Yale

Introduction to Social Network Analysis with R

Kayleigh Bohemier, Breanne Chryst, and Anderson Zhang

April 22, 2016

Outline

Introduction

Vertices/Nodes and Edges

- Wikipedia defines Social Networks as *"a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors."*
- In network parlance the social actors (people, groups, nations) are node or vertices and the ties between them are also called edges.

Directed and Undirected

- Directed graphs have edges originating from one node and ending in another. An example is twitter, if I follow you, there is an edge originating from me and ending at you, but you may not follow me back.
- Undirected graphs have reciprocal edges. An example is Facebook, if I am your friend, you are also my friend.
- The decision of whether a network is directed or undirected is an important one and can have major implications for the analysis. Careful thought should be spent in deciding on this point.

Adjacency Matrices

- The data for a social network can be organized as a matrix, with non-zero entries in the i, j^{th} entry of the i^{th} node shares an edge with the j^{th} node.
- If the network is undirected this matrix is symmetric.
- If the edges have weights, the entries in the non-zero entries are the weights for the edges.
- If the edges are not weighted, the non-zero entries are simply one.

Adjacency Matrices

	A	B	C	D	E	F	G	H	I	J	K	L
1		Bert	Adam	Batman	Sherlock	Ivy	Watson	Moriarty	Robin	Joker	Eve	Ernie
2	Bert	0	1	0	0	0	0	0	0	0	1	1
3	Adam	1	0	1	1	1	1	1	1	1	1	1
4	Batman	0	1	0	0	1	0	0	1	1	1	0
5	Sherlock	0	1	0	0	0	1	1	0	0	1	0
6	Ivy	0	1	1	0	0	0	0	1	0	1	0
7	Watson	0	1	0	1	0	0	0	0	0	1	0
8	Moriarty	0	1	0	1	0	0	0	0	0	1	0
9	Robin	0	1	1	0	1	0	0	0	1	1	0
10	Joker	0	1	1	0	1	0	0	1	0	1	0
11	Eve	1	1	1	1	1	1	1	1	1	0	1
12	Ernie	1	1	0	0	0	0	0	0	0	1	0
13												

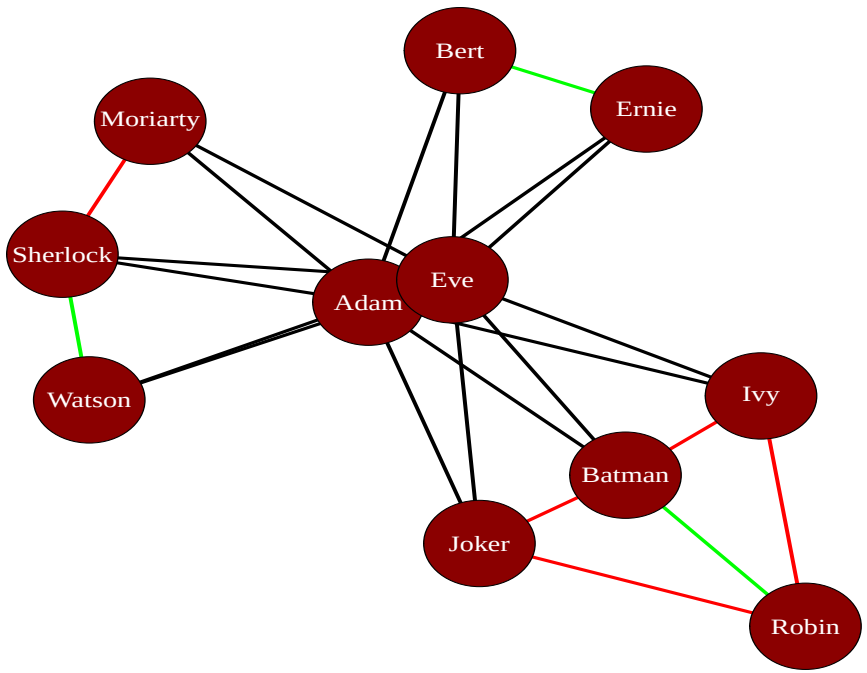
Edgelists

- Data can be stored as an edgelist, specifying the nodes that have a tie between them in the graph.
- If the network is directed the first entry is the start of the edge and the second is the end of the edge (i.e. there is an arrow connecting the two nodes pointing in the direction of the second).
- For undirected networks there is only one entry per connection and the order doesn't matter.

