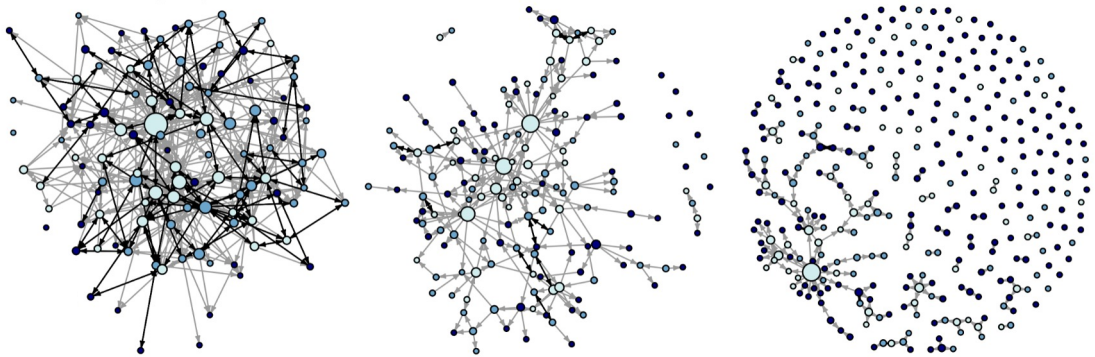
(A two-hour)
Introduction to Social Networks



Elspeth Ready and Daniel Redhead

Skillshare Workshop

(A two-hour)
Introduction to Social Networks

Outline

I. Basic concepts, data structures and operations
II. Overview of sampling strategies
III. A general orientation network theory & methods

I: Concepts, structures, operations

Basic network terms

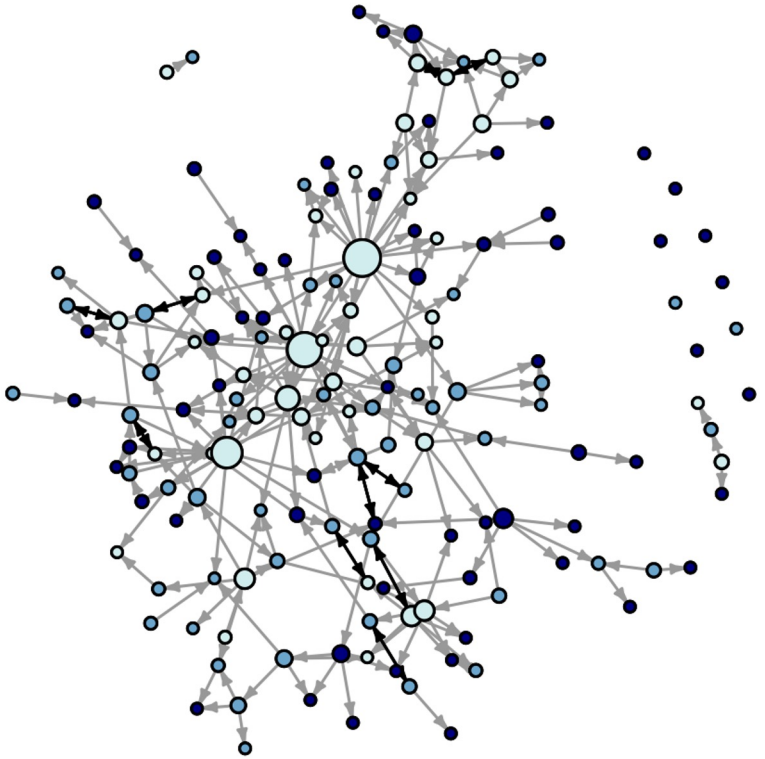
Nodes: or “vertices (vertex),” the entities in a network (e.g., individuals, households, places…)

Edges: the relationships or “ties” among entities (e.g., some quantity exchanged, a reported friendship)

- **Unweighted** (present/absent) or **weighted** (valued). Can be negative!
- Edges may also be directed (A → B ≠ B ← A) or undirected (usually drawn without arrowhead)

“**Ego**” often used to refer to a focal node and “**alter**” to refer to ego’s connections

Dyad refers to a pair of nodes; **triad** to a set of three



Data structures

Edgelist

Ego	Alter	Weight
A	B	1
A	C	2
B	A	1
…	…	…

Advantages: Compact, especially for sparse **graphs**

Disadvantage: Potential to lose **isolates**

Data structures

Matrix representations (a.k.a. sociomatrix, adjacency matrix). Nodes are rows/columns, edges are cell entries.

Diagonal cells represents ties to self: usually not allowed, so diagonal entries are (usually) zero

Less lossy than edgelists, but memory intensive for large networks

	A	B	C	D
A	0	1	2	0
B	1	0	3	0
C	2	3	0	0
D	0	0	0	0

Undirected networks have a symmetrical matrix (i.e., the top and bottom triangle are mirrored)

	A	B	C	D
A	0	1	2	0
B	1	0	3	0
C	0	2	0	0
D	0	0	0	0

Directed networks have an asymmetrical matrix representation

Network properties

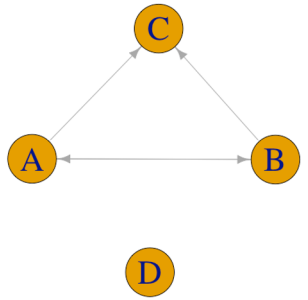
Density: proportion of all possible ties that exist

- Undirected possible ties = n*(n-1)/2
- Directed possible ties = n*(n-1)

Reciprocity: in directed networks, are both A → B and B ← A present? Yes/no at dyad level, often summarized as a proportion at the network level. Tricky in directed networks.

