealed network—including ties from people who were sampled to people nominated but not sampled—to decide where to sample next, NSM keeps its "eyes open" to try to find unexplored parts of the network using the topography of the revealed network as a guide. In contrast, other link-tracing sampling methods, such as Respondent Driven Sampling (RDS) are more of an "eyes closed" approach because they do

[2] For example, if the last four digits of phone numbers are randomly assigned, then there is a 1 in 10,000 chance that two people whose last names begin with "WH" and whose numbers end in 5444 would be mistaken for one another.

not consider where they are in the network as they move to new respondents. With eyes open, NSM can make adjustments that adapt to the new information about the underlying population social network that each additional sample participant reveals.

In contrast to the search mode, the List Mode is very straightforward. The list mode uses the list of all people who are in the revealed network (either because they were interviewed themselves, or because someone has nominated them on a network roster). It randomly samples from this list to choose the next respondent. As the survey progresses, the "list" used by the List Mode tends to grow, because each newly sampled case can contribute information about their friends and acquaintances in the network who were not previously known to the researchers. Eventually, in a finite population, everyone in the population being sampled will be on the list, and the List Mode will be equivalent to simple random sampling (SRS). Random sampling in the List Mode is conducted with replacement, so previously interviewed cases may be selected, in which case NSM increases the weight given to such people's responses but does not re-interview them.

In practice, the basic steps of an NSM survey are:

(A) During an interview, collect network data as part of the survey, including minimally identifying information on who the respondent knows or associates with in the population, such as the last four digits of a cellphone number, and the first four letters of their first name or last name. The social network data is collected on the "network roster" section of the interview.

(B) After completing an interview, separate the network roster from the rest of the survey data, and input data collected in this roster into Stata.

(C) In Stata, run the program (the "sampling algorithm") that updates the revealed network data and randomly selects the next (nominated) person to sample.

(D) Return to Step A, and interview the person who was selected in Step C unless that person has already been interviewed in which case, increase their weight by one. Repeat this process until the desired sample size has been reached.

3. A description of how NSM works (in more detail, but not as much detail as the paper).

[Add this later, based on the workshop at the training session]

4. How to run NSM: A "Hands-On" Tutorial.

In this section, we will go step by step through how to run NSM using the programs that are provided as part of this handbook. In section 4.1 we go through an example of the output from running NSM in Stata, and then in section 4.2 we describe in detail how to run it on your own. In section 4.2 we show how NSM can be conducted using Excel spreadsheets to enter the relevant data from new interviews and then running the NSM programs in Stata to manage the data, update the network, and select subsequent people to interview.

You can follow along with the examples without actually running the programs, but this tutorial is set up to run an example of how NSM works, using two example networks.

Computer requirements to run the tutorial:
1. To do the "hands on" part of the tutorial, you will need to have Stata 14 installed on your computer.
2. In addition to Stata, you will need to be able to open a Microsoft Excel spreadsheet.
3. To create the network plots, you will need to have R installed on your computer, and figure out the path to execute an Rscript command. For example, "C:\Program Files\R\R-3.3.0\bin\" is the directory path where Rscript.exe might be located on your hard drive.

Glossary
This provides a list of definitions for terms that are used.

Populations requiring link tracing sampling designs:
NSM is designed for populations that cannot be surveyed with traditional designs. Generally, there are two groups of such individuals.

Hidden populations are those whose members are difficult for researchers to identify because population membership is delineated on the basis of engagement in stigmatized or illicit activities. A defining feature of hidden populations is that they either entirely lack or do not have a complete sampling frame, which prevents selection of sampling units with known probability of selection as per conventional probability sampling approaches.

Rare populations are those whose members may be so rare that finding them in the general population through screening methods is expensive and inefficient.

Social Network Terms:
Node—in a social network, this is a point on the network (i.e., a person).
Edge—a tie between two nodes. For example, the tie or connection between two people who know each other is an "edge".
Seed – an initial respondent in the study population who is selected (potentially non-randomly) before the actual sampling begins. Ideally, the initial seeds should be as socially distant as possible in the social network of the study population (for example, two people who are friends should not be both selected as seeds). If, for example, you were aware of structural divisions or clusters in the network of the study population (this could be the case of Chinese immigrants of different provincial origin), then it would be a good idea to select initial seeds from different clusters.
Ego – In a network tie (i.e., an edge), this refers to the person who is the source of the connection.
Alter- In a network tie, this is the person who is the destination of the connection. For example, if A nominates B, then A is the ego, and B is the alter.

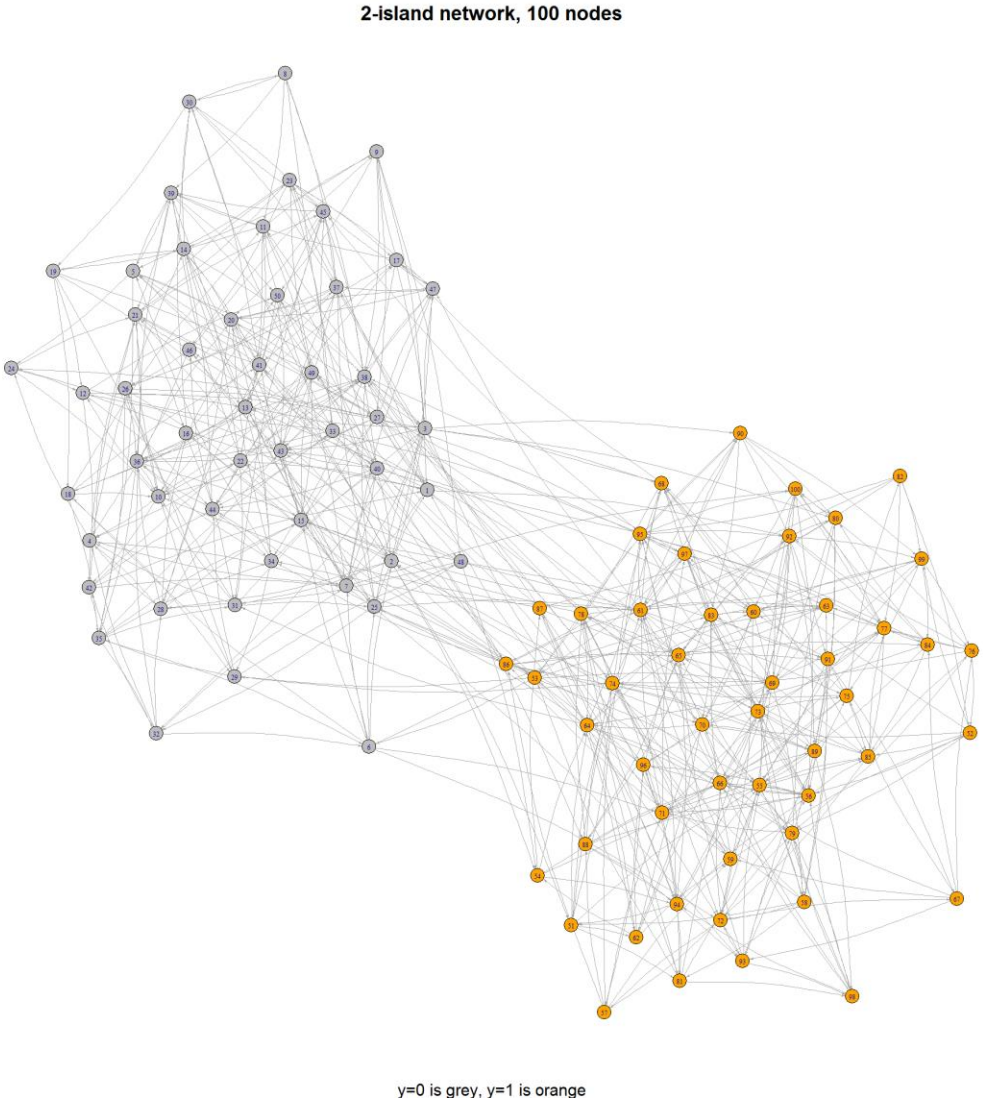
Terms specific to NSM:
Step—in the NSM program, this indicate the current interview number. I.e., the number of "steps" into the sampling process.
Proportion less than 2 nominations (plt2nom) –this is the proportion of nodes in the sampled network that are unsampled and have only been nominated 1 time. This is a measure of how well explored the network is. As this number declines towards 0, it means that most of the nodes in the network have been discovered, because new interviews are only listing people who have already been nominated by other people. The logic here is based on the "capture-recapture" approach used in ecology for estimating the size of a population, that is assessing how many nodes in one sample are also part of a second sample.
Simulated sampling—this refers to running the NSM programs in test mode on a known network. This is useful to make sure that everything is working properly (and to test how well the NSM sampling algorithm works. See Section 6 for more details).

