STEW-MAP Network Cleaning Steps
Michelle Johnson, updated July 2020

This guide will show you how to clean the network data you have gathered for the STEW-MAP project. To start, we provide a glossary of network terms referred to in the guide.

Glossary of Network Terms:
Sender: Organization that names another organization as a collaborator.
Receiver (or Alter): Organization named by sender as a collaborator. Alter is another term used to mean receiver.
Edge: A tie between 2 entities. In STEW-MAP, this is a collaborative tie between two organizations.
Edgelist: A spreadsheet of edges between senders and receivers.
Node: A node is an entity included in a network. In STEW-MAP, nodes are organizations and include senders and receivers.
Nodelist: A spreadsheet of nodes and associated attributes. In STEW-MAP, this included survey results for groups that took the survey.
Master List: A spreadsheet of all groups across all networks, typically created by expanding off of the initial sampling frame of groups the survey was sent to.

Overview of Network Cleaning Steps:
The final product of your analysis will be a table called an edgelist, which will list all of the individual connections between the “nodes” in your network (in this case, organizations). In the completed edgelist each row will contain a link between a **sender** (the group that took the survey, or respondent) and a **receiver** (the group named in the survey, also sometimes referred to as an alter). At its most basic, an edgelist will be comprised of 2 columns:
1) unique ID number for sender and 2) unique ID number of receiver / alter.

Including 2 additional columns for 3) sender organization name and 4) receiver organization name can also be useful as reference, especially during the cleaning process.

This *edgelist* can then be linked with a *nodelist*. A node list is a list of all entities within a network, which can also contain other attributes about a group, like organization names and other responses from the STEW-MAP survey.

Additionally, this process will create a *master list*, which is a list of all entities across all networks for a location, not just a single network. For example, in STEW-MAP surveys, there is often a general *collaborate* network, but also *resources* and *knowledge* networks, based on the three types of network questions in the STEW-MAP survey.

Tables 1-3 demonstrate examples for edgelists, nodelists, and master lists:

Table 1. Example Master list for an example location

UniqueID	GroupName	Website	Notes	Network_Collaborate	Network_Knowledge	Network_Resources
121	Generic Community Garden			1	1	0
454	Generic2 Community Garden			1	0	0
101	Neighborhood 1 Street Tree Group			1	1	1

Table 2. Example Edgelist columns for an example network

UniqueID_ Sender	GroupName_Sender	UniqueID_Receiver	GroupName_Receiver	GroupName_Receiver_Std	Website	Notes
121	Generic Community Garden	454	Generic2 Comm. Gdn	Generic2 Community Garden		
121	Generic Community Garden	101	Street Tree Group – Neighborhood 1	Neighborhood 1 Street Tree Group		
454	Generic2 Comm. Gdn	101	N1 St Tree Grp	Neighborhood 1 Street Tree Group		

Table 3. Example Nodelist for example network

UniqueID	GroupName	ContactPerson	ZipCode	PrimarySiteType	PublicMap
101	Neighborhood 1 Street Tree Group	Jane Smith	10001	Street Tree	1
121	Generic Community Garden	Jane Doe	11359	Community Garden	1
454	Generic2 Community Garden	John Smith	10001	Community Garden	1

Below we summarize each of the steps you will take to produce these final products.

Step 1: Adding each receiver group to an individual row in a new spreadsheet.

Depending upon your survey software, your survey responses should be in a spreadsheet, and they may provide responses as one row per responding group (sender group) (see Table 4a) or multiple rows per sender group (Table 4b). Other data structures may also apply.

Table 4a. Examples of unformatted survey responses for network questions (separate columns for each network question, responses separated by commas)

UniqueID_Sender	GroupName_Sender	Collaborate	Knowledge	Resources
121	Generic Community Garden	Generic2 Comm. Gdn, Street Tree Group – Neighborhood 1, 2nd Park Group	2nd Park Group	Generic2 Comm. Gdn
101	N1 St Tree Grp	Generic Community Garden		Citywide Street Tree Group
454	Generic2 Comm. Gdn	N1 St Tree Grp	N1 St Tree Grp	

Table 4b. Examples of unformatted survey responses for network questions (each named group on a separate row, network type formatted by checkboxes)

UniqueID_Sender	GroupName_Sender	PartnerGroup	Collaborate	Knowledge	Resources
121	Generic Community Garden	Generic2 Comm. Gdn	x		x
121	Generic Community Garden	Street Tree Group – Neighborhood 1	x		
121	Generic Community Garden	2nd Park Group	x	x	

To manually code network data, you will want to be able to create edgelists for each network question. You may want to have the network data setup as one spreadsheet tab per network question (e.g., collaborate, resources, knowledge) or bundle the responses for all network questions, making sure to have a column(s) that indicates to which network(s) the individual row applies.

