Social Networks: A Fast Tour From People to Groups!

Matthew A. Hoover

02 May 2017

Welcome!

- Data scientist at Gallup
- Ph.D. in public policy
  - Dissertation on the effects of children's social networks on education in rural Afghanistan
  - Research on how social networks affected individual decision-making
- Previous life, 15 years in international development along with additional work for a healthcare startup

What is social network analysis?

- Understanding the structure, composition, and purpose of people's social networks, whether in-person or online
- It helps answers questions from "how do my friends and acquaintances affect my behaviors" to "from whom can I seek support in a given situation"
- **Structure** identifies how ties connect people in certain ways – are there mutual ties, triangles, cycles?
- **Composition** describes the characteristics of people that are connected – are they the same gender, about the same age?
- **Purpose** of a particular network varies – is it a support network, drug seeking/using network, professional network?

What we'll cover today

- Build up the conception of a network from an individual
- Introduce network measures: degree, centrality, triangles, and isolates
- Discuss analyses: community structure, ERGMs, SAOMs
- Visualize networks and how/why that's important
- DIY: Building a network

What we'll cover today

- Build up the conception of a network from an individual
- Introduce network measures: degree, centrality, triangles, and isolates
- Discuss analyses: community structure, ERGMs, SAOMs
- Visualize networks and how/why that's important
- DIY: Building a network
- **The point-of-view for this talk are human networks, so the scale is considerably smaller – and more tractable – than large computer or web networks, like Facebook or Twitter**

Let's start small: The individual

- The individual plays the key role in most econometric analyses
- However, misses relational information between people
  - Relations can guide actions or behaviors (*influence*)
  - Actions can determine relations (*selection*)
- In networks, an individual is called a **node**
  - Terminology borrowed from graph theory, as another representation of a network is a graph
  - Each node can have a series of attributes: age, gender, beliefs, career
  - Note, networks do not have to be of people only

A simple question: What is a network?

- 'I know it when I see it': Simply put, it's a collection of entities (nodes) connected in some way (edges)
- That collection of entities – depending on how they are connected – do or do not make a network... moreover, depending on the connection type, it forms different networks

A simple question: What is a network?

- 'I know it when I see it': Simply put, it's a collection of entities (nodes) connected in some way (edges)
- That collection of entities – depending on how they are connected – do or do not make a network. . . moreover, depending on the connection type, it forms different networks
- The size of a network can differ dramatically; for example. . .
  - Friendship ties of high schoolers *in real life*
  - Friendship ties of high schoolers *on Instagram*

A simple question: What is a network?

- The type of network can differ too...
  - A *one-mode* network is of a single entity type, e.g., dogs connected to other dogs through breeding
  - A *two-mode* network is of two separate entities, e.g., people connected to beers they drank
  - A two-mode network can be ‘flattened‘ to create a person-to-person or beer-to-beer network

A simple question: What is a network?

- The type of network can differ too...
  - A *one-mode* network is of a single entity type, e.g., dogs connected to other dogs through breeding
  - A *two-mode* network is of two separate entities, e.g., people connected to beers they drank
  - A two-mode network can be ‘flattened‘ to create a person-to-person or beer-to-beer network
- Finally, the scope of the network can differ
  - A *complete* network looks at the whole network, e.g., doctors’ referrals to other doctors in a hospital
  - A *personal* or *ego-centric* network focuses on the constellation of nodes around an entity in particular, e.g., whom jazz musicians have sessioned with in the past
  - Both networks have their advantages and disadvantages

The seeds of a network: Dyads and triads

- The start of a network is the connection of two nodes through a edge: A *dyad*
- This already adds complexity – is the network *undirected* or *directed*?

The seeds of a network: Dyads and triads

- The start of a network is the connection of two nodes through a edge: A *dyad*
- This already adds complexity – is the network *undirected* or *directed*?
  - An undirected network means there is no directionality in the tie – the tie is the same for both nodes, e.g., two servers that are connected to one another
  - A directed network means ties have a sender and a receiver and the connection flows one-way only, e.g., followers on Twitter
  - Of course, in a directed network and edge can be bi-directional, which is generally seen as a 'stronger' tie than a uni-directional connection

The seeds of a network: Dyads and triads

- The start of a network is the connection of two nodes through a edge: A *dyad*
- This already adds complexity – is the network *undirected* or *directed*?
  - An undirected network means there is no directionality in the tie – the tie is the same for both nodes, e.g., two servers that are connected to one another
  - A directed network means ties have a sender and a receiver and the connection flows one-way only, e.g., followers on Twitter
  - Of course, in a directed network and edge can be bi-directional, which is generally seen as a 'stronger' tie than a uni-directional connection
- A core building block in networks are *triads* or the grouping of three nodes together in some way.

