LISS2108

Statistical Inference for Social Networks Analysis

Session 2. Permutation-based tests and autoregressive models

Santiago Quintero

santiago.quintero_suarez@kcl.ac.uk

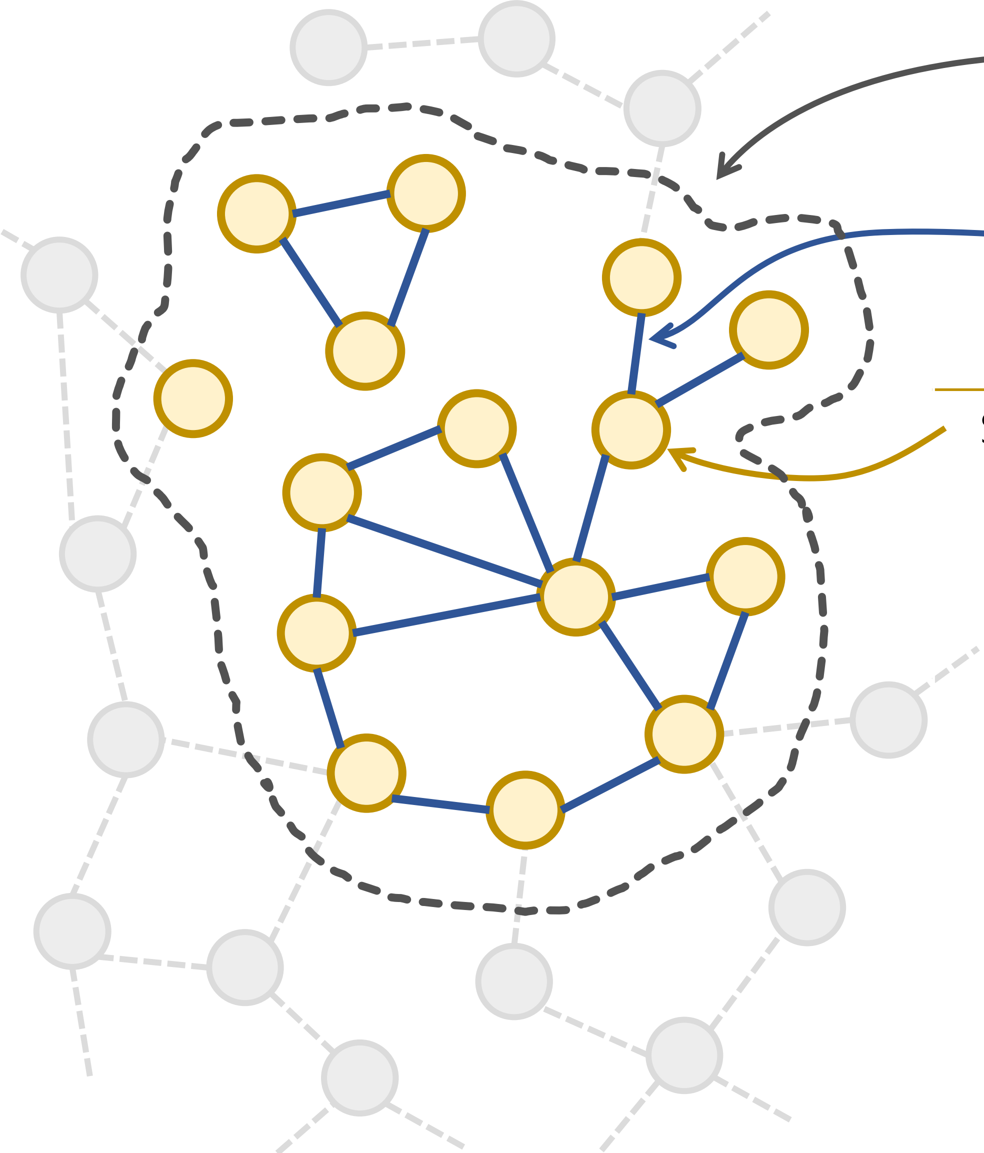
March 13th 2025

Session outline

1. Review: Descriptive Network Research
2. Explanatory Network Research
   - Networks as independent and dependent variables
   - Challenges of statistical inference in networks
3. Inference at the group level
4. Inference at the node level
5. Coding practice