Running the NSM programs in Stata.

There are two test networks provided in this handbook.

The first of these is the "Two-Island Network" presented in Figure 1. This is a simulated network designed to illustrate the problem of clustering.

Figure 1: The Two-Island network



This network consists of 100 nodes, divided into two clusters based on the key dependent variable Y. The grey nodes are cases where Y=0 and the orange nodes are cases where Y=1.

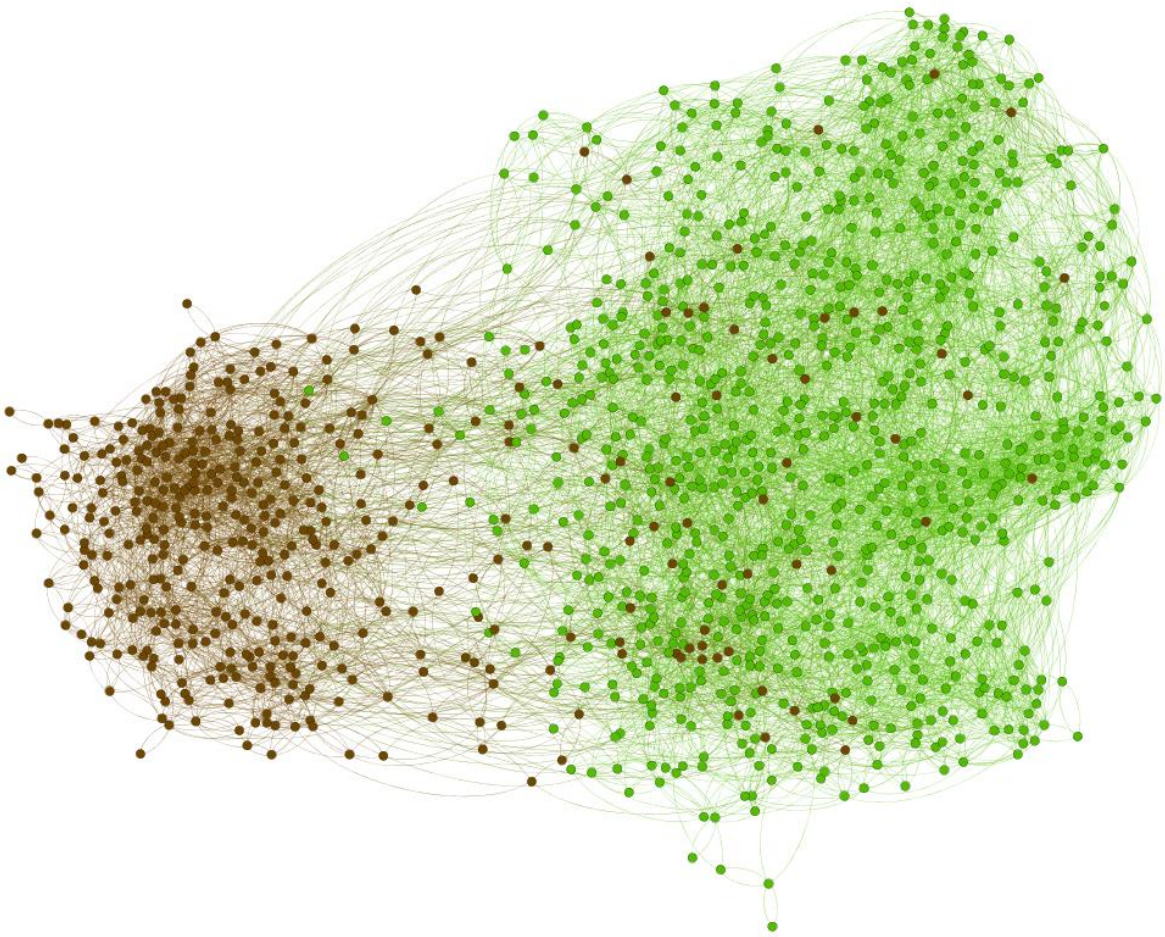
This kind of network, where the nodes are highly clustered and segregated on the basis of a variable such as Y, is difficult for link-tracing approaches to sample, because the sampling process can easily get stuck in one of the clusters unless one randomly selects a bridge node linking the grey cluster to the orange cluster.

Note that in this case, the clustering is based on an observed variable Y. This could be the result of a social process, such as homophily, where individuals have preferences for social interaction with people who share similar attributes to their own. In general, however, the clustering illustrated in Figure 1 could be due to variables that are not observed (or collected) by the researchers. For example, clustering might exist on respondents' propensity to engage in risky behaviors that are not measured by the researcher. Hence, using observed data on Y to move to different parts of the network may not be sufficient to sample the network efficiently. NSM, in contrast, relies on information about the structure of the network itself—i.e. the clustering observed in Figure 1—to try to move to different parts of the network, so it is less susceptible to homophily on unobserved characteristics.