Transitivity: ratio of triangles to connected triples. This is sometimes also called the clustering coefficient.

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	0	0	0	0
D	0	0	0	0



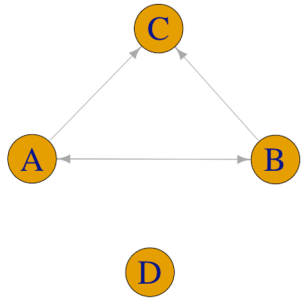
	Directed	Undirected (top half)
Density	4/12 = 0.33	3/6 = 0.5
Reciprocity	2/4 = 0.5	N/A
Transitivity	1	1

Centrality measures

Degree: the number of other nodes a node is connected to

- Variations for directed networks: in-degree, out-degree, Freeman degree (total number of unique alters, not necessarily equal to sum of in and out degree)

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	0	0	0	0
D	0	0	0	0



	In	Out	Freeman
A	1	2	2
B	1	2	2
C	2	0	2
D	0	0	0

Centrality measures

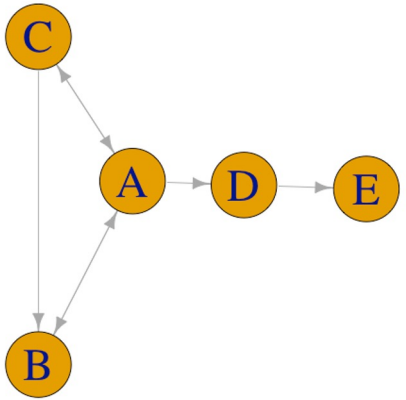
Betweenness centrality: number of shortest paths in a network that pass through a node.

Eigenvector Centrality: node centrality given by the eigenvector corresponding to the largest eigenvalue of the matrix. A node connected to many nodes who themselves have high scores will have a higher score.

Other centrality measures are mostly based on distance metrics (**closeness**: 1/lengths of shortest paths) or eigenvectors (**PageRank**, Katz, prestige score).

Cautions:

- Most measures are highly correlated with degree
- Isolates may have infinite/NaN values

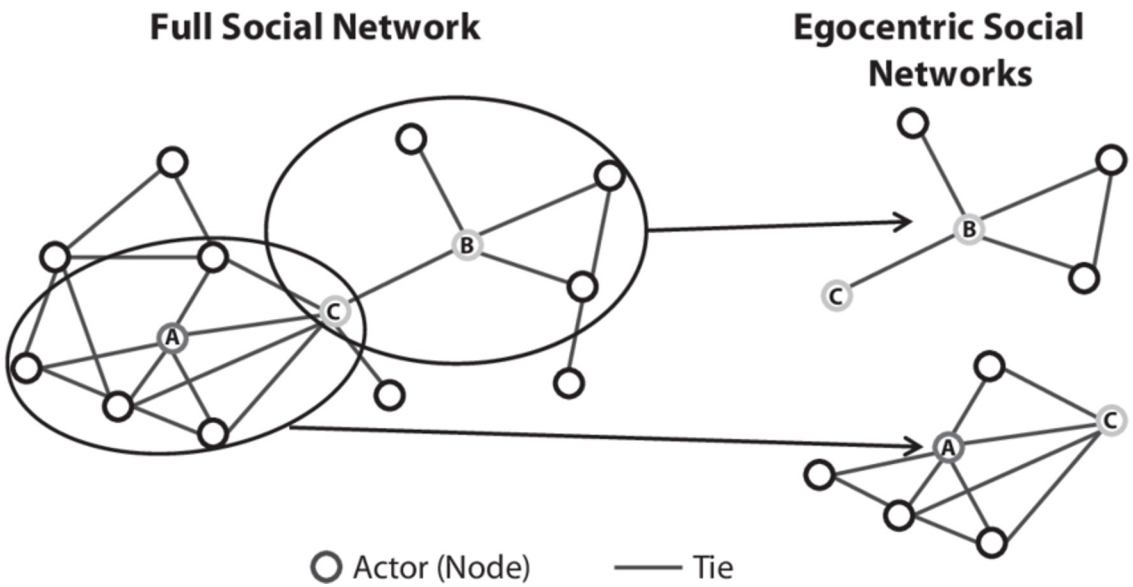


Node	N shortest paths
A	4
B	0
C	0
D	3
E	0

Ignoring direction!

Network types

Figure: Skiera et al. 2015



A **sociocentric network** represents the connections within an entire group.

- Setting group boundaries obviously a perennial problem

An **egocentric network** is the network surrounding a focal individual

Egocentric networks can be extracted from sociometric data; in some cases sociometric networks can be *inferred* from egocentric data.

Network types

Multimodal networks: multiple node "types," (e.g., people and organizations).

