An Introduction to Social Network Analysis in R

Vania Wang[1]

Broom Center for Demography, UC Santa Barbara

June 10th, 2019

[1] vaniawang@ucsb.edu

Outline

Topic

Introduction to social network analysis

Network data collection

An introduction to graph theory and network measures

Core analytical metrics of networks

An overall heuristic of the SNA process

Today's activities

References and helpful resources

What are network models?

- Networks are formal representations of dependencies and relations among defined entities
- They are mathematical constructs defined by an edge (E) and node (V) set
- These networks can be directed or undirected
- Attributes can also be assigned to nodes or edges

Why use network models?

- In many statistical models, the assumption of independence is chief
- However, what if we're interested in effects where dependence among etities is of interest?
- This is where network models come in: they provide a formalized model to represent relationships among a set of entities.

Epistemology of social network analysis

The history of social network analysis is interdiscplinary. The need to represent social entities using networks was expressed by sociologists. The analytical techniques were pioneered by physicists and computer scientists.

- The field of social network analysis emerged in the 1930's, introduced by psychiatrist Joseph Moreno and psychologist Hellen Jennings who pioneered the sociometry, or the study of social interactions.
- Between the 1930's and 1970's, the field was largely fronted by sociologists and and psychologists. But from 1970 to the 1990's, physicists and computer scientists helped apply graph theoretical methods to the study of social networks.
- For more information on the history of social network analysis, please read a wonderful overview by Linton Freeman (Freeman 2014).

Examples of social networks

- Kinship networks
- Social media, social influence, and digital connection
- Sexual networks for HIV modeling

Topic

Field work and empirical data collection

Sociocentric

- Data collection centers around the population, and how its constituent individuals are connected
- Requires a boundary definition to the population under study
- Identified relationships are surveyed

Egocentric

- Data collection centers around the individual and their connections
- Identified network relationships are not intentionally surveyed

Analytical differences between sociocentric and egocentric data

Sociocentric

- Limited by population size and logistical complexity of data collection (expensive and difficult to implement)
- *But*, network structure can be empirically recreated through the data.
- Therefore, network structure is recreated and *not* inferred.
- A wonderful paper by Helleringer and Kohler demonstrate the usage of sociocentric data collection and analysis (Helleringer and Kohler 2007).

Analytical differences between sociocentric and egocentric data

Egocentric

- Limited by completeness and validity of network information
- *But*, the expanse of data collection is larger in scope (easier to implement)
- Network structure is inferred by a suite of statistical models called *exponential random graph models* (ERGMS).
- Please refer to this paper for a good example on egocentric network analysis (Jenness 2016).

Topic

Core graph components

Graphs representing social relations consist of two core components: actors and ties

Actors

Also known as nodes or vertices. These refer to individual entities that are related with other entities (of the same type) in some way. Oftentimes in SNA, this refers to a population or communities of people.

Ties

Also known relation or edge. These represent relationships between actors. In sexual network modeling for HIV, every tie would represent a defined sex act. In a twitter network, the tie could be a follower/following relationship (*would this be a directed or undirected graph?).

Assigning attributes to graph components

Now that we've established that social networks are composed of two components, actors and ties, additional complexities can be layed onto these base entities:

Weighted ties

Ties can have weights, or representations of how weak or how strong that tie is.

Attributes on actors and ties

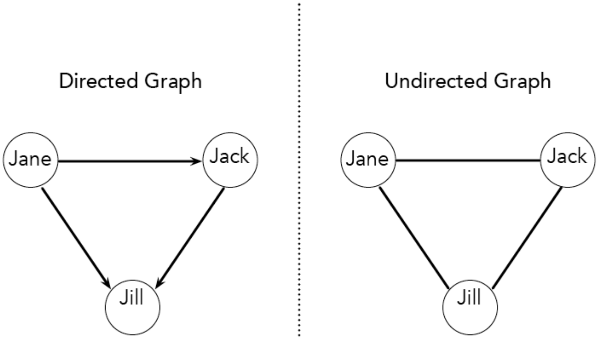
Actors and ties can both be assigned attributes. Within the context of modeling social processes, actors may have the following attributes: *race, age, SES, HIV status*. Tie attributes be, type of social interaction (re: Twitter, follower or following relationships) or sentiment (re: enemy vs friendship).

Directed vs undirected graphs

The distinction between undirected and directed graphs lies in how relationships are represented within your model. It's important to consider, does the relationship have directionality?

An illustrated example: directed vs undirected graphs

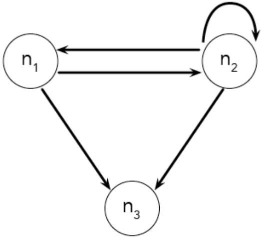
What is the meaning of friendship?