The seeds of a network: Dyads and triads

- A triad, even though only three nodes, is already an interesting structure. It can exhibit hierarchy, closeness, or little relation between nodes
- With many nodes in a network, one technique to use is a *triad count*, which counts the number of triads for the various possible formations
- Triads, depending on how they are formed and interconnect, help to determine the more advanced structure of a network, including cliques and sub-graphs

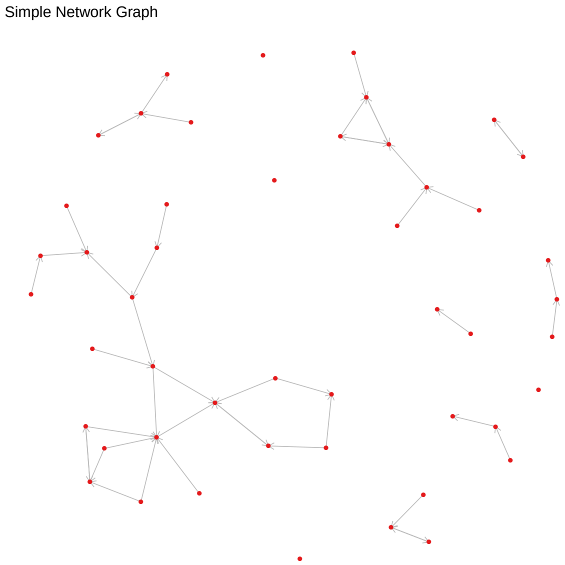
The seeds of a network: Dyads and triads

- A triad, even though only three nodes, is already an interesting structure. It can exhibit hierarchy, closeness, or little relation between nodes
- With many nodes in a network, one technique to use is a *triad count*, which counts the number of triads for the various possible formations
- Triads, depending on how they are formed and interconnect, help to determine the more advanced structure of a network, including cliques and sub-graphs
- An outgrowth of triads are *isolates*, which are nodes in a network that are not connected to any other node
- Isolates often exhibit unique behavior or have attributes that differ from others within the network
- A network itself can be composed of many isolates, which is also interesting – depending on the purpose of the network, this could be expected or potentially problematic

The total network

- From triads, the network builds into more complex structures that can be broken down to isolates, dyads, and triads
- At this point, network measures and statistics become important
- The most basic is *density*, a measure on the network itself – of all possible connections, how many are present?
- At the node level, there are measures of *centrality*:
  - *Degree* (*in-degree* and *out-degree* for directed networks): The number of ties for each node in a network; in-degree/out-degree centrality can be called popularity and activity, respectively
  - *Betweenness*: A measure of position of a person – do they sit 'between' others or not; nodes will high betweenness centrality are 'bridges' to other parts of the network
  - *Eigenvector*: Measures the connections' connections, that is, nodes have higher values if their connections are well-connected and those connections are well-connected and so on
- Let's take a minute or two to look at a network and then take a look at some of these network summary measures

A directed network



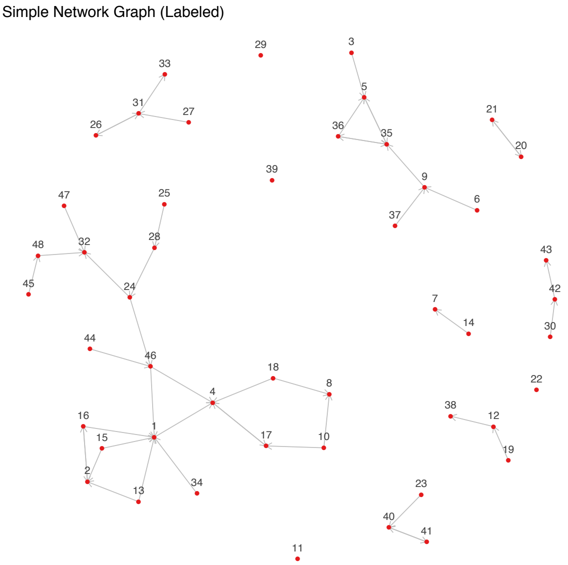
Network summary statistics

Table 1: Directed Network Summary

Measure	Value
Number of Nodes	48
Number of Edges	48
Number of Isolates	4
Network Density	0.021
Degree Centralization	0.044