Bipartite networks, where nodes of type 1 associated to node(s) of type 2 are very common

- E.g., ties-by-association, or person by event data
- Can be represented in 2-mode form (P x E) or as a 1-mode projection (E x E or P x P), obtained through:

$$\mathbf{A} * \mathbf{A}^{\mathbf{T}} \text{ or } \mathbf{A}^{\mathbf{T}} * \mathbf{A}$$

P x E	Journal club	Dept. seminar	Xmas party
Sally	1	0	0
Sue	1	0	1
Sam	0	1	1
Stephen	0	1	1

P x P	Sally	Sue	Sam	Stephen
Sally	1	1	0	0
Sue	1	2	1	1
Sam	0	1	2	2
Stephen	0	1	2	2

Network types

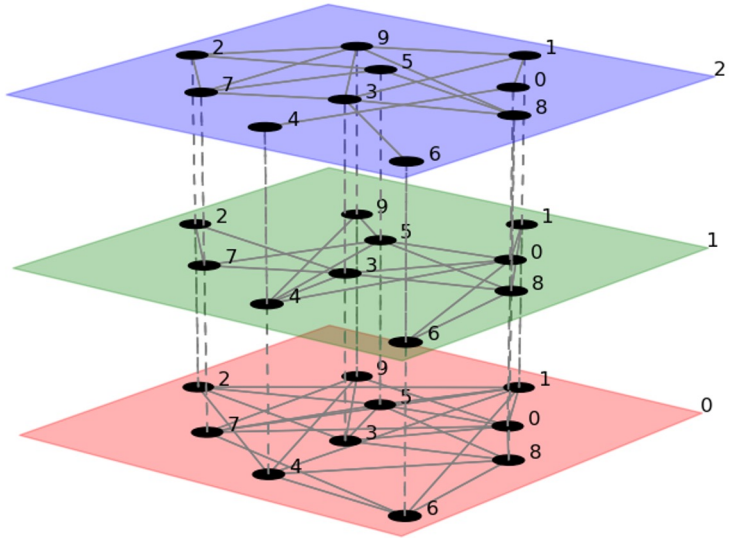
Multilayer/multiplex networks:

Multiplex (inter-layer edges connect the same nodes; e.g., time-slices of the same network)

In principle, interlayer edges could connect different nodes (e.g., food web)

Multi-relational networks (different types of edges) is usually what we mean

The same data can be represented many different ways, and the terminology is messy. **Describe your network well.**



Software (for manipulation/visualization)

The classics: UCINET (more anthro?), Pajek (more soc?)

R: igraph, statnet suite (network, sna), STRAND, tidygraph

*BUT often easiest to manipulate network data (as dataframes or matrices) in basic R.

Gephi: interactive visualization

Python: igraph, NetworkX

Further reading:

Wasserman and Faust. 1994. *Social Network Analysis: Methods and Applications.* Cambridge University Press. (SNA “Bible”)

Scott, J. 2017. *Social Network Analysis*, fourth edition. SAGE Publications. (Short & easy)

Borgatti, Everett and Johnson. 2018. *Analyzing Social Networks.* SAGE Publications.

Also: analytictech.com/networks/

*Note some analysis methods in these books out-of-date; but concepts/ideas pertinent

II: Overview of sampling strategies

Ego-networks

An **egocentric network** is the network surrounding a focal individual

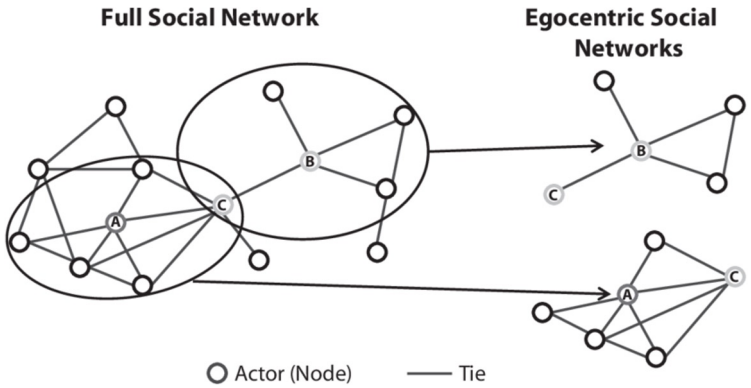
- Individual questionnaires: who do you interact with? (more later)
- Focal follows of an individual

Advantage: easier sampling!

Disadvantages: (limited) information on position of person and their contacts within broader network. Informant accuracy/recall.

A simple list of alters doesn't get you far

- Need to also document characteristics of alters
- Ideally, about interactions between alters (do X and Y know each other)?



Bott, E. 1955. Urban families: Conjugal roles and social networks. *Human Relations,* 8:4.Bidart, C. and Charbonneau, J. 2011. How to generate personal networks: Issues and tools for a sociological perspective. *Field Methods*, 23(3): 231–247.

Smith, J.A., 2020. The continued relevance of ego network data. *The Oxford Handbook of Social Networks*, p.170.