1. Review

Representing a network



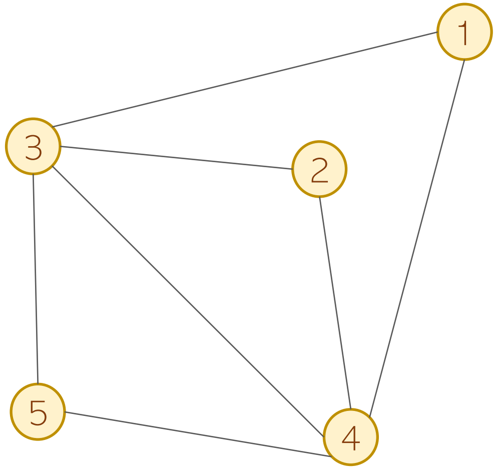
Nodes (vertices)	Ties (edges)	Boundary
Students / Academics	Advice / co-authorship	Classroom / Discipline
Politicians	Partisanship	Parliament
Countries	Alliances	A region / UN
Businesses	Trade	An industrial sector
Words	Used together	All tweets on a topic
Organisations	Employee exchange	A country

(Agneessens, 2023)

Representing a network

We are better off by getting used to thinking of networks as (adjacency) matrices!

- Total number of vertices: $\boldsymbol{n}$
- Total number of edges: $\boldsymbol{m}$
- Generic vertices: $\boldsymbol{i}$ and $\boldsymbol{j}$
- Generic edge: $\boldsymbol{(i, j)}$
- Adjacency matrix: $\mathbf{A}$ of size $\boldsymbol{n}$ by $\boldsymbol{n}$
- Particular edge in matrix: A_{ij}

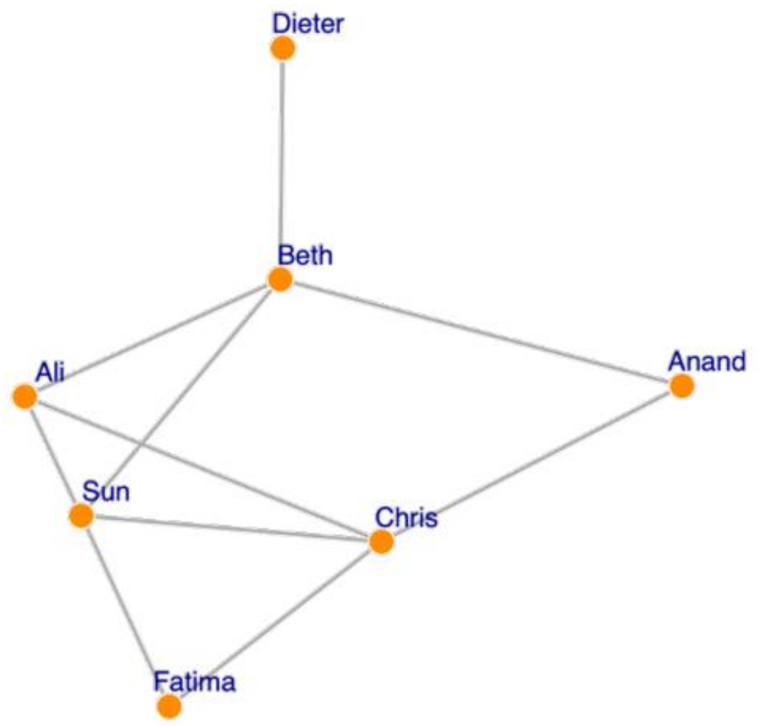


$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Representing a *undirected* network

We are better off by getting used to thinking of networks as (adjacency) matrices!

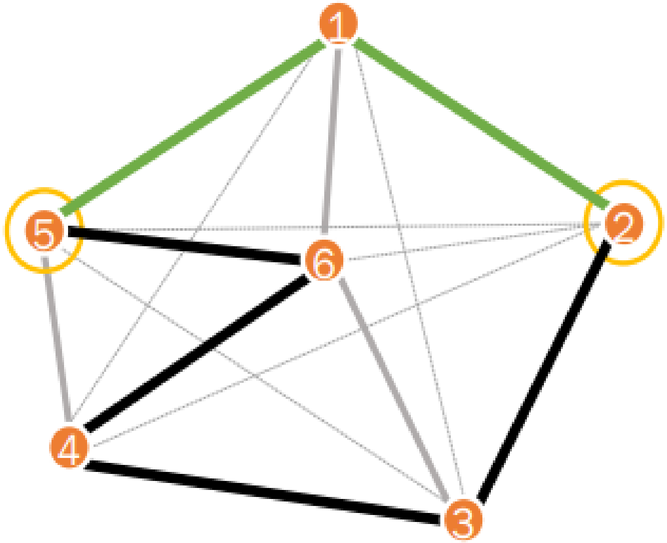
	Anand	Beth	Chris	Dieter	Sun	Ali	Fatima
Anand	0	1	1	0	0	0	0
Beth	1	0	0	1	1	1	0
Chris	1	0	0	0	1	1	1
Dieter	0	1	0	0	0	0	0
Sun	0	1	1	0	0	1	1
Ali	0	1	1	0	1	0	0
Fatima	0	0	1	0	1	0	0