	A	B	C	D
1	Source	Target	Weight	Type
2	Bert	Ernie	1	Partners
3	Adam	Eve	1	Partners
4	Batman	Joker	1	Enemies
5	Sherlock	Watson	1	Partners

```
> get.edgelist(g)
      [,1]       [,2]
[1,] "Bert"     "Ernie"
[2,] "Adam"     "Eve"
[3,] "Batman"   "Joker"
[4,] "Sherlock" "Watson"
[5,] "Batman"   "Robin"
[6,] "Batman"   "Ivy"
[7,] "Robin"    "Ivy"
```

Network Graph



Publicly Available Datasets

- UC Irvine has a nicely curated list of publicly available network data at `https://networkdata.ics.uci.edu/resources.php`
- Stanford Large Network Dataset Collection: `https://snap.stanford.edu/data/`
- Of course there are many more resources throughout the internet.

API

- Poplication Program Interface (API) allow users to interact with programs. Facebook and Twitter are two of the main sources that come to mind when thinking about social network data. You may be able to access some data from these sites through their API.
  - Twitter: `https://dev.twitter.com/overview/api`
  - Facebook:
    `https://developers.facebook.com/docs/graph-api/overview`

From the Field

- Gather your own social network data through observations, literature, etc.
- When gathering your own data it is important to think about what constitutes a node and an edge in your data, before you start collecting the data.

Basic Social Network Analysis (SNA) in R

Intro to R

- R is both a statistical package (like SPSS, Stata, SAS, etc.) and a programming language.
- More accurately, R is an environment within which you can do statistical programming and graphics.
- R provides an environment and a language that allow you to compute, analyze, and graph your data - and so much more.
- R is an object-oriented programming language.
- Everything you do and create in R, including statistical models, functions, graphics, and output, is an object stored in memory that can be manipulated, saved, and reused for greater efficiency and power.

Brief Intro in R

- Lets try out some R!

Reading network Data into R

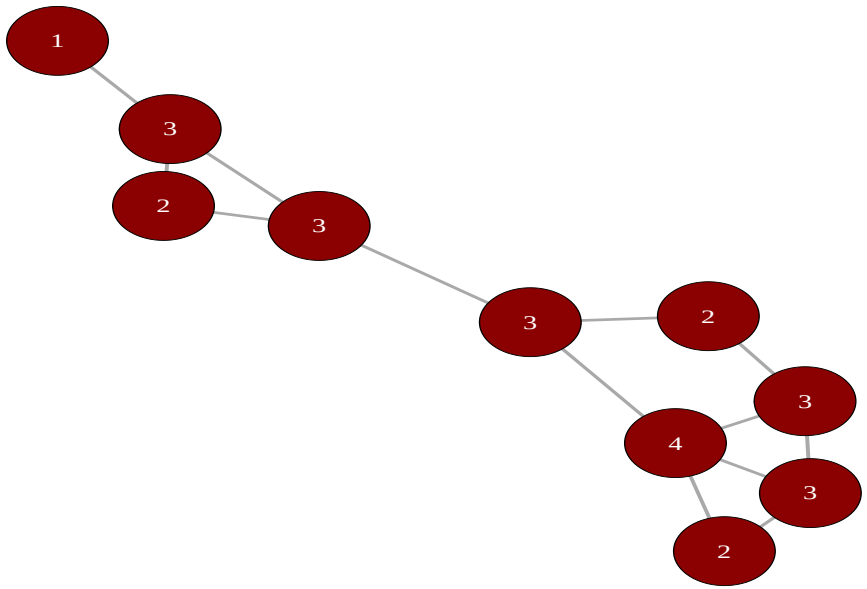
- Reading network data into R depends on the data file type.
- **read.csv** can be used for reading in .csv files
- **read.table** can be used for reading in .txt files
- For other file types, you may need packages like **foreign**

Local Network Statistics

- Centrality: a measure of importance for vertices in the network.
- Common measures of centrality include degree, closeness, betweeness, and eigenvector centrality.