Respondent-driven sampling (RDS)

Similar to snowball sampling

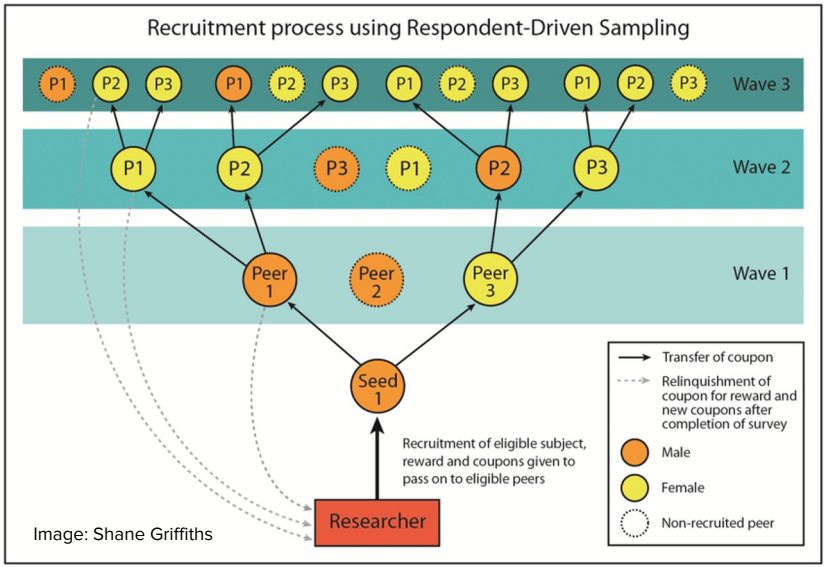
- Classic technique for “hidden” or “**hard-to-reach**” populations (e.g., drug users, sex workers)
- Most literature in epidemiology

Advantages: convenient,confidentiality, any data better than no data (e.g., for id-ing a problem)

Disadvantage: selection bias

Argued that ***under certain assumptions***, bias from the initial sample is attenuated wave by wave

For inference to broader networks: **To be used only under supervision of a trained professional.**



Gile and Handcock. 2010. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology,* 40:285–327.

Gile et al. 2018. Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application,* 5, 65-93.

Whole-network sampling

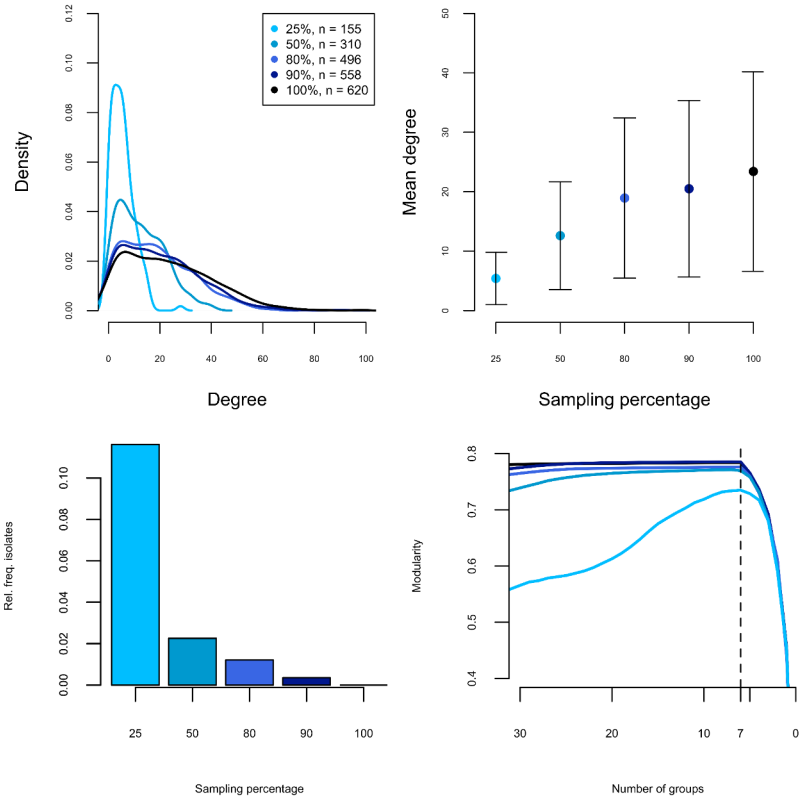
Attempt to sample all nodes in a network

- Boundaries of the relevant sphere can be artificial/difficult to define (what if the connections that interest you end up being beyond your sampling frame?!)
- Often there are practical size limitations

Consequences of missing data **depend on the network structure and research question**; there is no single rule for how much coverage is “enough.”

Are nodes **missing-at-random**?!

There are methods to impute missing network data but this is very difficult.



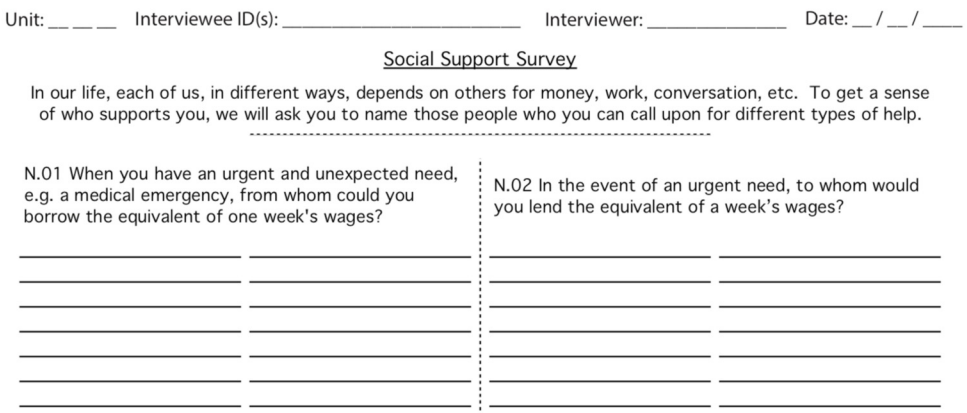
See: Smith, J. A., Moody, J., & Morgan, J. H. 2017. Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48, 78-99.