(Power, 2021)

Describing networks

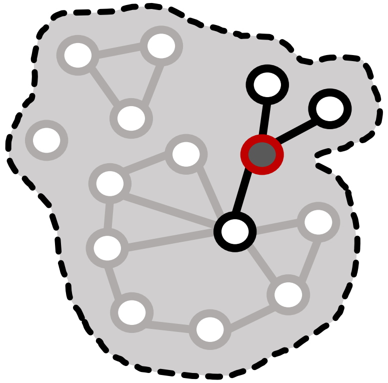
- Total number of vertices: n `vcount()`
- Total number of edges: m `ecount()`
- Degree : $k_i = \sum_{j=1}^{n} A_{ij}$ * `degree()`
- Mean degree: $c = \frac{2m}{n}$ * `mean(degree())`
- Density: $\rho = \frac{2m}{n(n-1)} = \frac{c}{n-1} \approx \frac{c}{n}$ * `graph.density()`
- Diameter: longest geodesic path length `diameter()`



*Different in directed networks

(Power, 2021)

Levels of analysis



Node level

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (direct connections)

How much a node fills a structural hole (broker)

How do node characteristics determine its position in the network

<u>Key focus:</u>
The position of an individual node on the network.
The relation of that individual node to other nodes

(Agneessens, 2020; 2023)

Levels of analysis

Key focus:
How to nodes are connected to each other

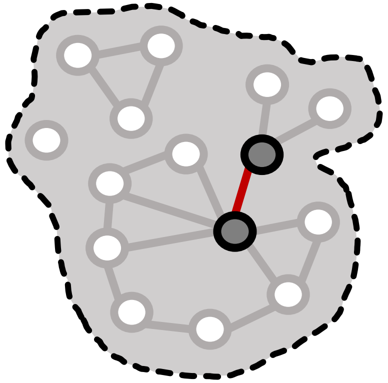
Dyad level

Presence or absence of a tie between two nodes

Distance between two nodes

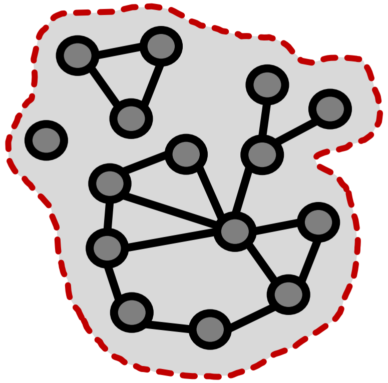
Number of common connections

Whether connected nodes are similar (homophily)



(Agneessens, 2020; 2023)

Levels of analysis



Group level

Density or centralisation

Number of cliques

Level of "core-peripheriness"

Level of average homophily in a group

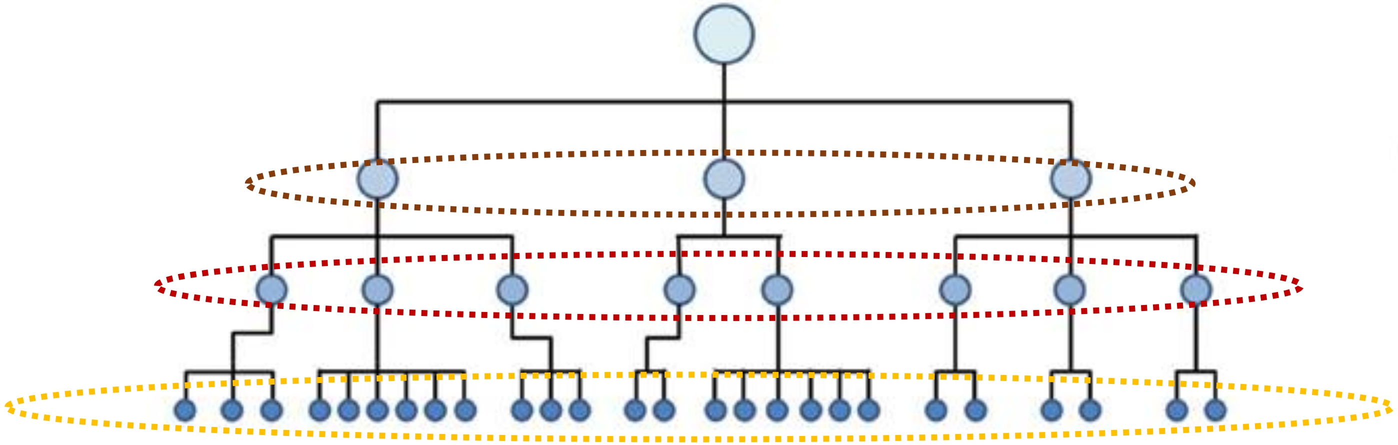
Key focus:
Structure of the group as a whole