Figure 2 shows a plot of the second example network, which simulates the network structure in a highly segregated school-based social network. This network has 1,278 nodes. The green nodes represent white students and the brown nodes represent non-white students. 67% of the students in the school are white. As with the two-island network above,

the partial name and phone information in these data used for the “hands on” sampling are fictitious and have been randomly generated.

Figure 2: Simulated school network. ("field_test_data")



4.1. An example of running NSM in "test mode".

If you run the NSM programs in test mode, they will sample from one of these two test networks, which will simulate the process of actually using NSM to sample in a real survey.

In particular, if line 18 of the Stata do file "test_nsm_4seeds.do" is:
`global testnetwork=1`
(this is the default), then NSM will run in "test mode" using simulated sampling on the Two Islands Network, pictured above in Figure 1. If you run this do file with `testnetwork=1`, then you will get output similar (but not identical) to the output described in this section.

The details of how to run the NSM programs in Stata are discussed below in Section 4.2.

In this section (4.1), we will explain what happens when the NSM programs run, and then, in section 4.2 we will show how to conduct a "simulated interview" to see how one would use an Excel file to enter data from an actual interview into the NSM process. Here, we will show portions of the log file "test-first example.log" which is the result of running the NSM programs on the Two Islands Network.

There are four main parts to the output from running NSM in Stata.
(A). The record of the interview and updating the network.
(B). Selecting additional cases to sample using the NSM sampling algorithm.
(C). A list of the "data log", which provides statistics of the sampled network and the survey process.
(D). Graphs of the network at various stages in the sampling process.

Partway down the log file (which you can open up in the Stata browser, or in any text editor [we recommend EditPad Lite 7 as a good free program to view log and text files in].

4.1.A. The record of the interview and updating the network

```
interview #  1
this is the person who is being interviewed

     +------------------------------+
     | id   phone   first   intvnum |
     |------------------------------|
  1. |  1     113       n         0 |
     +------------------------------+

these are his/her contacts

     +-----------------------------------------------+
     | zid   y   phone   first   interv~w   intvnum |
     |-----------------------------------------------|
  1. |  90   1    4348       f          1          1 |
  2. |  32   0    8895       c          1          1 |
  3. |  13   0    7184       l          1          1 |
  4. |  47   0    3179       s          1          1 |
  5. |  68   1    9753       a          1          1 |
     |-----------------------------------------------|
  6. |  25   0    2941       b          1          1 |
  7. |   7   0    2380       x          1          1 |
  8. |  43   0    4412       s          1          1 |
  9. |  80   1     158       c          1          1 |
     +-----------------------------------------------+
```

"id" is the id number of this respondent. It is created in the NSM program.

"phone" and "first" are examples of partially identifying information collected on the survey. See section 4.3 for details on modifying these.

"intvnum" is the interview # that this person was first encountered in the survey. 0 indicates that this person was a "seed".

These are people that id 1 listed on his/her network roster.

"zid" is a variable on the Two Islands network that indicates the ID on the full network…this wouldn't be available in a real interview.

"y" is the value of the key dependent variable.

After recording the interview, the NSM program updates the sampled network data.

```
these edges will be added to the network
```

	e_id	e_phone	e_first	e_intv~m	e_y	f_id	f_phone	f_first	f_intv~m	f_y
1.	1	113	n	0	0	5	158	c	1	1
2.	1	113	n	0	0	6	2380	x	1	0
3.	1	113	n	0	0	7	2941	b	1	0
4.	1	113	n	0	0	8	3179	s	1	0
5.	1	113	n	0	0	9	4348	f	1	1
6.	1	113	n	0	0	10	4412	s	1	0
7.	1	113	n	0	0	11	7184	l	1	0
8.	1	113	n	0	0	12	8895	c	1	0
9.	1	113	n	0	0	13	9753	a	1	1

This is a list of new “edges” that will be added to the sampled network data. The prefix “e_” indicates a variable for an ego and the “f_” prefix indicates a variable for an alter.

For example, in the first row, ego has id 1 (this is the respondent in the first interview), and the alter has id 5. All of the alters in this list have f_intvnum equal to 1, because they were first encountered in the data during the first interview.

The ids for the alters have been updated by the NSM programs based on the partial identifying data (phone and first initial).

The subsequent “interviews” follow a similar pattern. For example, The second interview is:
interview # 2

	id	phone	first	intvnum
1.	2	2166	h	0

```
these are his/her contacts
```

	zid	y	phone	first	interv~w	intvnum
1.	27	0	1099	v	2	2
2.	7	0	2380	x	2	2
3.	92	1	9828	y	2	2
4.	11	0	7304	h	2	2
5.	38	0	978	z	2	2
6.	46	0	6852	a	2	2
7.	74	1	1225	s	2	2
8.	55	1	5144	b	2	2
9.	12	0	4831	w	2	2

4.1.B. Selecting additional cases to sample using the NSM sampling algorithm

The simulated survey starts with 4 seeds (these are ids 1-4). Once these four initial seeds have been interviewed, the NSM sampling algorithm randomly selects the next person to interview. The NSM sampling algorithm starts out in the Search Mode, and then switches to the List Mode once the network has been explored.

Here is a description of the output in the log file associated with the Search Mode (this begins after interview 4 in "test-first example.log).

```
Sampling to add a new node to the sampling queue

Number of interviews so far (sampling with replacement): 4
current step number : 4
Current size of the network: 30
current plt2nom : .7666666666666667


method for next selection (1=search mode, 2=list mode): 1

selected node 2

selected node 21

selected node, adding to sampling queue 21

     +--------------------+
     | f_phone   f_first |
     |-------------------|
  1. |    9828         y |
     +--------------------+
listing current sampling queue

     +-------------------------------------------------------------------------------------------+
     | intvnum   phone   first   zid   y   contac~o   refused   sampled   id   method   queuepos |
     |-------------------------------------------------------------------------------------------|
  1. |       2    9828       y    92   1          1         0         0   21        1           6 |
  2. |       2    5144       b    55   1          1         0         0   17        2           5 |
  6. |       0    5815       c     2   0          1         0         .    4        0           4 |
     +-------------------------------------------------------------------------------------------+
```

The current size of the network is the number of nodes (sampled and nominated people) who are part of the data.

The variable "plt2nom" is the proportion of nodes that have received fewer than two nominations and haven't been sampled. A high number indicates that much of the network is still unexplored.

This is the person that was just selected by the Search Mode to be interviewed. He/she is placed in the sampling queue (see the next list)

This is the current sampling queue, which is the list of people who are scheduled to be interviewed. The final variable "queuepos" indicates the position on the sampling queue. The bottom one (queuepos = 4) has just been sampled in interview #4. The next person to be sampled (for the 5th interview) has queuepos = 5.