Eliciting network data: Name generators

Most common method in human studies.

Potential pitfalls:

- Informant recall (people don't remember their interactions well)
- Priming/repeating
- Question order effects (fatigue)
- Truncated number of responses (deliberately or **accidentally**)



Unit: __ __ __ Interviewee ID(s): _____ Interviewer: _____ Date: __ / __ / ____

Social Support Survey

In our life, each of us, in different ways, depends on others for money, work, conversation, etc. To get a sense of who supports you, we will ask you to name those people who you can call upon for different types of help.

N.01 When you have an urgent and unexpected need, e.g. a medical emergency, from whom could you borrow the equivalent of one week's wages?

N.02 In the event of an urgent need, to whom would you lend the equivalent of a week's wages?

A critical consideration: How will you **match the named alters** to generate the network?

- Usually need to collect some additional identifying information (age/gender, residence location…)
- In many places people have multiple names/nicknames

Eliciting network data: Name generators + roster

Sometimes you may have a full list of network members (e.g., in a small village or a classroom), so you can go through each potential alter with the respondent

Advantage: Improves accuracy/recall?

Disadvantage: Quickly becomes intractable. Fatigue, order effects (of Qs and alters) remain?

	Q1	Q2	Q3
Adam			
Abel			
Abraham			
Aaron			
Achim			
...			

Eliciting the network data: Observational data

E.g., classic ethological methods with humans and non-humans.

- Scan sampling
- Focal sampling

But also includes: social media data, GPS “ping” data.

Sampling is fraught with difficulties and potential biases due to limitations of our observation and due to organisms/behavior itself

What counts as a “tie”?
Observational data are often associations (i.e., individuals present at a event/location), i.e., bipartite data, and we assume that the co-association is meaningful.

How should interactions be counted?
Amount of time? Number of instances? …

How do deal with imbalanced sampling? E.g., differential visibility of individuals…

Altmann, J. 1974. Observational study of behavior: Sampling methods. *Behaviour,* 49:227–265.

Sampling: summary

Proceed with caution

Seeing something in the literature does not mean it's a good idea.

Sampling (patterns) of interactions AND nodes is hard.

One network representation cannot answer all questions about the importance of network structure in a system.

Unfortunately, researchers new to SNA often have questions that their data cannot answer

- Does not mean the data is useless
- Ego-network representations can still be very useful

Ideally, **use methods that account for the uncertainty** in the data

III: A general orientation to network theory & methods

What is your research question?

The basic conceit of SNA is that structure matters.

But there may be no true “social network” that exists that determines behavior; network data are just an abstracted representation of (a sample of) certain interactions or relationships.

The researcher has to make decisions about what the relevant “network” is to guide data collection/sampling, and modeling. What structure matters, and why does it matter?

- Past research can be a guide (but be careful!)
- But for many questions, researchers have to develop the reasoning of what aspects of a network are important and why
    - Draw on existing theory and suggest how to measure that structurally

What is your estimand?

Node-level

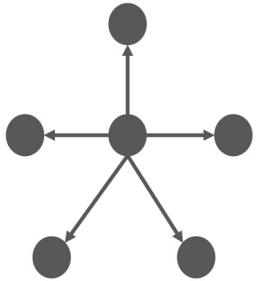
- Attribute-related outcomes (e.g., health, happiness)
- A feature of an individual's position within the network(e.g., in-degree, out-degree)

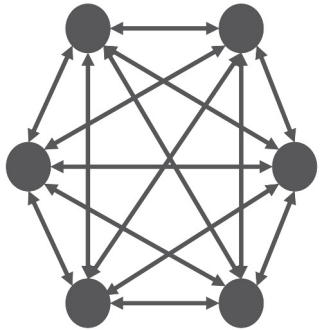
Dyadic

- Attribute-related outcomes (e.g., emotional closeness)
- Patterns of homophily (e.g., gender, ethnicity)

Higher-order

- Attribute-related outcomes (e.g., group success)
- Clustering & community detection
- Global network structure