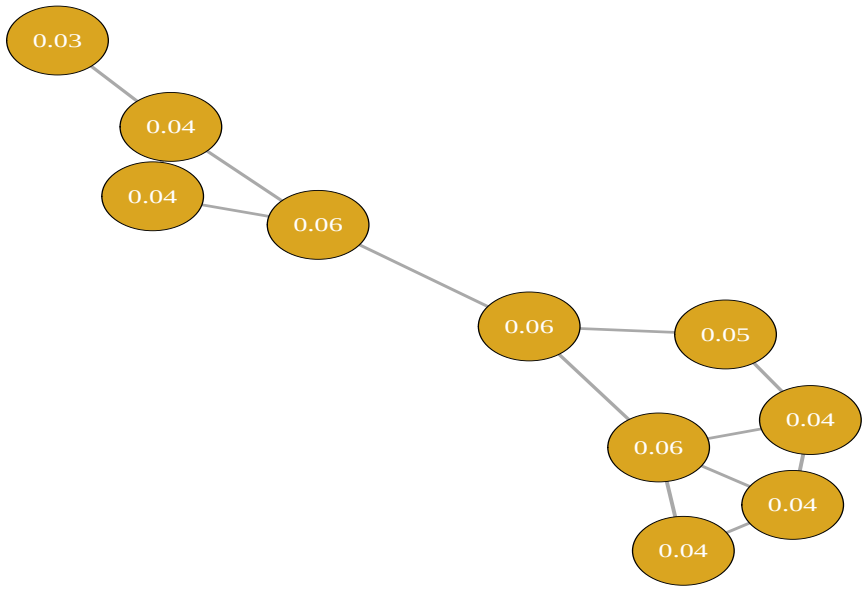
Degree

- The degree is the number of edges for a node (number of alters or social connections in the network).



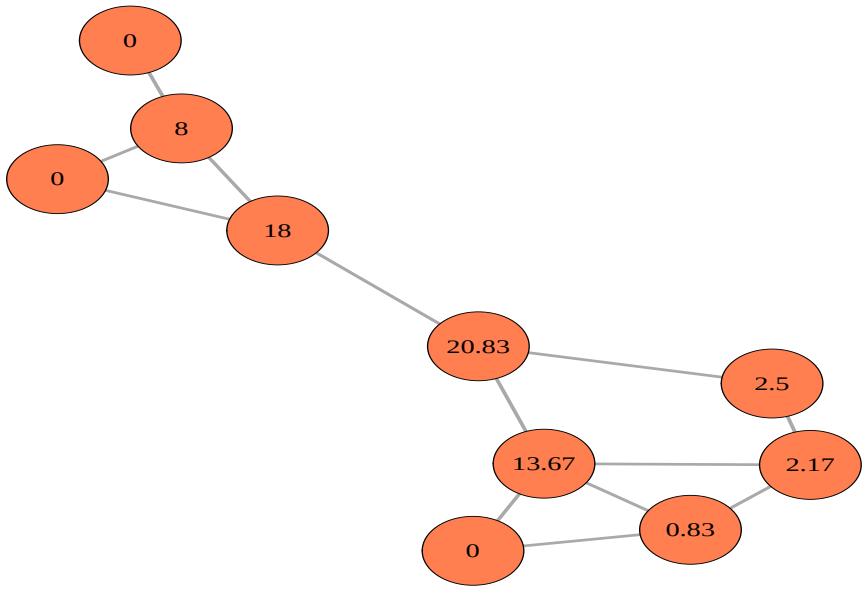
Closeness

- Closeness is the number of nodes "close" in geodesic distance or shortest path between vertices.