4.1.C. A list of the "data log", which provides statistics of the sampled network and the survey process.

At the end of updating the data for each interview, the NSM program provides a list of statistics about the sample. This is called the "data log".

We'll jump ahead in the log file to near the bottom to look at the data log after the 30th interview.

This is the data log after step (interview) 30:

	interv~w	netsize	plt2nom	est_pop	ymean	search~e	evensa~g
1.	1	10	.9	99.99998	0	.	0
2.	2	19	.8421053	120.3333	0	.	0
3.	3	24	.8333333	144	0	.	0
4.	4	30	.7666667	128.5714	0	.	0
5.	5	35	.7428572	136.1111	.2	0	0
6.	6	39	.7179487	138.2727	.3333333	.5	0
7.	7	45	.7555556	184.0909	.4285714	.6666667	0
8.	8	49	.7346939	184.6923	.5	.75	0
9.	9	51	.6666667	153	.5555556	.8	0
10.	10	54	.5925926	132.5455	.6	.8333333	0
11.	11	54	.5	108	.6363636	.8571429	0
12.	12	57	.4912281	112.0345	.5833333	.875	0
13.	13	59	.4915254	116.0333	.6153846	.8888889	0
14.	14	59	.440678	105.4848	.6428571	.9	0
15.	15	60	.4333333	105.8824	.6666667	.9090909	0
16.	16	63	.4126984	107.2703	.625	.9166667	0
17.	17	66	.4242424	114.6316	.5882353	.9230769	0
18.	18	67	.4179105	115.1026	.6111111	.9285714	0
19.	19	68	.4117647	115.6	.6315789	.9333333	0
20.	20	71	.4225352	122.9512	.6	.9375	0
21.	21	71	.4084507	120.0238	.5714286	.9411765	0
22.	22	75	.4266667	130.8139	.5909091	.9444444	0
23.	23	75	.3866667	122.2826	.6086956	.9473684	0
24.	24	76	.3552631	117.8775	.5833333	.95	0
25.	25	76	.3421053	115.52	.6	.952381	0
26.	26	76	.3157895	111.0769	.6153846	.9545454	0
27.	27	77	.2727273	105.875	.5925926	.9565217	0
28.	28	79	.278481	109.4912	.5714286	.9583333	0
29.	29	80	.275	110.3448	.5862069	.96	0
30.	30	81	.2469136	107.5574	.5666667	.9615384	0

The variables in the data log are:
interview: the number of interviews completed.
netsize : the number of people in the network (both sampled and nominated)
plt2nom : the proportion of people in the network unsampled and nominated only once.
est_pop : the estimated size of the population. This is calculated as netsize/(1-plt2nom). See Mouw and Verdery (2012) for more details on this calculation.
ymean : the mean of the key 0/1 dependent variable in the survey.
search_mode : the proportion on new respondents (after the initial seeds) that were selected using the search mode.
evensampling : whether "even sampling" has been turned on. In the list mode, this ensures that people who have recently been added to the sampling list are sampled at the same cumulative rate as people who were added earlier.

For example, after the 30th interview, 81 people are in the current network, the estimated size of the actual population is 107.56, and the current mean of Y is 0.567.

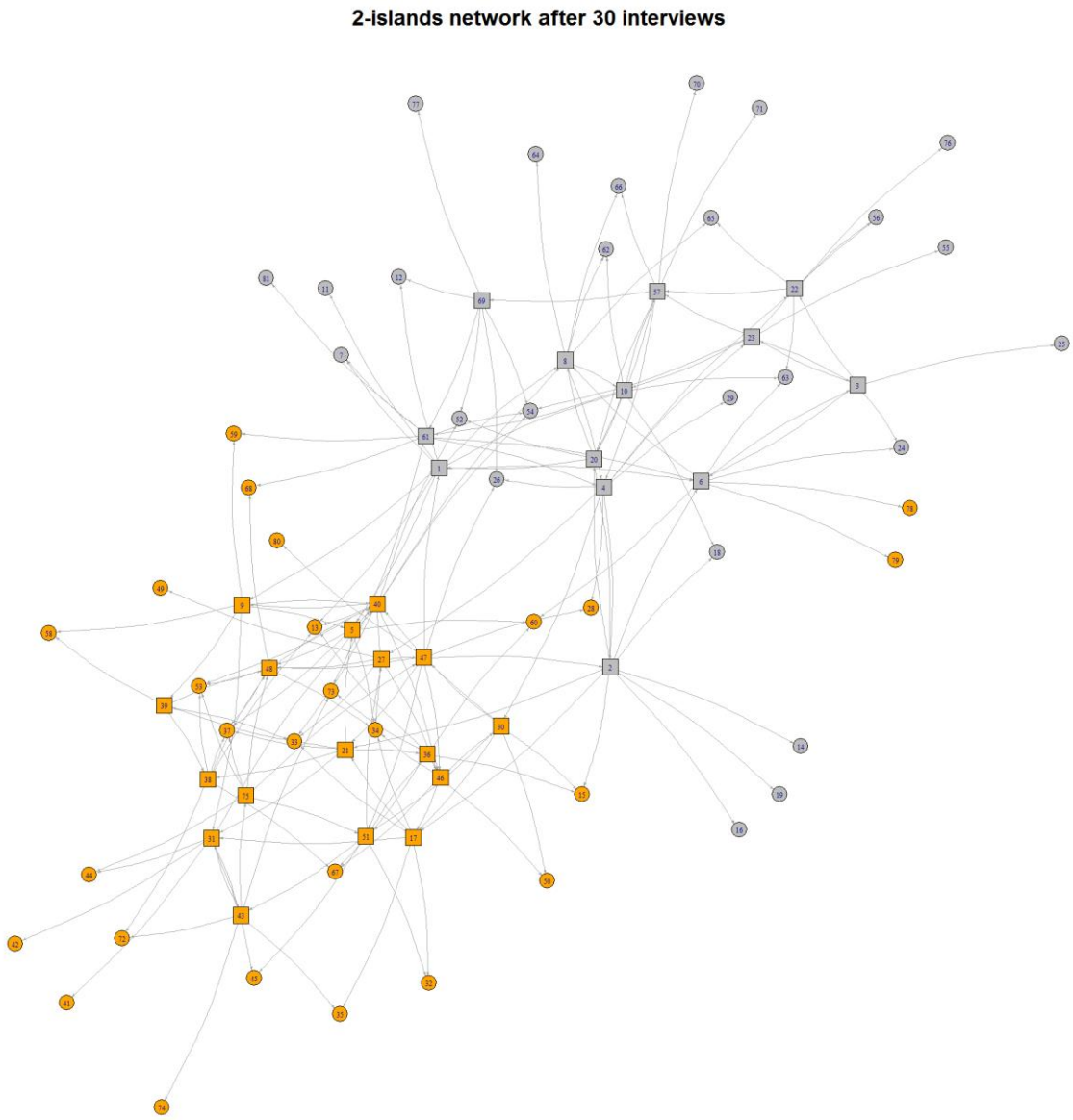
D. Graphs of the network at various stages in the sampling process.