- Think about the following as we look at the graph again:
  - Which nodes have high degree centrality?
  - What about betweenness centrality? Why?
  - Thoughts on high eigenvector centrality?

A directed network



Relating network information back to the individual

- Let's pause for a second: So these measures may be *interesting*, but what can be done with them beyond descriptive statistics?
- There needs to be a way to relate network summaries – and the statistics we will discuss next – to the individual
- There are two choices in particular:

Relating network information back to the individual

- Let's pause for a second: So these measures may be *interesting*, but what can be done with them beyond descriptive statistics?
- There needs to be a way to relate network summaries – and the statistics we will discuss next – to the individual
- There are two choices in particular:
  - The first is using node-level summaries as individual-level measures
  - These measures can be used as covariates in modeling efforts (remember to normalize!)

Relating network information back to the individual

- Let's pause for a second: So these measures may be *interesting*, but what can be done with them beyond descriptive statistics?
- There needs to be a way to relate network summaries – and the statistics we will discuss next – to the individual
- There are two choices in particular:
  - The first is using node-level summaries as individual-level measures
  - These measures can be used as covariates in modeling efforts (remember to normalize!)
  - Second, if there are individuals from multiple networks, then think about a hierarchical model, if the outcome is amenable
  - Works well for analyses utilizing personal networks or similar complete networks measured at the same time

Relating network information back to the individual

- Let's pause for a second: So these measures may be *interesting*, but what can be done with them beyond descriptive statistics?
- There needs to be a way to relate network summaries – and the statistics we will discuss next – to the individual
- There are two choices in particular:
  - The first is using node-level summaries as individual-level measures
  - These measures can be used as covariates in modeling efforts (remember to normalize!)
  - Second, if there are individuals from multiple networks, then think about a hierarchical model, if the outcome is amenable
  - Works well for analyses utilizing personal networks or similar complete networks measured at the same time
- **Network analysis has suffered because people often can't figure out what to do with them in practice**

Moving from descriptives: Identifying structure in a network

- Beyond network descriptive statistics and visualizations, we can use algorithms to determine underlying structure in networks
- The value of structure is it helps to sub-divide a network into smaller, (more) cohesive groups

Moving from descriptives: Identifying structure in a network

- Beyond network descriptive statistics and visualizations, we can use algorithms to determine underlying structure in networks
- The value of structure is it helps to sub-divide a network into smaller, (more) cohesive groups
- Some examples of community structure:
  - Clients and personal relationships in commercial sex workers' lives
  - Grade levels within a high school
  - Political affiliation of Twitter users

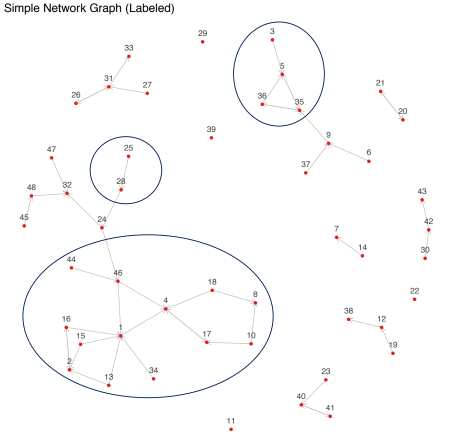
Moving from descriptives: Identifying structure in a network

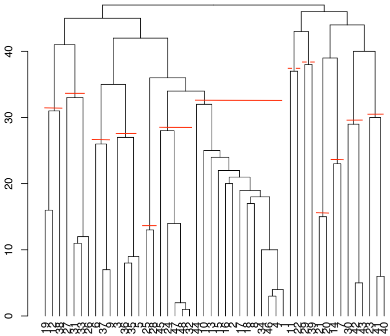
- Beyond network descriptive statistics and visualizations, we can use algorithms to determine underlying structure in networks
- The value of structure is it helps to sub-divide a network into smaller, (more) cohesive groups
- Some examples of community structure:
  - Clients and personal relationships in commercial sex workers' lives
  - Grade levels within a high school
  - Political affiliation of Twitter users
- Of course, the algorithms only do the math – the analyst needs to understand what the results mean

