The Comparative Networks Dataset

CASSIE MCMILLAN, SAYALI PHADKE, MITCHELL GOIST, AND MATTHEW DENNY SUNDAY 12TH MARCH, 2017

This document introduces a dataset currently comprising 304 social and biological networks that have been curated in such a way as to facilitate comparative studies of these networks. Where applicable, networks are accompanied by node-level covariate data, and detailed metadata is also available for each network. In particular, we have hand-coded each network into one of a small number of broad categories (exchange, friendship, biological, etc.) to facilitate comparison within and across types of networks. These networks are drawn from a number of sources, all of which are documented in the metadata, and are available as R Lists of adjacency matrices, iGraph objects, and network data objects compatible with the Statnet suite of R packages. The data and documentation can be accessed at this website: mjdenny.com/Comparative_Networks.html.

Our goal in developing this dataset and website was to improve access to large numbers of network datasets in a common and well documented format. A number of previous studies have looked at samples of multiple networks, but have relied on sources that are specific to the author's field of study, and involve re-collecting and reorganizing existing network datasets (Albert et al., 2000; Albert and Barabási, 2002; Gupte et al., 2011; Mones et al., 2012; Shizuka and McDonald, 2012; Corominas-Murtra et al., 2013; Helbing, 2013; Mones, 2013). We believe this represents a great deal of wasted effort, and this project was designed to prevent this sort of wasted effort in the future. Furthermore, we add a substantial amount of value to these network datasets by adopting a common metadata format, and by producing additional metadata such as network categorizations which can be used to analyze subsets of the dataset which are germane to a particular study. In service of this goal, we are actively looking for additional contributors to this project and dataset.

1 Data Sources and Collection

We sourced our data from a number of existing websites and individual studies. Our primary focus during data collection was on websites where network data had already been partially aggregated (such as the UCINet website) so as to maximize our efforts to collect as many networks as possible. Most of our data are downloaded from web pages and catalogued using a standardized data cataloguing procedure which is outlined below. We created a separate folder for each contributor so as to minimize the possibility of overwriting eachother's work, and to prevent errors introduced by a single coder from propagating to the entire dataset.

1. Download the network and any relevant node level covariates and save these raw files in the `Source_Files` subdirectory. It is important to save the raw data in case we find an error in data ingestion process, so we can re-import it. Make sure to use a descriptive/unique name such as `Source_Data_1.whatever` and note the file name in the `Source_Network` field in the `YourName_Network_Metadata.csv` file. Give a similar

name to the node level covariates source file (if applicable), and record the name of this file in the Source_Node_Level_Covariates field in the YourName_Network_Metadata.csv file.

2. Now convert the source network (and node-level covariate data file if applicable) to an adjacency matrix and store it in the Network_CSVs subdirectory (similarly convert any node-level covariate data to a csv, and store in the Metadata_CSVs). Please use a standardized name such as Matt_Network_1.csv (and Matt_Metadata_1.csv) for each network file. Store these file names in the Network and Node_Level_Covariates fields in the YourName_Network_Metadata.csv file.

3. Record metadata in the YourName_Network_Metadata.csv file.

4. Record the number of nodes in the network in the Number_of_Nodes field.

5. Record the type of edge (communication, friendship, load, conflict, etc.) in the Edge_Type field.

6. Record the node type (student, country, cell, dolphin, computer, etc.) in the Node_Type field. Try to be reasonably general here so if the nodes are teachers at a Catholic school, just say they are teachers (same goes for edges).

7. Try to give the network a broad classification and record it in the Network_Type field. The list of Categories: Association, Biological, Ecological, Exchange, Friendship, Kinship, Perception, Support, and Transportation. If you do not believe that the network you are working with fits into one of these categories, mark it with a label you feel is more appropriate and we will discuss this label as a group.

8. If applicable, include the url where you downloaded the file from in the Source_URL field. This will help if we need to go back and clarify anything, so it is essential.

9. Include a one sentence to one paragraph description of the network in plain English in the Description field.

One potential challenge in taking the approach of downloading network data from existing online repositories was that some networks referenced on one site could be duplicated on another. We took steps to mitigate this possibility by conducting several meetings to go over the source data collected by each member of our group, and by checking all networks in our dataset against each other by hand and using an automated approach. The automated approach consisted of comparing all networks against each-other and looking for networks that had an identical number of nodes, and proportion of non-zero edges. If we found any networks that were identical on these two traits, we then checked them by hand to ensure that the networks were not identical. We chose this approach over simple exact matching of the adjacency matrices to catch networks that might have been scaled differently (so the adjacency matrices would not be equal) but still represented the same relations between the same actors.