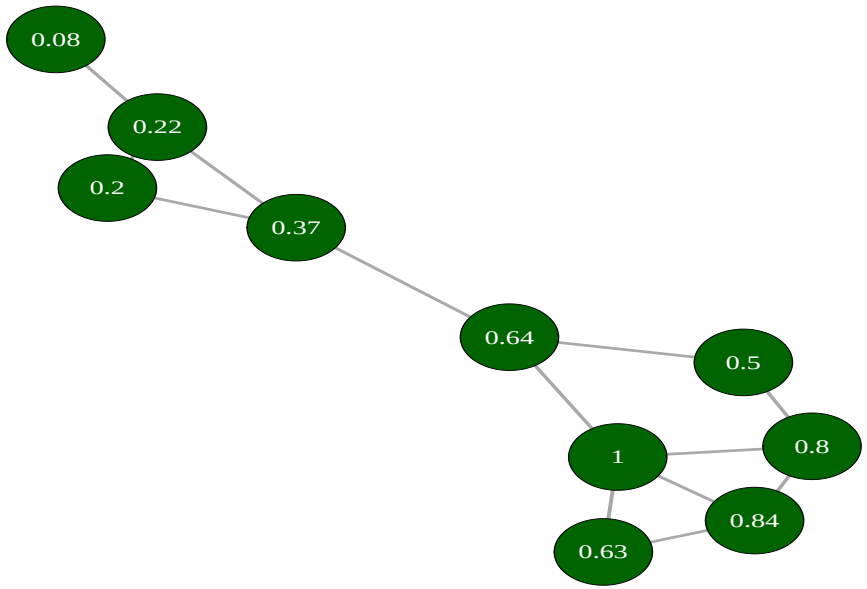
Betweeness

- Betweeness measures the extent to which the vertex connects disparate groups.



Eigenvector

- Eigenvector centrality uses the first eigenvector to create a centrality measure that is a function of its neighbors centrality.

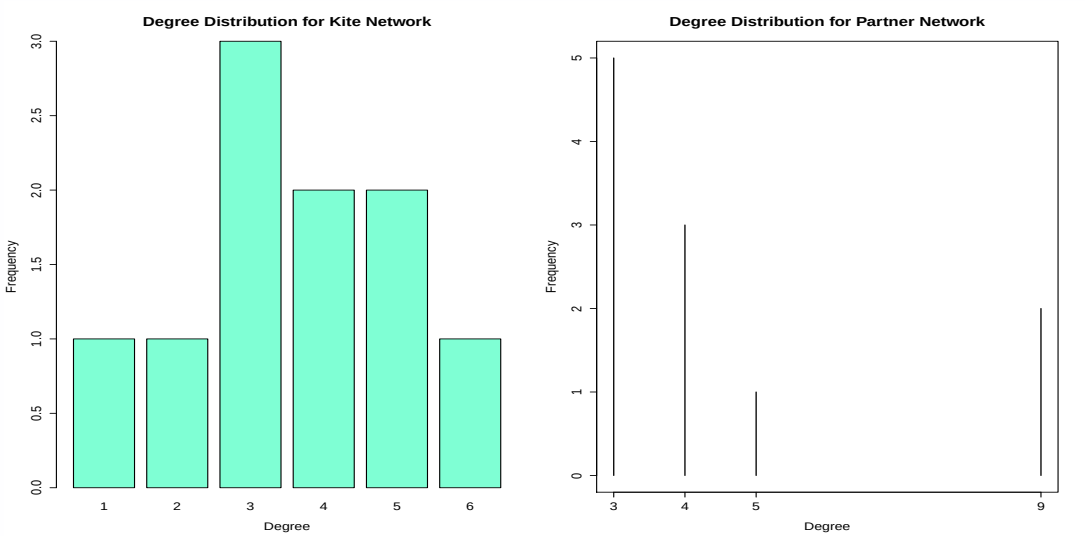


Global Network Statistics

- Average Shortest Path: the average shortest path between vertices in the graph.
- Connectivity/Density: the proportion of the possible edges that are realized in the graph.
- Average Degree: mean of the degrees in the network.

Global Network Statistics

- Degree Distribution: the distributions of degrees for the network. this summarizes the connectivity of the network.



Global Network Statistics

- Clustering Coefficient/Transitivity: the number of connected triangles divided by the number of connected pairs in the graph.
- Partitioning: refers to separating the graph into k disjoint sets of vertices, usually based on the connectivity within partitions.
- Assortivity/Homophily: describes the tendency for "birds of a feather flock together."

Try it out