Network position: Out-degree

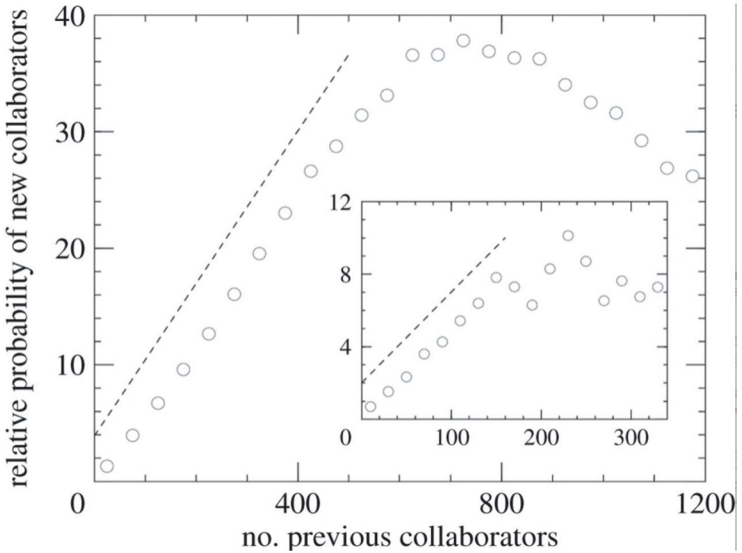
- In directed networks, the variation in the amount of ties that individuals ‘send’ may be important.

- Take, for instance, research questions about food sharing and provisioning of material support.
    - What individual attributes drive variation in out-degree? Could it be associated with the ability to provide support (e.g., status, wealth?

- *Variation* in out-degree is not the only feature to matter: individuals may occupy the *same* (structural equivalence) or *similar (*stochastic equivalence*) structural roles* based on their out-degree (or degree in undirected networks.
    - Equivalence may also be driven by other positional measures.

Network position: In-degree

- Understanding differences in in-degree distributions is a common question within social network analysis.
    - These questions often involve some individual attribute (e.g., social status).
- In-degree is often considered a metric of popularity or power within groups.
- Theory in network science highlights the role of '*preferential attachment*' in individuals' choices of who to form ties with (Barabási & Albert, 1999; Newman, 2001)
    - This is often referred to as the '*Matthew effect*' and represents the notion that '*success breeds success*' (e.g., academic collaboration networks).
- There are many models for generating networks characterised by different forms of preferential attachment.



Perc (2014)

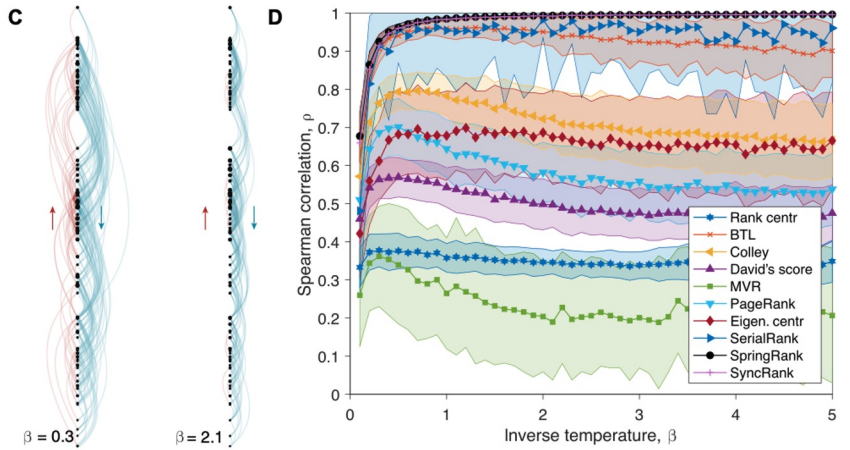
Network position: Betweenness, brokerage & shortest paths

- Small world networks are a popular idea of "*six degree of separation*".
  - The number of intermediary nodes (with connecting ties) or steps needed to connect any one arbitrarily chosen individual to any other.
  - Milgram's (1967) small world experiment.
  - The shortest path (or network diameter) is often smaller than expected. This could mainly be due to higher than random clustering within social networks (Uzzi, 1996).

- The individuals who sit within paths connecting two other individuals are thought to be *'brokers'* and have high '*betweenness centrality*' (Burt, 2007).
  - Many researchers use this form of centrality as a metric of power and brokerage of information.
  - However, be careful with betweenness!

More complex ideas about network position

- The features that bring about such power may be more complex than just counting an individual's out/incoming ties or betweenness.
  - These ideas stem from theory on social capital (Lin, 2000).
  - For instance, does the position of the individuals who that individual is connected with matter (e.g., the in-degrees or betweenness of their alters)?
- There are a great deal of centrality measures that have been derived–many of which capture these more complex dependencies.
  - It's important to implement the correct metric for your specific research question and context.



De Bacco, Larrimore & Moore (2014)

Reciprocity

- Many applied research questions are interested in understanding dyadic features within a network.
- An important feature is '*network reciprocity*'
  - In cross-sectional contexts this just the patterning on reciprocal ties within the sample.
  - Longitudinal and time-series data may be better able to describe the process of individuals being more likely to reciprocate ties.
- What reciprocity means depends on the research context.
  - For instance, in direct observations of food sharing networks, reciprocity may suggest that individuals are more likely to support those who have previously supported them.
  - In friendship networks, reciprocity metrics would signpost the number of reciprocated nominations (i.e., that both individuals consider one another to be friends). Capturing some similarity in the perceptions of relationships.