After conducting a thorough review of our data, we did not find any overlap, mostly because of careful choices of which data sources to pursue. In the future, we will need to continue to be very careful when adding networks to the dataset to ensure we do not introduce overlapping networks. A list of network data sources we considered for this

project is included below. Note that we did not make use of all of these resources due to time and network size constraints, and that many of them contained overlapping links to data. In the future, we intend to exhaust this list and continue to update it with new resources.

1. The UCINet data repository.
2. Stanford's SNAP Lab.
3. Tore Opsahl's website.
4. The National Institute of Standards and Technology: Complex Networks Data Sets.
5. Mark Newman's website.
6. The Koblenz Network Collection.
7. The Arizona State University Network Data Archive.
8. Stanford's SoNIA group website.
9. Princeton's International Networks Archive.
10. The Gephi network dataset archive.
11. The LINK Group at Semmelweis University.
12. Alex Arenas' website.
13. Jake Hoffman's network data repository.
14. Harvard Dataverse.
15. The Abdul Latif Jameel Poverty Action Lab Dataverse.
16. The SEINA network data webpage.
17. The ICPSR data webpage.
18. The Duke Network Analysis Center's network data webpage.

2 Data Harmonization and Formatting

Our goal for this project was to harmonize the data such that it follows a common format (to facilitate running the same analysis across multiple networks), and to provide the data as R Lists of adjacency matrices[1], iGraph objects, and network data objects compatible with the Statnet suite of R packages. We chose `.Rdata` objects as the format for making the data available because R is free and open source, and is the main platform for the dissemination of new methods for analyzing network data. A list of data objects generated by our

1. Metadata for all networks can be downloaded as a [.csv] or an [.RData] object.

[1] Compatible with all network analysis software packages, including the GERGM package (Wilson et al., 2017).

Figure 1: Example List entry for a network in the dataset.

```
List of 3
 $ network        : int [1:50, 1:50] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:50] "V1" "V2" "V3" "V4" ...
 $ node_level_data:'data.frame':50 obs. of  4 variables:
  ..$ Alcohol: int [1:50] 3 2 2 2 3 4 4 4 2 4 ...
  ..$ Drugs  : int [1:50] 1 2 1 1 1 1 3 3 1 1 ...
  ..$ Tobacco: int [1:50] 2 3 1 1 1 1 1 3 1 1 ...
  ..$ Sports : int [1:50] 2 1 1 2 2 2 1 2 2 2 ...
 $ metadata       :List of 9
  ..$ name        : chr "Girls' Friendships 1"
  ..$ directed    : logi TRUE
  ..$ valued_edges: logi FALSE
  ..$ category    : chr "friendship"
  ..$ source_URL  : chr "https://sites.google.com/site/ucinetsoftware/datasets/50women"
  ..$ description : chr "This is a friendship network of a cohort of girls attending a school in Western Scotland..."
  ..$ num_nodes   : int 50
  ..$ node_type   : chr "student"
  ..$ edge_type   : chr "friendship"
```

2. An R List object containing networks represented as (dense) sociomatrices (adjacency matrices) with one list entry per network can be downloaded here as an [.RData] object (approximate file size 483Kb). See below for a full description of each list entry.

3. An R List object containing networks represented as `igraph` network objects with one list entry per network can be downloaded here as an [.RData] object (approximate file size 7.93Mb). See below for a full description of each list entry.

4. An R List object containing networks represented as `network` objects compatible with the `Statnet` libraries with one list entry per network can be downloaded here as an [.RData] object (approximate file size 2.22Mb). See below for a full description of each list entry.

Each entry in one of the data list objects follows the same general structure (Illustrated in Figure 1). They are all R List objects of length 304, with each entry representing a network, and with the List indices matching up to row indices in the Metadata file. As depicted below, each entry in these list objects is itself a List object with three fields: `$network`, `$node_level_data` (which is `NULL` if node node level covariates were available for the network), and `$metadata`. The `$network` field varies based on the way the network is represented (as a numeric matrix, `igraph` object, or `statnet` network object), but the other two fields remain identical across the three different representations of the network. The `$node_level_data` is represented as a `data.frame` if it is available and the rows of the `data.frame` correspond to the rows/columns of the network. The `$metadata` field holds much of the same information as the stand-alone metadata file, and is designed to make it easy to filter networks inside of a loop by checking its values. This structure was designed to facilitate efficient automated analysis across multiple networks using loops or the `apply()` family of functions.