Community detection algorithms

- Many variations; one of the more well-known algorithms is Girvan-Newman, which is a hierarchical method for detecting community structure
    1. Calculate betweenness
    2. Remove edge with highest betweenness
    3. Repeat 1 and 2 until no edges are left
- Other algorithms exist, utilizing different rules for structuring
- Some will only work with undirected networks
- A community detection algorithm is not *the answer* it is *an answer* to help better understand what's happening in a network.

Girvan-Newman on our simple network

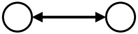
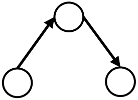
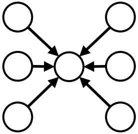




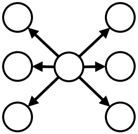
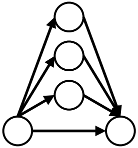
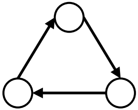
Modeling networks, not individuals

- Up to now, we've discussed operations that help identify network features for *individual* analyses
- There are *network-based* analyses that are possible now, with the advent of greater computing power
- Exponential random graph models (ERGMs)
  - Cross-sectional
  - Identifies structural and compositional elements of a network
    1. Start with network representation
    2. Use MCMC to remove/add random edges
    3. Take 'snapshot' of network
    4. Repeat 1 through 3 a given number of times
    5. Calculate how likely/unlikely given structural/compositional characteristics are in network representation, given all other network possibilities

ERGM terms

Statistic	Visualization	Formula	Description
EDGES		$\sum_{i,j} y_{ij}$	Sum of all ties in network
MUTUAL		$\sum_{i<j} y_{ij} y_{ji}$	Sum of all reciprocated ties in network
TWOPATH		$\sum_{i \neq j \neq k} y_{ij} y_{jk}$	Sum of all paths containing exactly one in-degree and one out-degree
GWIDEGREE		$\sum_{i=0}^{n} e^{-\alpha y_{+i}}$	Indegree distribution, accounting for decrease in marginal utility of each additional nomination received

ERGM terms

Statistic	Visualization	Formula	Description
GWODEGREE		$\sum_{i=0}^{n} e^{-\alpha y_{i+}}$	Outdegree distribution, accounting for decrease in marginal utility of each additional nomination sent
GWESP		$e^{\theta_t} \sum_{i=1}^{n-1} \left\{1 - \left(1 - e^{-\theta_t}\right)^{i}\right\}$ $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$	Transitive triplet distribution, accounting for decrease in marginal probability of closing triplet
CTRIPLE		$\sum_{\substack{i \neq j \neq k \\ i < j,k}} y_{ij} y_{jk} y_{ki}$	Sum of all cyclic triples in network
MATCH		$\sum_{i,j} y_{ij} \mathbb{1}\{D_i = D_j\}$	Sum of dyads matched on specified attribute

ERGM estimates on our simple network

Table 2: Simple network ERGM parameters

Parameter	Estimate	Std. Error	*p*-value
edges	-15.73	3.928	$<$0.001
mutual	2.465	0.597	$<$0.001
twopath	-0.385	0.214	0.072
gwidegree	-2.790	1.136	0.014
gwodegree	12.864	3.912	0.001
gwesp	0.927	0.528	0.079
ctriple	0.686	1.455	0.637
gender (match)	3.437	1.011	0.001