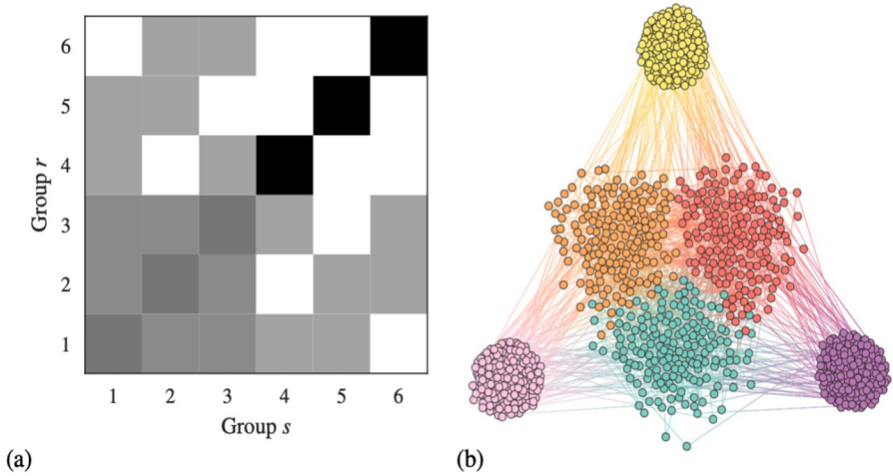
Homophily (& Heterophily)

- Individuals tend to create and maintain relationships with others who are the same, or similar to themselves, on a given attribute (McPherson, Smith-Lovin & Cook, 2001).
    - This could be, for instance, gender–which is often empirically observed within friendship networks.
- These notions of homophily are similar to ‘assortativity’ within the evolution of cooperation literature–with cooperators choosing partners who are also cooperators (e.g., Wang, Suri & Watts, 2012).
- In certain contexts, individuals may also prefer to make ties with those who are dissimilar to themselves.
    - This may, for example, be observed within collaborative teams–where there is complimentarity in attributes that promotes learning and improves payoffs.
- It is important to note, however, that the empirically observed patterns of similarity based on a potentially time-varying attribute (e.g., status, personality, obesity) may be caused by two different processes (Shalizi & Thomas, 2011; :
    - Individuals creating ties with those who are similar to themselves (‘*network selection*’ or *‘homophily’*).
    - Individuals becoming similar to one another because they share a ties (‘*network influence’* or *‘contagion’*).

Network structure: Groups with observed labels

- Building upon notions of homophily, observable grouping of individuals may guide the formation of social ties.
    - For instance, individuals who are the same ethnicity may form marriages, and certain ethnicities within a given sample may be more likely to marry than other ethnicities.

- Some of the most common generative network models–*Stochastic Blockmodels* (e.g., Holland, Laskey & Leinhardt, 1983; Karrer & Newman, 2011)–aim to capture this '*community structure*'.

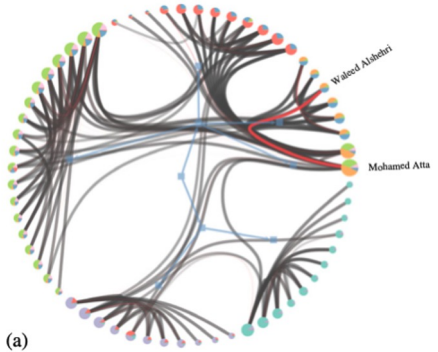


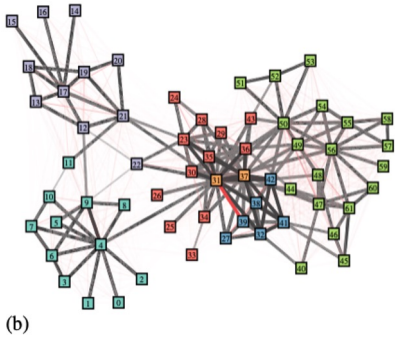
Piexoto (2019)

Network structure: community detection

- Where there is no observable group labels, community structure may be inferred through the patterning of observed ties.
  - This problem is often referred to as '*community detection*', and there are many algorithms do this (Lancichinetti & Fortunato, 2009).

- There are many applied research questions that require community detection.
  - For example, network scientists tried to determine the structure of communication between terrorists involved in the 9/11 attacks.





Piexoto (2018)