3 Network Categorization and Descriptive Statistics

The data comprise 304 social and biological networks. We began by coding these networks into one of (currently) 11 broad categories: Association, Biological, Ecological, Exchange, Friendship, Kinship, Perception, Support, and Transportation. These categories are described in greater detail below and they allow us to study variation in the properties of these networks across different types of nodes and ties. While we make no claims that these categories are definitive, they serve as a basis for making comparisons between networks, or for looking at particular types of networks.

1. **Association:** This category primarily captures relationships of group co-membership including the number of movies actors have co-starred in, whether two students went to the same school, or the number of scenes two characters in a book shared.
2. **Biological:** This category includes metabolic, protein, and gene interaction networks. This category of networks is distinguished from ecological networks by the nodes, which are not autonomous in this classification.
3. **Ecological:** This category includes interactions, flows, and relationships among animals and ecosystems. Some examples include dominance relationships among cattle, hens, and female sheep, the count of interactions between kangaroos, a monkey-grooming network, and the carbon flow network.
4. **Exchange:** This category includes trade relationships at the national and local levels. Examples include the the volumes of raw materials exchanged between countries and the Taro exchanged among 22 households in a Papuan village, as well as a number of communication networks.
5. **Friendship:** This category records friendship relations between people in a number of different contexts (both in person and online). Some examples include the self assessed friendship networks of high school and college students, prison inmates, bank employees, and monks.
6. **Kinship:** This category includes networked familial relationships, often recorded over a long time period.
7. **Perception:** This category includes networks that were collected by asking respondents to give their perception of romantic, social, friendship, etc. relationships between a group of their peers or subordinates.
8. **Support:** This category primarily includes networks of social support and advice giving. Some examples include advice giving networks in several firms, a law office, and the Harry Potter books, as well as legislative co-sponsorship networks.
9. **Transportation:** This category includes transportation links between cities and countries. For example, one of the networks in this category records whether there is a direct flight between two U.S. cities.

Table 1 provides descriptive statistics for networks in each of these categories (as well as the entire dataset). These include the minimum, median, and maximum number of nodes in networks assigned to that category, the average proportion of non-zero edges in

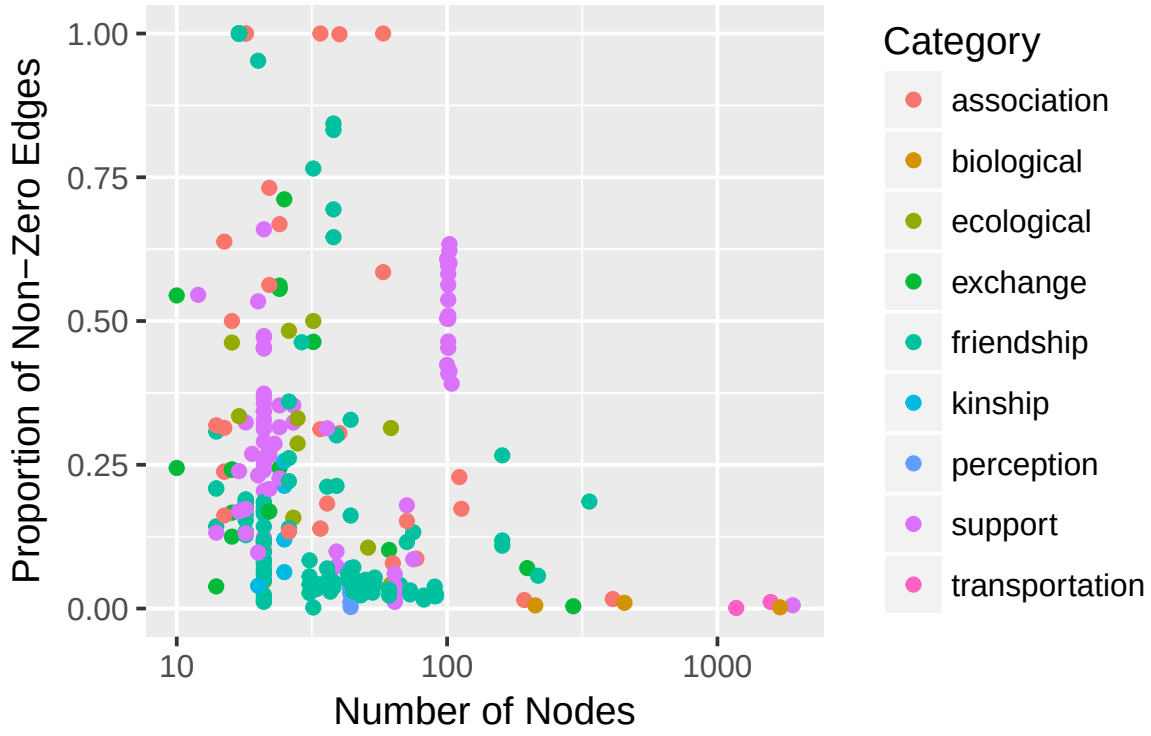
networks assigned to that category, and the count of networks assigned to that category. Support and Friendship networks are the two largest categories, and currently make up over half of the dataset. One important aspect of the dataset is that it does not include any particularly large networks. This is a conscious choice designed to ensure that most forms of statistical analysis can be applied to these networks, and to ensure that the resulting aggregate file sizes would not be prohibitively large. If the reader is interested in perform comparative studies using very large networks, we suggest they look at the data available through Stanford's SNAP Lab. In addition to practical concerns associated with storing and analyzing very large networks, we also believe that there are likely to be substantial differences in the way that network processes operate at the scale of millions of nodes as opposed to the scale of tens of nodes. This focus on smaller networks is reflected in a median network size of just 34 nodes in the dataset, with a maximum network size of under 2,000 nodes.