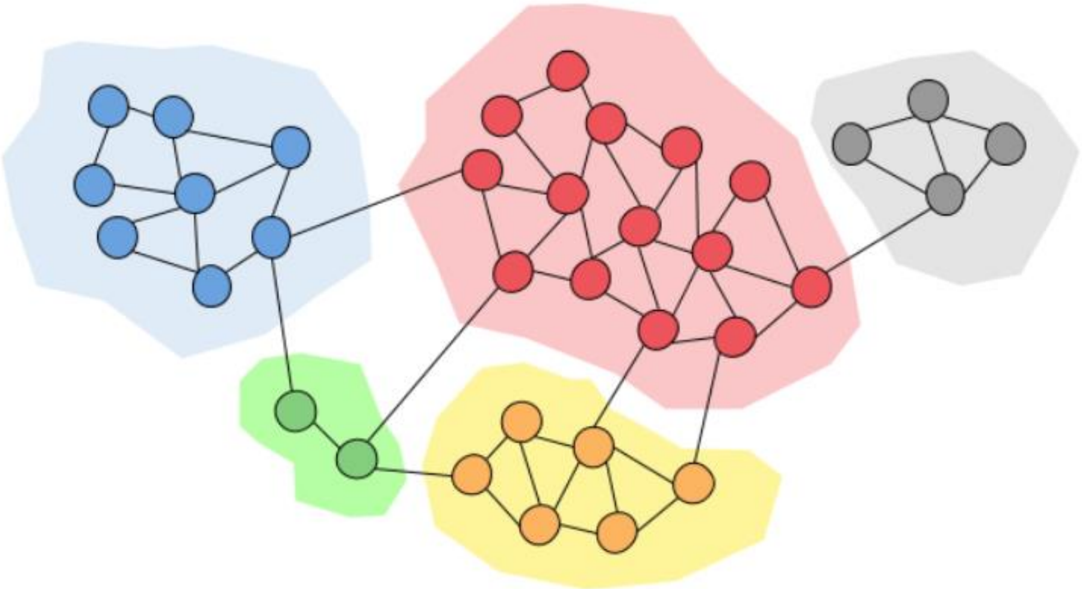
(Agneessens, 2020; 2023)

Describing networks

There are many ways to descriptively analyse a network

- Centrality measures (most important nodes?)
- Structural equivalence (are two nodes “structurally equivalent”?)
- Community detection (are there meaningful sub-partitions of the network?)





Describing networks

There are many ways to descriptively analyse a network

- Centrality measures (most important nodes?)
- Structural equivalence (are two nodes "structurally equivalent"?)
- Community detection (are there meaningful sub-partitions of the network?)

The main goals of *Descriptive Network Research* are:

1. Identifying the properties of a group (e.g., density or level of centralisation of a group; number of sub-communities)
2. Identifying the position of a node in the network (e.g., most central node; most powerful node)
3. Defining the relationship between two nodes (e.g., distance between nodes; similarities between dyads)

(Agneessens, 2020)

2. Explanatory Network Research

Explaining networks

Antecedents of networks
(Networks as *Dependent variable*)

Uncovering **why** specific nodal-, dyadic-, and group-level network properties **emerge**

Consequences of networks
(Networks as *Inependent variable*)

Uncovering the **effects** of such nodal-, dyadic-, and group-level network properties

(Agneessens, 2020; 2023)

Modelling networks

Antecedents of networks
(Networks as *Dependent variable*)

Group characteristics

- Group size
- Composition
- Formal structure

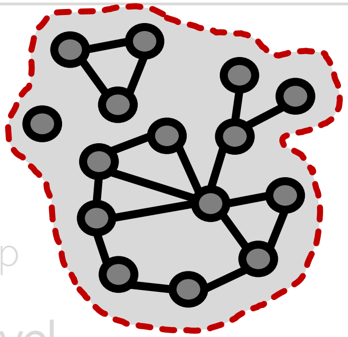
Group level

Density or centralisation

Number of cliques

Level of “core-peripheriness”