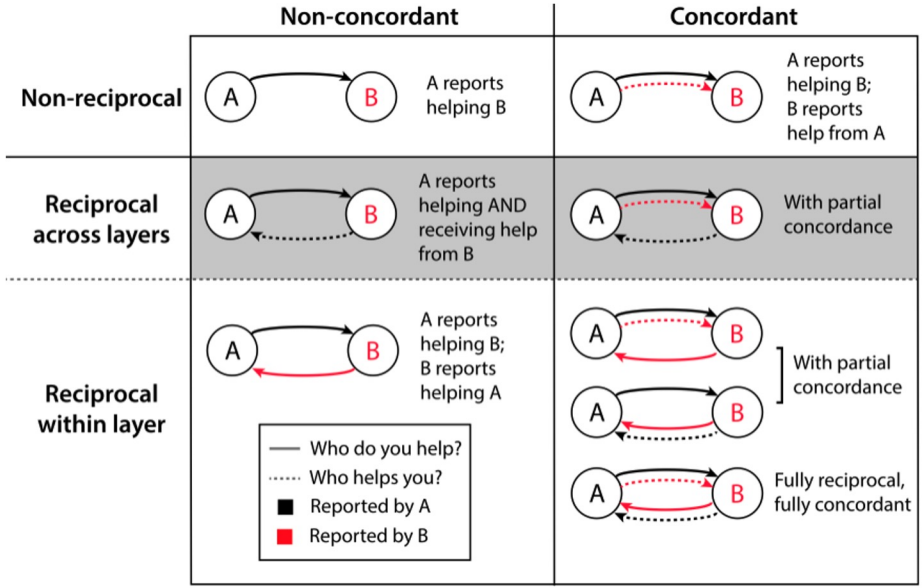
Applying SNA

- There are many statistical tools that are available for applied researchers to examine the research questions outlined above:
  - The social relations model (dyadic analysis): Kenny & La Voie, 1984.
  - Stochastic blockmodelling (community structure): Piexoto, 2018.
  - The combination of the two: Redhead, McElreath & Ross, 2022.
  - Exponential Random Graph Models: Robins, Pattison, Kalish & Lusher, 2007
  - Latent Space Models: Hoff, Raftery & Hancock, 2002.
  - Stochastic Actor-Oriented Models (Longitudinal network analysis): Snijders, Van de Bunt & Steglich, 2017.

- Choice in the types of models depends on the research questions, contexts and assumptions.

Revisiting data collection

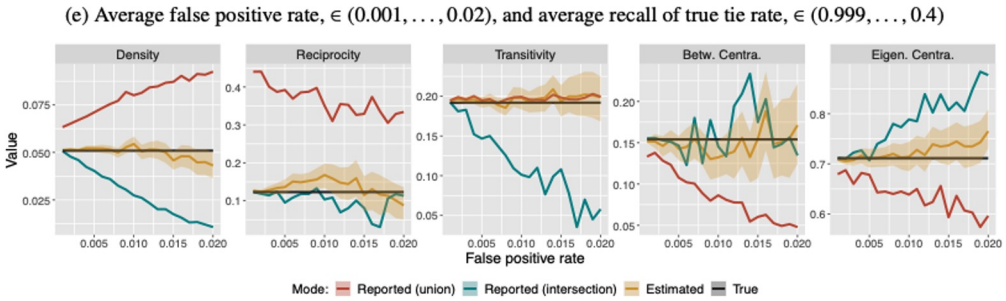
- Network data are commonly ‘double-sampled’.
  - This allows researchers to gain more information about every reported tie within a network.
- Double-sampling is important because individuals may not be reliable in their reports of their relationships:
  - They may falsely report ties.
  - They may forget ties.
  - The may duplicate their reports across prompts.
- But how do researchers treat and analyse double-sampled network data?
- There seem to be two standard approaches in the social sciences for dealing with double-sampled network data:
  - Taking the union of nominations: i.e., coding a tie as present if at least one individual reports it.
  - Taking the intersection of nominations: i.e., coding a tie as present if both individuals report it.



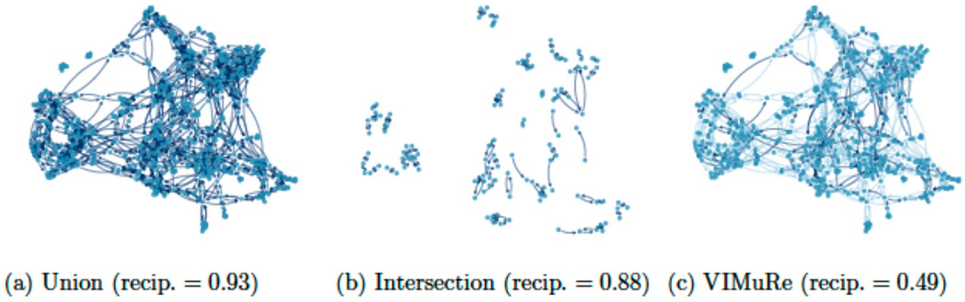
Ready & Power (2021)

Understanding & Incorporating measurement error

- Recent evidence suggests that these approaches to treating double-sampled social network data cause very different conclusions to be drawn:
  - The union of ties often over-estimates many global network features.
  - The intersection often under-estimates them.
- The networks resulting from these different approaches look incredibly different.



Redhead, McElreath & Ross (2022)



De Bacco et al. (2022)

Summary and takeaways

- Social network theory & methods provide powerful tools for answering a vast array of research questions.
- While there are a great deal of tools available for applied research, the choice of measures and metrics is really important.
- Everything is hard to do.

Further readings:

Bianconi, G. (2018). *Multilayer networks: structure and function*. Oxford university press.

Estrada, E., & Knight, P. A. (2015). *A first course in network theory*. Oxford University Press, USA.

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A first course in network science*. Cambridge University Press.

Newman, M. (2018). *Networks*. Oxford university press.

Recommended resources

Intro workshop materials:
eehh-stanford.github.io/SNA-workshop/

Rethinking Chapter 14; Lecture 15:
tinyurl.com/yc5fhy5n

We will post this and other material on a couple of different websites/repositories:

- elspethr.github.io
- https://github.com/danielRedhead
- https://github.com/ctross/STRAND
- https://github.com/ctross/DieTryin