Table 1: Descriptive statistics by category.

Category	Min. # Nodes	Median # Nodes	Max. # Nodes	Prop. Non-Zero Edges	# Networks
Association	14	34	410	0.41	29
Biological	212	453	1706	0.01	3
Ecological	16	28	62	0.28	11
Exchange	10	24	293	0.30	17
Friendship	14	31	336	0.25	123
Kinship	20	25	25	0.14	5
Perception	44	44	44	0.05	39
Support	12	22	1899	0.31	75
Transportation	1174	1374	1574	0.01	2
All Networks	**10**	**34**	**1899**	**0.25**	**304**

Figure 2 plots the size of the network (on a log scale) against the proportion of non-zero edges for each of the 304 networks in our dataset, with nodes colored by category. As we can see, there is a great deal of heterogeneity in the proportion of non-zero edges across different categories and network sizes, with communication networks showing some of the highest variability across these dimensions.

Figure 2: Plot of network size (on a log scale) against the proportion of non-zero edges for each networks in our dataset, with nodes colored by category.



References

Albert, Reka and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(January), 2002. http://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47.

Albert, Reka, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(July):378–381, 2000. http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html.

Corominas-Murtra, Bernat, Joaquín Goñi, Ricard V Solé, and Carlos Rodríguez-Caso. On the origins of hierarchy in complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(33):13316–21, aug 2013. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3746874{&}tool=pmcentrez{&}rendertype=abstract.

Gupte, Mangesh, Pravin Shankar, Jing Li, S. Muthukrishnan, and Liviu Iftode. Finding hierarchy in directed online social networks. *Proceedings of the 20th international conference on World wide web - WWW '11*, page 557, 2011. http://portal.acm.org/citation.cfm?doid=1963405.1963484.

Helbing, Dirk. Globally networked risks and how to respond. *Nature*, 497(7447):51–9, may 2013. http://www.ncbi.nlm.nih.gov/pubmed/23636396.

Mones, Enys. Hierarchy in directed random networks. *Physical Review E*, 87(2):022817, feb 2013. http://link.aps.org/doi/10.1103/PhysRevE.87.022817.

Mones, Enys, Lilla Vicsek, and Tamás Vicsek. Hierarchy measure for complex networks. *PloS one*, 7(3):e33799, jan 2012. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3314676{&}tool=pmcentrez{&}rendertype=abstract.

Shizuka, Daizaburo and David B. McDonald. A social network perspective on measurements of dominance hierarchies. *Animal Behaviour*, 83(4):925–934, 2012. http://dx.doi.org/10.1016/j.anbehav.2012.01.011.

Wilson, James D., Matthew J. Denny, Shankar Bhamidi, Skyler Cranmer, and Bruce Desmarais. Stochastic weighted graphs: Flexible model specification and simulation. *Social Networks*, 49:37–47, 2017. http://arxiv.org/abs/1505.04015.