The NSM programs create graphs of the network at regular intervals during the sampling process. These graphs are created using the igraph command in R, which is run from within Stata. These graphs are saved as *.png files (although this option can be changed by modifying the R script file that creates the graphs). Section 4.3 provides more details on this process. For these graph to be created, you must have R installed on your computer in addition to Stata.

Here are the snapshots of the sampled network after 10 and 30 interviews (you can compare these to the graph of the full network in Figure 1 above):

Figure 3:

Figure 4:



square = sampled node, circle= nominated, y=0 is grey, y=1 is orange

Section 4.2: How to run the NSM programs.

This section illustrates how to run the NSM programs and provides an example of entering data after an interview.

4.2.1 Open and run the do file "test_nsm_4seeds.do".

We will start by running the first 30 interviews using simulated sampling on the Two Islands network, then we will enter the data for the 31st interview.

First, open Stata. Make sure that the current working directory for Stata is the same directory where you have put the NSM programs and example data.

To see what the current Stata working directory is, type "pwd" at the command prompt. To change the working directory, you can type "cd c:\mydirectory" where c:\mydirectory is the name of the directory you want to move to. Make sure that you are in the directory where the NSM programs and data are by typing "dir" at the Stata prompt. You should see the do files there.

Next, open the Stata do file editor. To do this, click on the "window" tab, then "do file editor", the "new do file".

Then, open the do file "test_nsm_4seeds.do".

Make sure that the 18th line of the do file has "global testnetwork=1". In addition, the second to last line of the do file should read "nsm_test 30". This will run the simulated sampling for 30 interviews.

Then, in the do file editor, go to the "Tools" tab, and click "execute (do)" → this will run the do file.

After the do file has completed, you will see the log file, "test.log' which will have output similar (but not identical) to the output discussed above in section 4.1.

4.2.2 The Excel files for the sampling program.
In the working directory where you ran the "test_nsm_4seeds.do" file, you will now see several Excel spreadsheets.

These are:

Name	Contents
interview_list.xls	This is a list of the interviews that have been recorded.
interviewsheet.xls	This is the raw data from the interviews. In Section 4.2.3 we show how you add data to this spreadsheet.
nodesheet.xls	This is a list of the nodes (people) in the data, either because they were interviewed or nominated on a network roster.
test_log.xls	This is a spreadsheet version of the data_log discussed above in 4.1.C. The variables on this spreadsheet are the same as in 4.1.C.
nsm_queue.xls	This is a list of people who are in the queue to be interviewed.

Important note: if you open these Excel files, make sure you close them before running the NSM program. The program will load these files into Stata, and, when it is done, it will write the new ones back to Excel. If they are open in Excel, you will get a sharing violation error (at least in Windows) and the program will stop in Stata.

Click on “interview_list.xls”
The **interview_list.xls** spreadsheet keeps track of the interviews that have been recorded. Here is a look at the first 6 from our test run.

intvnum	id	Phone	first	zid	Y	method	interview
0	1	113	n	3	0	0	1
0	2	2166	h	1	0	0	2
0	3	3124	v	4	0	0	3
0	4	5815	c	2	0	0	4
2	17	5144	b	55	1	2	5
2	21	9828	y	92	1	1	6

The variables are:
intvnum : the interview number this person was first added to the data. “0” indicates an initial seed.
id : the person’s id number (this may change each time the simulated sampling is run).
phone and first : these are the partially identifying variables. These can be changed, as described below.
zid : this is the id in the full network (not the sample). It is a fixed ID that identifies people in the full network, and it is specific to the Two Islands network.
y : the key dependent variable of the respondent.
method : the sampling mode used to select this person to interview. “0” indicates an initial seed. “1” indicates the Search Mode and “2” indicates the List Mode.
interview : the interview number.

Next, click on “interviewsheet.xls”

The **interviewsheet.xls** spreadsheet is the raw data from the interviews. Here is a look at the first two interviews from our test run:

interview	role	intvnum	phone	first	y	zid	id	contact_info
1	1	0	113	n	0	3	1	1
1	2	1	7184	l	0	13		1
1	2	1	3179	s	0	47		1
1	2	1	158	c	1	80		1
1	2	1	8895	c	0	32		1
1	2	1	4348	f	1	90		1
1	2	1	4412	s	0	43		1
1	2	1	2941	b	0	25		1
1	2	1	9753	a	1	68		1
1	2	1	2380	x	0	7		1
2	1	0	2166	h	0	1	2	1
2	2	2	1225	s	1	74		1
2	2	2	1099	v	0	27		1
2	2	2	6852	a	0	46		1
2	2	2	9828	y	1	92		1
2	2	2	7304	h	0	11		1
2	2	2	5144	b	1	55		1
2	2	2	978	z	0	38		1
2	2	2	4831	w	0	12		1
2	2	2	2380	x	0	7		1

The variables are:

interview : the interview number.

role : 1 is the respondent, and 2 indicates an alter (someone who was nominated by the respondent).

intvnum : the interview the person was first encountered. For alters, this is the current interview. This variable is used to differentiate specific nominations in the NSM programs.

phone, first : partially identifying variables. These can be different depending upon the survey instrument.

y : the key dependent variable.

id : the current ID for the person. This is blank for alters because the IDs are assigned after the raw data for the interviews are read in.

contact_info : this indicates whether or not the person has contact information (so that they can be contacted about participating in the survey). “1” indicates yes, and “0” indicates no.

Open the file "**nodesheet.xls**". This file keeps track of information about the nodes (i.e. people) in the network.

id	interview	intvnum	phone	first	s	zid	y	contact_info	refused	role	sampled	queuepos
1	0	0	113	n	1	3	0	1	0	1	1	1
1	1	0	113	n	1	3	0	1	0	1	1	1
1	21	21	113	n	1	3	0	1	0	2	1	1
1	29	29	113	n	1	3	0	1	0	2	1	1
2	0	0	2166	h	1	1	0	1	0	1	1	2
2	2	0	2166	h	1	1	0	1	0	1	1	2
2	4	4	2166	h	1	1	0	1	0	2	1	2
2	21	21	2166	h	1	1	0	1	0	2	1	2
2	29	29	2166	h	1	1	0	1	0	2	1	2

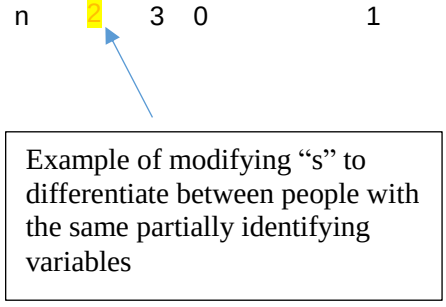
Most of the variables on this sheet are described above in one of the previous two sheets. The new ones are:
s : this is a variable that allows you to differentiate among people with identical phone/first information. This process is described below.
refused : indicates someone who refused an interview request. They will not be selected to be interviewed again.
sampled : "1" indicates the person has been sampled. "0" indicates that they have not.
queuepos: if the person has been, or is, in the queue to be interviewed, it will be listed in this variable.