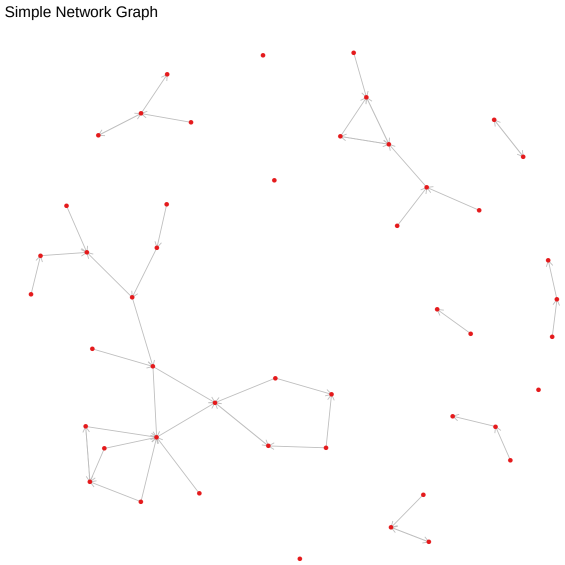
Modeling networks over time

- ERGMs can be thought of as a cross-sectional analysis; with networks, temporal analysis is possible as well
- Stochastic actor-oriented models (SAOMs) identify and measure the change in a network over time
- Help measure two processes within the network – *selection* and *influence*
- Selection: Are connections chosen based on a shared attribute?
- Influence: Do connections induce behavior change?
- Teen smoking: are friends chosen because they smoke (selection) or does smoking start because other friends are smoking (influence)
- Both processes can happen at the same time

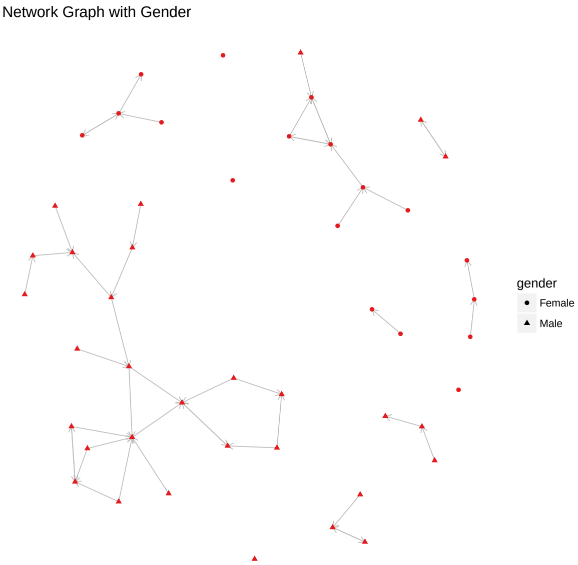
Using visuals for qualitative analysis

- Utilizing visuals can improve knowledge and understanding of a network
- Can utilize attributes on the network to help – size, color, shape of nodes; can even use edge attributes (size, color, transparency)
- As with any visual, be careful that what's added doesn't ultimately distract from the message

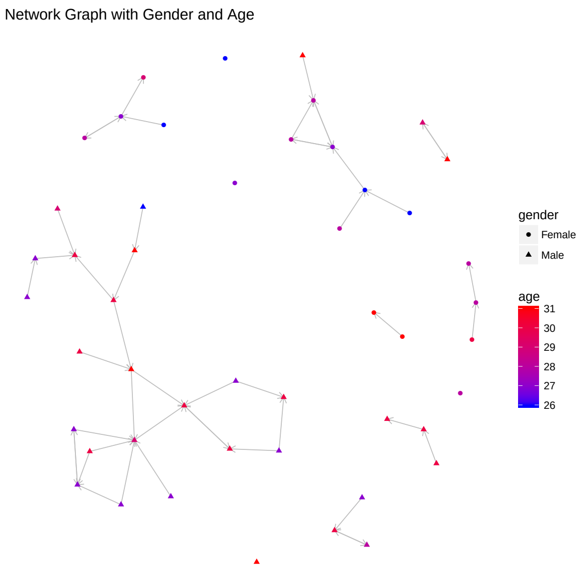
A plain network



A network with shape



A network with shape and color



Building networks

- It is expensive and time-consuming to collect network data
- However, building a network isn't too difficult; with API access and some code, network construction is certainly possible
- Most APIs will provide some information on users (nodes) and based on the site, there are probably ways to identify connections (edges)
- Depending on the edges, this may create a one- or two-mode network
- Examples of network possibilities:
  - Twitter followers (one-mode)
  - Friends on Instagram (one-mode)
  - Hashtags used in tweets (two-mode)
  - *Beers checked-in on Untappd* (two-mode)
- Let's take a look at building a one-mode network by collecting data from an API on a two-mode network and figure out some interesting things along the way

Creating a network from Untappd

- Untappd is a mobile app that allows people to track the beers they are drinking
- Its API provides access to user details as well as the beers that users check in (along with other information that we'll ignore for this exercise)
- So let's say I am interested to know not just *who* my friends are (which I already have due to friend connections), but *which of my friends have similar beer preferences*
- We could:
    - Pull all the beers I have checked-in along with all of my friends ids
    - For each friend, grab all the beers they've checked-in
    - Create an edgelist of person-to-beer
    - 'Flatten' the two-mode network to a one-mode person-to-person (or, beer-to-beer) network, where the edges are the number of beers in common two nodes have checked in

What does the network look like?