Perceived friendship

An interaction

Data structures and representations of network data: creating the overall graph



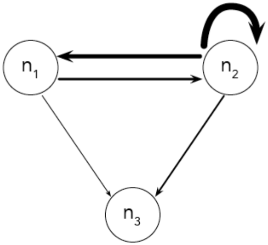
Edge list

n_1	n_2
n_1	n_3
n_2	n_1
n_2	n_2
n_2	n_3

Adjacency matrix

	n_1	n_2	n_3
n_1	0	1	1
n_2	0	1	1
n_3	0	1	0

Data structures and representations of network data: layering complexity



Edge list and weights

n_1	n_2	weight 3
n_1	n_3	weight 1
n_2	n_1	weight 5
n_2	n_2	weight 10
n_2	n_3	weight 2

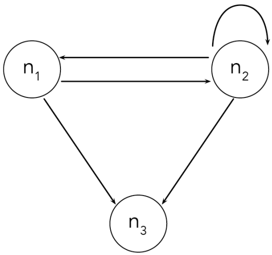
Nodal attributes

n_1	age: 10	sex: F
n_2	age: 19	sex: M
n_3	age: 28	sex: F

Topic

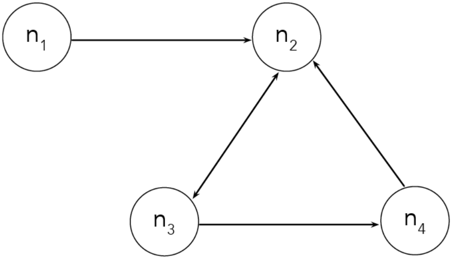
Metric 1: Degree centrality

- For an undirected graph, the degree of a node is the number of adjacent nodes.
- For a directed graph, we distinguish between indegree and outdegree.
- What is the indegree and outdegree of node n_1 in the following graph?



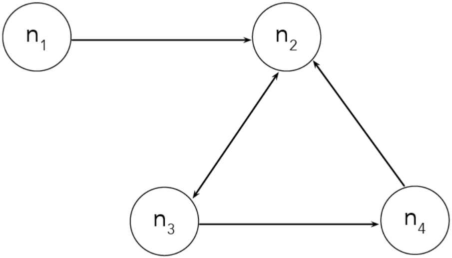
Metric 2: Betweeness centrality

- This is the proportion of shortest paths between all other pairs of nodes that the given node lies on.
- This is calculated on a node-by-node basis.
- What is the betweeness centrality of node n_1? What about n_2?



Metric 3: Geodesic distance

- The number of edges in the shortest path connecting two nodes.
- What is the geodesic distance between n_1 and n_4?

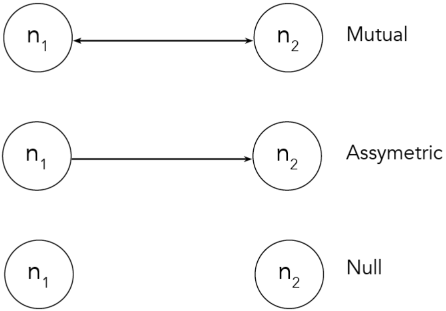


Metric 4: Network density

- This is the number of ties, expressed as a percentage of the number of possible ties.
- This is slightly different between directed and undirected graphs.
- For directed graphs: $E/[N(N-1)]$
- For undirected graphs: $0.5 * E/[N(N-1)]$

Metric 5: How many dyads in the network?

- Dyads are defined as pairwise ties.
- In the `sna` package, the `dyad.census` function counts all mutual (`M`), assymetric (`A`), and null (`N`) dyads.



Topic

From abstraction to analysis

Social network analysis is fundamentally a modeling process. It's therefore helpful to approach the study of social networks through a precise framework that permits quantitative analysis of social processes.

1. Conceptualize your model: Social relationships obviously are not *literally* actors and ties. But it's useful to think specifically about what a tie or actor means in your model.
2. Collect your data: After you've conceptualized your model, should you go with egocentric and sociocentric data collection?
3. Analyze and visualize your results: In the next section of this workshop, we'll go over basic data analysis techniques of social network data using the `network` and `GGally` packages.

Topic

Exercise 1: A smaller network example (live coding example)

For exercise 1, we will analyze two artificial networks. We will be using the files `e1_edge-attr.csv`, `e1_ad-matrix.csv`, `ex1_vertex-attr.csv`. The tasks for this exercise are,

1. Load network data into R
2. Calculate and find analytical metrics
3. Visualize the network using the `ggnet2` and `gplot` functions

Exercise 2: Working with larger networks, Google Plus network (group work)

In exercise 2, we will take a large edge list dataset representing a Google Plus friendship network. Although large, the dataset is composed of actors and ties *only*. No attributes are included in network. The tasks for this exercise are,

1. Both groups will load the network data into R
2. Calculate and plot degree measures
3. Understand the significance of small world networks

Topic

References

The following list includes references included in the presentation and other helpful resources.

- Social Network Analysis: Methods and Applications (Wasserman, Faust | 1994): This is *the core text* on social network analysis.
- Exponential Random Graph Models for Social Networks (Lusher, Koskinen, Robins | 2013): A foundational text on ERGMs. This is pretty dense on the statistical component of the modeling process, but still digestable given prior stats knowledge.

Tutorials for further learning

The study of social networks is broad! We only touched a modicum of the field today. Here are some other helpful resources on self-learning SNA in `R`.

- The `igraph` package in `R`: In this tutorial, I introduced the `network` package for representing network information. Another competing (in terms of popularity) is the `igraph` package, which has a lot of the same functionality. Chief among these functions is its ability to assess network homophily.
- The `statnet` package in `R`: The statnet analytical package is based on ERGMs. The developers of this package have very helpful suite of tutorials on using the package. Access them here.
- `networkD3`: This is a JavaScript-based package in R that can create really cool interactive graph visualizations.