Every time a person appears on the interview sheet they also appear on the nodesheet. For example, id 1 is on the sheet four times, because they appeared in the initial seeds, then as the respondent in the first interview, and as an alter in interviews 21 and 29.

The reason that the nodesheet stores the data on individuals that way is that it makes it possible to keep people separate who have the same partially identifying information. This is done by modifying the variable "s" on the nodesheet.

For example, if additional information that you collected in interview 29 led to the conclusion that the current ID 1 in row 4 above was different then the ID 1's in rows 1-3, then you would modify the variable s for this row from 1 to 2.

id	interview	intvnum	phone	first	s	zid	y	contact_info	refused	role	sampled	queuepos
1	0	0	113	n	1	3	0	1	0	1	1	1
1	1	0	113	n	1	3	0	1	0	1	1	1
1	21	21	113	n	1	3	0	1	0	2	1	1
1	29	29	113	n	2	3	0	1	0	2	1	1



Now, the network can be updated and row 4 will be assigned a different ID than rows 1-3. The NSM program checks and updates the ID variable every time the nodesheet file is updated manually. (This is because the program loads the data on the nodes from the Excel spreadsheet and then recreates the ID variable).

In general, the possibility of multiple people with the same partially identifying information will depend upon what type of information you collect on the survey. It is possible to check for this manually if the size of your sample is small. We

have written a program that does this for data we collected in Mexico.[3]For data we collected in Tanzania[4], we collected last four digit of cell phone in addition to first 3 digits of last name and 1st initial. We were pretty confident that this combination was unique.

Next, click on the file "**nsm_queue.xls**". This file shows a list of everyone who has been in the queue to be interviewed for a particular interview. We will look at the top of this file to see who is in line to be interviewed (note that this will look different in your data because it will be a new sample):

intvnum	phone	first	zid	y	contact_info	refused	sampled	id	method	queuepos
30	7349	v	18	0	1	0	0	81	1	32
24	8574	f	16	0	1	0	0	56	1	31
27	1462	q	6	0	1	0	0	61	1	30
26	4723	h	64	1	1	0	0	47	1	29

ID 56 is the next case to interview (for interview 31)

This indicates that ID 56 is the next case to interview (for interview 31), and ID 81 will be interview 32.

4.2.3.1: What to do if someone declines to do an interview? First, delete them from the queue by deleting the row of the Excel sheet that they are on. Then, in the nodesheet.xls, change the variable "refused" from 0 to 1 indicating that they have declined to participate. This will then update the network data automatically and they will not be selected for future interviews.

Section 4.3 How to add data from an interview.

After running the do file "test_nsm_4seeds.do" you will have 30 simulated interviews, which represents an initial exploration of the network. Click on the network picture ("2-islands network_30.png") to get a visual sense of where you are in the sample.

Now, we want to practice adding data to the interviewsheet.xls that we discussed above in Section 4.2. Although this is just a simulated, or "pretend", survey and interview, it will be the same as the process of adding data from a real interview to the NSM sample.

To conduct the 31st interview, type the following at the Stata prompt:

```
practice_interview
```

The program will then select the next person in the sampling queue (see the description of nsm_queue.xls above).

You will then see output showing the results of the interview (i.e., the network roster). Here is the output for the example run:

[3] Ted Mouw, Sergio Chavez, Heather Edelblute, and Ashton Verdery. 2014. "Binational Social Networks and Assimilation: A Test of the Importance of Transnationalism" *Social Problems*. 61(3): 1-31.

[4] Merli, M. Giovanna, et al. "Sampling migrants from their social networks: The demography and social organization of Chinese migrants in Dar es Salaam, Tanzania." Migration studies 4.2 (2016): 182-214.

```
. practice_interview
this program simulates an interview
use the interviewsheet in excel to input the data from the interview
see the example in the training manual
this program is similar to the automatic interviews in the test mode
except that it doesn't automatically update the data after the interview

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        step |         30        15.5    8.803408          1         30
interview #  31
this is the person who is being interviewed

     +------------------------------+
     | id   phone   first   intvnum |
     |------------------------------|
  1. | 56    8574       f        24 |
     +------------------------------+

these are his/her contacts

     +-----------------------------------------------+
     | zid   y   phone   first   interv~w   intvnum |
     |-----------------------------------------------|
  1. |  25   0    2941       b         31        31 |
  2. |  30   0     309       b         31        31 |
  3. |  31   0     171       w         31        31 |
  4. |  50   0    3247       v         31        31 |
  5. |  10   0    8530       a         31        31 |
     |-----------------------------------------------|
  6. |  37   0    2777       t         31        31 |
     +-----------------------------------------------+

interview data in the form it should be entered into the interviewsheet
note if multiple interviews have been completed, update the interview number sequentially
on the interviewsheet

    interv~w   role   intvnum   phone   first   y   zid   id   contac~o
          31      1        24    8574       f   0    16   56          1
          31      2        31    2777       t   0    37    .          1
          31      2        31    8530       a   0    10    .          1
          31      2        31    3247       v   0    50    .          1
          31      2        31     171       w   0    31    .          1
          31      2        31     309       b   0    30    .          1
          31      2        31    2941       b   0    25    .          1
file practice_interview.xls saved
```

The results of the interview have been saved in “practice_interview.xls”. Open this in Excel, and then paste the results into the interviewsheet.

For example, here are the contents of practice_interview.xls:

Interview	role	intvnum	phone	first	y	zid	id	contact_info
31	1	24	8574	f	0	16	56	1
31	2	31	2777	t	0	37		1
31	2	31	8530	a	0	10		1
31	2	31	3247	v	0	50		1
31	2	31	171	w	0	31		1
31	2	31	309	b	0	30		1
31	2	31	2941	b	0	25		1

The order of the variables is in the same order as the variables in interviewsheet.xls.