- Two-mode network of user and beer id's
- Represented as an edgelist

User	Beer
3171525	1468817
3171525	12955
3171525	127175
645175	1623868
645175	1452189
645175	878224

Once flattened to a user-to-user network

- What do these values in the table represent?
- Why is the matrix symmetrical?
- How could this information be used in a network visualization?

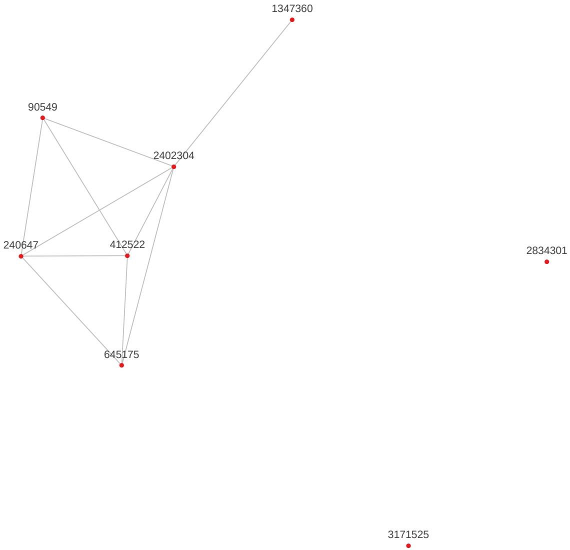
	90549	240647	412522	645175	1347360	2402304	2834301	3171525
90549	150	1	24	0	0	1	0	0
240647	1	150	3	7	0	3	0	0
412522	24	3	150	1	0	4	0	0
645175	0	7	1	126	0	3	0	0
1347360	0	0	0	0	150	1	0	0
2402304	1	3	4	3	1	150	0	0
2834301	0	0	0	0	0	0	14	0
3171525	0	0	0	0	0	0	0	3

Some descriptive statistics on the network

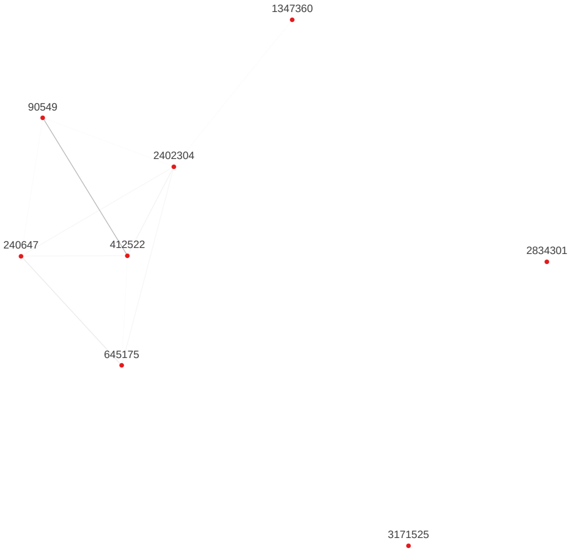
Table 3: Directed Network Summary

Measure	Value
Number of Nodes	8
Number of Edges	10
Number of Isolates	2
Network Density	0.357
Degree Centralization	0.476

Beer network visualized



Beer network with tie weight



Thank you!

- Thank you for your time!
- All materials are on my GitHub page: github.com/mhoover
  - `presentations/spdc_may2017`
  - `ggnet`
  - `untappd`
- Good references:
  - **Social Network Analysis: Methods and Applications** by Stanley Wasserman and Katherine Faust (the maroon book)
  - **Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications** edited by Dean Lusher, Johan Koskinen, and Garry Robins (the black-and-green book)
- Questions?