Often, a respondent, or sender, group will name multiple receiver groups. To get to an edgelist that has multiple rows (or collaboration links) per sender group, you will need to format the data. How you complete this step will depend on how the survey responses are formatted in your survey.

If all named groups (e.g., receivers) are present in a single row/cell of data (Table 4a), they will need to be split onto multiple rows:

- For each respondent (or sender), copy the unique ID and group name into 2 columns of a new spreadsheet (e.g., Sender Unique ID and Sender Name).
- For each individual group named by that respondent, add to a third column (e.g., Receiver Name).
- After group names are standardized, a unique ID column will be populated for the receivers (e.g., Receiver Unique ID).
- While doing this process, be sure to track which survey responses have been formatted and which have not.

If all named groups (receiver) are present on separate rows of data in the survey responses (e.g., one respondent has 10 rows of named groups), format the data so you have 4 columns (Table 4b):

- unique ID for sender,
- group name of sender,
- unique ID of receiver (this will be populated after group names are standardized in step 2), and
- group name of receiver.

Ideally, you will clean network data after the survey is completely closed, but with some modifications, you can conduct this cleaning process in an iterative fashion, being careful with version control.

Step 2: a) Standardize group names and b) assign UniqueID (where needed)

Group names identified by responding groups may differ in spelling or abbreviation. A single standardized version is needed for each group name in the network.

Some general suggestions for standardizing names include the following:

- Delete leading or trailing white spaces
- Remove double white spaces
- Change all names to either Title Case, lower case, or UPPER CASE for consistency
- Scan for special characters
- Scan for Arabic numerals that might match to text (ie 6th could match to “sixth”)

You may likely need several passes through the data to fully standardize names.

You can start with your initial sampling frame to make a master list of all groups:

Create a list of existing standardized group names and Unique IDs by combining lists of standardized names and Unique IDs for respondent groups and partner groups that were included in the initial STEW-MAP sampling frame (the list you used to distribute your survey). This will be a master list – and may include more group names than you have in a single network (Table 1).

In the edgelist that you started in Step 1, add a column titled “GroupName_Receiver_Std” (Table 1) next to the group name of the receiver provided in the survey, Use the standardized names you already have in your master list to populate this column. It can be manual or through some sort of formula, filtering, or sorting. For any groups not already on the master list, add them to the master list and create a standardized name and UniqueID.

To confirm the group’s correct name and existence, you can use Internet searches and if a group has a website and/or social media accounts, add those in an additional column (e.g., a Website column).

For entries you can’t confirm are groups through websearches or other research, add a Notes column and document categories of potential issues (e.g., Individual, Can’t Locate Group, Not Enough Information). We define groups as two or more people; if just a single person is named, that receiver would be excluded from the network list. Additionally, there may not be enough information from the response to clearly identify a group. Registered non-profits will be searchable at the IRS and non-profit databases; informal community groups After passing through the edgelist multiple times, you may want to exclude these rows from the final dataset.

Step 3: Add new unique IDs where needed

For groups named in the network (alters) *that are not already part of the initial sampling frame*, assign unique ID numbers to all remaining groups, in both the edgelist and the master list.

Step 4: Compile nodelist for a specific network

From a particular edgelist (e.g., collaboration, knowledge, resources), filter the dataset so you only have unique groups.

In EXCEL, you can stack the sender and receiver columns together and then run the advanced filter in EXCEL, with the unique entries box checked.

In R, you could apply one of the following code snippet to create a nodelist:

- unique(dat$UniqueID_Sender)
- nodelist <- unique(dat$UniqueID_Sender)
- write.cvs(data.frame(unique(dat$UniqueID_Sender)), MyCityYEAR_nodelist.csv, sep=',’), etc.
- distinct() would also work well in this application

This is your nodelist and additional data from the survey or elsewhere can be joined to each group here. For example, you may want to identify the sector of each group (e.g., civic, government).

A nodelist is different from a master list, as it is only for a single network, like collaborate or resources. The master list is used to ensure each group has a unique ID number, and its presence across networks is known.

Resources:

Gephi software (freeware) has some good tutorials on edge and nodelists and how to import them into Gephi: https://gephi.org/users/supported-graph-formats/spreadsheet/

Kumu software is also used for network visualization: https://kumu.io/

Other analytical software options include packages in R. Here are some overviews of social network analysis in R:

https://www.jessesadler.com/post/network-analysis-with-r/

https://statnet.org/trac/raw-attachment/wiki/Resources/introToSNAinR_sunbelt_2012_tutorial.pdf

http://doogan.us/netdata.html