After pasting it into the bottom of interviewsheet.xls, the bottom of the sheet (interviews 30 and 31) looks like this:

30	1	27	1462	q	0	6	61	1
30	2	30	5815	c	0	2		1
30	2	30	7753	f	1	83		1
30	2	30	6147	d	1	71		1
30	2	30	2380	x	0	7		1
30	2	30	2980	x	1	72		1
30	2	30	2941	b	0	25		1
30	2	30	7349	v	0	18		1
30	2	30	9091	j	0	15		1
30	2	30	2941	b	0	25		1
31	1	24	8574	f	0	16	56	1
31	2	31	2777	t	0	37		1
31	2	31	8530	a	0	10		1
31	2	31	3247	v	0	50		1
31	2	31	171	w	0	31		1
31	2	31	309	b	0	30		1
31	2	31	2941	b	0	25		1

Interview 30

Interview 31

(Note that the top of the sheet is shown above in Section 4.2).

Now, save this file, and close it.

Then, at the Stata prompt, type:

```
nsm_test 1
```

The number after nsm_test indicates the number of "steps" (i.e., interviews) to process. Running this command will process the interview that you just added, and update the node and network data. In addition, if there are fewer than two people in the sampling queue, the sampling algorithm will add an additional person to the sampling queue.

Now, you could do another practice interview and repeat this process, which would be interview #32, by again typing

```
practice_interview
```

at the Stata prompt.

Then, paste the results of the interview into the interviewsheet, save it, and type

```
nsm_test 1
```

again at the Stata prompt.

However, if instead of doing another practice interview you type

```
nsm_test X
```

where X is a positive integer, then the program will run the next X interviews. Each time you run this command, the program first checks the interviewsheet to see if more interviews have been added manually. When the program is in test mode (meaning it is running on a known network, rather than on an unknown network in a real survey) it continue with automatic interviewing if no new interviews have been entered manually. For example, typing nsm_test 29, gets us to the 60th interview of the survey, and resulted in the following picture of the sampled network:

Figure 5

After 60 interviews in this practice run, 97 people are in the sampled network, and the mean of Y in the sample is 0.5.

Section 4.4: Details of the NSM start up do file.

The Stata do file "test_nsm_4seeds.do" sets the basic variables and parameters needed to run the NSM programs. In this section, we describe the contents of this file. The do file can be opened in a text editor or in the Stata do file editor, which is explained above in Section 4.2.1

If you run this do file it will load the NSM programs into Stata. See Section 4.2.1 for instructions on running the do file.

4.4.1. Stata macros

There a macro is a parameter that can be accessed to run a program. In Stata, a "global macro" is set by a command such as line 18 of the do-file:

```
global testnetwork=1
```

which sets the macro for testnetwork to 1.

Global macros are referenced by using $ + the name of the macro, or ${+the name of the macro}.

For example, on line 19 of the do file,
`${testnetwork}` refers to the contents of the global testnetwork, which is set to 1 in line 18.

4.4.2 Key parts of the "test_nsm_4seeds.do" file

Lines 1-62 of the do file set the system parameters (macros) that you will need to understand to run the NSM programs.

These are explained in detail here:
$testnetwork = 1 uses the Two Islands Network data, = 2 uses the "field test" network data.
$network_data sets the network to use (for simulated sampling). If you are sampling an unknown network in the field, then this global is not used.
$pc differentiates between windows and mac computers.
$backup = 1 sets automatic backups of the main data files on.
$test = 1 runs the programs in "test" mode (which is simulated sampling on a known network).
$rpath is the path to Rscript.exe on your computer. This must be set to run R from Stata to create the network graphs.
$sample_data is the name that Stata will save the current network data to.
$datalog is the name of the data file that keeps track of the overall statistics of the survey.
$ids is a list of partially identifying variables in the data. For example, in the two networks provided with this tutorial they are "phone first". These can be changed to adapt to the information on a different survey.
$fids is the same as $ids, but with "f_" before each of the variables.
$eids is the same as $ids, but with "e_" before each of the variables.
$key_y is the main dependent variable. This is not part of the NSM sampling algorithm, but it can be set to provide information about the characteristics of the sample.
$y is a list of variables, including the $key_y, to keep in the data. These are not part of the NSM sampling algorithm.
$fy and $ey just append "f_" and "e_" in front of these macros.
$evensampling is set to 0 for the start of the survey.

On line 62 of the do file, the command
do nsm_programs
will run the do file nsm_programs.do, which loads the NSM programs into Stata.

Lines 65-74 of the do file reset the NSM data files. These data files are discussed above in Section 4.2.2. Stata moves back and forth between the Stata and Excel version of these files.

Lines 89-94 of the do file will stop the do file if $test is not equal to 1. $test=1 means that NSM will run in test (simulated sampling) mode. Otherwise, if it is not 1, then you are running NSM to conduct an actual survey (on an unknown network), which is what we turn to next in Section 5.

4.5 Graphing the network data during the sample.
In the do file "nsm_programs.do", the Stata program "plotigraph" will use the data from the current sample to graph the network in R.

To do this, you need to have R installed on your computer (it is a free program). You need to set the correct path to the file Rscript.exe in the $rpath global as described above in section 4.4.2.

In addition, you need to install the following R packages:
foreign
igraph
You can install these packages by typing
install.packages("foreign")
install.packages("igraph")
at the R prompt.

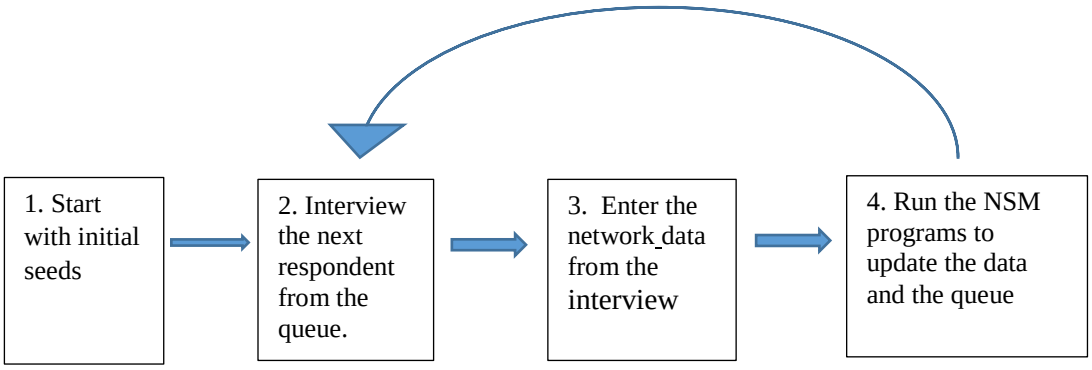
The program plotigraph in Stata will call the R syntax file "nsmgraph.R" which should be placed in the same directory that you put the Stata NSM programs.

4.6 NSM programs file.
The Stata do file nsmprograms.do contains the programs needed to run NSM. This do file is run by nsm_test_4seeds.do as discussed in Section 4.4.2. There are two main programs in this do file:
1. "nsm_after_interview" which processes data from new interviews and updates the network. It is the data management program.
2. "nsm_sample_field" which uses the NSM sampling algorithm to randomly select a new case to sample, based on the current data on the sampled network.

Section 5: Flow chart of steps and starting a new survey.

Section 5.1: In this section we summarize the steps involved in doing an NSM survey, building on the description of the NSM programs in Section 4.

Figure 6: Flow Chart of the Steps in an NSM Sample



Description of the steps

Initial preparation.
Start with a finite study population that is connected by a social network. For an example of a network roster and questions about contact information, see Appendix A.

1. Initial seeds. Start with initial contacts (seeds). The assumption is that these seeds may be non-randomly selected. A general guideline is to start with around 5 initial seeds. Put these seeds into the sampling queue, which is the nsm_queue.xls spreadsheet discussed above in Section 4.2.2.

2. Interview the next respondent. Interview the next person in the sampling queue from the "nsm_queue.xls" spreadsheet.

3. Enter the network data. After completing the interview, enter the network data from the interview into the spreadsheet interviewsheet.xls (see Section 4.2.2 for a description of this spreadsheet, and Section 4.3 for entering the results of the interview. [Note: make sure you save and close the Excel files before running the Stata programs]

4. Run the NSM programs. After entering the data in Step 3, run the NSM program by typing
`nsm_test 1`
at the Stata prompt (see Section 4.3). This will update the network data based on the interviewsheet.xls you just modified in Step 3, and it will add a new case to the sampling queue. Then, return to Step 2 and interview the next person on the sampling queue. [Note: before running this command, make sure you have run the do file "test_nsm_4seeds.do" (or its equivalent) with $test=0 to load the programs into memory. See section 4.4 for details on this do file.

5. How to add more people to the sampling queue.
In addition to running `nsm_test 1`, you can add more people to the sampling queue by typing
`nsm_sample_field`
At the Stata prompt (make sure you have run the do file "nsmprograms.do" to load this program into memory. It is a good idea to have only a few people in the sampling queue at a time, especially at the start of a sample. You should update the information on the network after every interview. Limiting the number of people in the queue until the network has been explored will improve the precision of the samples. See Mouw and Verdery (2012) for more discussion on this.

For example, if you wanted to start a survey of an unknown network with two seeds, the Excel file "nsm_queue.xls" might look like this (see Section 4.2.2 for a description of this spreadsheet):

intvnum	phone	first	y	contact_info	refused	sampled	id	method	queuepos
0	113	n	0	1	0		1	0	1
0	2166	h	0	1	0		2	0	2

Then your first two interviews would be "113 n" and "2166 h" (using the partially identifying information to refer to the respondents). This is all you would need to start with, as the other spreadsheets will be updated as you add information from the interview (to interviewsheet.xls) and run the NSM programs to process the results from the interview.

Section 6: Testing the design effects of NSM using simulated sampling.

The programs in the do files nsm_programs.do and test_nsm_4seeds.do are designed to run NSM for a single sample.

If you want to test NSM by conducting simulated sampling on a known network multiple times to calculate the design effects (for that network), then Section 6 describes the use of a faster version of the NSM programs written in Stata's Mata language. Mata is faster than the regular Stata syntax, but it is harder to debug if something goes wrong, which is why we wrote the NSM programs in the regular Stata syntax as well.

In order to run simulated sampling NSM (and RDS—respondent driven sampling) with 500 replications on the simulated school network data depicted above in Figure 2, see the do file "run_newnetsample2017.do" for more details.

[More details will be added to this section].

Appendix A: Example network roster and referral questions.

This is an example of a sample script used to introduce the network roster questions:

Now, think about Chinese people you know in the area, including those whom you know only a little. You may know many Chinese people in this area. However, we only want to ask a few questions about 6 people that you know. These don't have to be your best friends, just people you talk to or interact with on a regular basis, people whom if you see them on the street, you know their name and they know yours.

Do you think you know 6 Chinese people in the local area? Please only limit your list to people who were born in China, Hong Kong, or Taiwan.

First, let's make a list of people with the first few letters of their names just to keep track of them. For this list of people you know, we would like to start by knowing the first three letters of their last name and the first initial of their first Chinese name and her first English name if they have one. So, if your friend's name was Fan Bingbing (like the famous Chinese actress), it would be: "FAN" for the first three letters of the last name, and "B-B" for the first initial of the first name. If your friend's English name were Paula, then P would be the first initial.

#	Enter first 3 letters of last name	First name initial	What place does this person currently live in	How are you related to this person? 1. Family 2. Friend 3. Coworker 4. Other 5. Classmate	What is this person's gender? 1. Male 2. Female	Roughly what is this person's age? 1. < 25 2. 25-34 3. 35-49 4. 50+	What is this person's highest educational attainment? 1. Primary 2. Secondary 3. Tertiary 4. Don't know	What is this person's occupation?	What province in China was this person born in? (if not China, write country)	How many years have you known this person? (if <1 write months/12)	How often do you talk to this person? 1. Once a year or less 2. 2-11x a year 3. Once a month 4. Once a week+	How close do you feel to this person? 1. Not very 2. Somewhat 3. Very 4. Extremely	Do you ever discuss personal matters with this person?	If you were to contact this person, would you contact him/her by: 1. Text 2. Email 3. Phone 4. In person 5. WeChat 6. FB Msg 7. Other
1														
2														
3														
4														
5														
6														

“As I mentioned before, in order to better understand the Chinese community you are part of, we would like to ask you to help us recruit the friends you have nominated.

All YOU need to do is to forward an invitation text or email to each of these friends which describes the survey and asks them if they would like to learn more about our survey. All YOUR FRIENDS have to do is to send us a text or call us saying that they would be willing to learn more about our survey and we will give them a €10 compensation in Amazon electronic gift cards just for wanting to learn more about the study.

We have found that referrals feel more comfortable if they are contacted both by the respondent and by the survey team. So we are also asking you for their contact info (email or cell number) so that we can follow up with them to verify that this is an academic, non-governmental, non-commercial study where their privacy and anonymity as a respondent is guaranteed and protected. For your help in providing us with the contact information of these six friends we will give you $30 in Amazon gift cards, and you will be helping your friends by providing them with the opportunity to earn up to $50 themselves by participating in our survey. Of course, your friends are completely free to decide whether or not they want to participate on their own, and their privacy as potential respondents is guaranteed by the guidelines of our study.

We depend upon the generous help of previous respondents to refer us to new respondents, and we are truly grateful for your time and assistance in helping us out with this study by recruiting your friends.

Would you be willing to refer us to the contacts you have listed?

INTERVIEWER: ASK FOR CONTACT INFORMATION FOR all contacts provided by the respondent

Interviewer: If R answers Yes, then ask: Can you give us your cell phone number and/or email address so we can send you the invitation text that you can send to your friends? Can you also give us your email address so we can send you the Amazon gift card?