Level of average homophily in a group



Dyad level

Presence or absence of a tie between two nodes

Distance between two nodes

Number of common connections

Whether connected nodes are similar (homophily)

Node level

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (direct connections)

How much a node fills a structural hole (broker)

How do node characteristics determine its position in the network

Consequences of networks
(Networks as *Inependent variable*)

Group outcomes

- Performance
- Group culture
- Intervention effect

(Agneessens, 2020; 2023)

Modelling networks

Antecedents of networks
(Networks as *Dependent variable*)

Dyadic characteristics

- Similarity (homophily)
- Peer-effects
- Sorrounding

Group level

Density or centralisation

Number of cliques

Level of “core-peripheriness”

Level of average homophily in a group

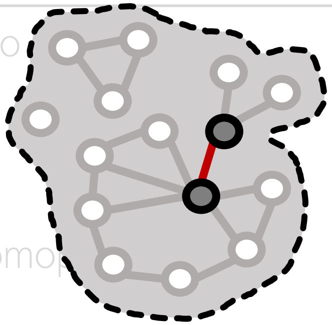
Dyad level

Presence or absence of a tie between two

Distance between two nodes

Number of common connections

Whether connected nodes are similar (homo



Node level

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (direct connections)

How much a node fills a structural hole (broker)

How do node characteristics determine its position in the network

Consequences of networks
(Networks as *Inependent variable*)

Dyadic outcomes

- Transfer

(Agneessens, 2020; 2023)

Modelling networks

Antecedents of networks
(Networks as *Dependent variable*)

Dyadic characteristics

- Similarity (homophily)
- Peer-effects
- Sorrounding

Group level

Density or centralisation

Number of cliques

Level of "core-peripheriness"

Level of average homophily in a group

We will discuss this level of analysis in the next to sessions

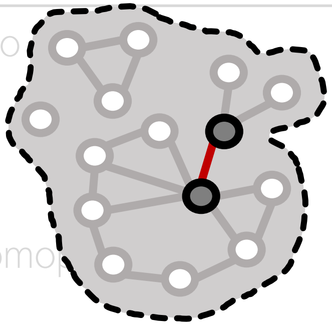
Dyad level

Presence or absence of a tie between two

Distance between two nodes

Number of common connections

Whether connected nodes are similar (homo



Node level

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (direct connections)

How much a node fills a structural hole (broker)

How do node characteristics determine its position in the network

Consequences of networks
(Networks as *Inependent variable*)

Dyadic outcomes

- Transfer

Modelling networks

Antecedents of networks (Networks as *Dependent variable*)

Consequences of networks (Networks as *Inependent variable*)

Group level

Density or centralisation

Number of cliques

Level of "core-peripheriness"

Level of average homophily in a group

Dyad level

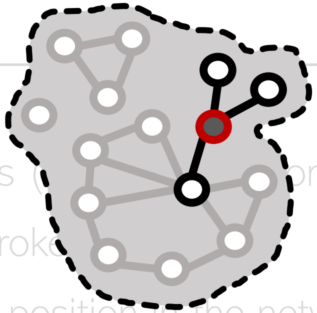
Presence or absence of a tie between two nodes

Distance between two nodes

Number of common connections

Whether connected nodes are similar (homophily)

Node level



Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (ons)

How much a node fills a structural hole (broke

How do node characteristics determine its position in the network

<u>Nodal (individual) characteristics</u>

- Age
- Values
- Reputation

<u>Nodal (individual) outcomes</u>

- Performance
- Beliefs
- Well-being

(Agneessens, 2020; 2023)

Modelling networks

1. Do groups with higher levels of identity cohesion mobilise more effectively?
2. Do local governments with higher levels of bureaucratic capacity engage more in collaborative policy delivery?
3. Are sleep patterns contagious?
4. Do pay scales affect the hierarchical structure of an organisation?
5. Do teams with more psychological safety share more information?
6. Do countries with strong trade relationships increase the likelihood of cultural assimilation?
7. What is the effect of friendships on mental health?

	Antecedents of networks (Networks = DV)	Consequences of networks (Networks = IV)
Group level	?	?
Dyadic level	?	?
Nodal level	?	?

What are the nodal units (vertices)?

What are the network relations (edges)?

What is the boundary

Modelling networks

1. Do groups with higher levels of identity cohesion mobilise more effectively?
2. Do local governments with higher levels of bureaucratic capacity engage more in collaborative policy delivery?
3. Are sleep patterns contagious?
4. Do pay scales affect the hierarchical structure of an organisation?
5. Do teams with more psychological safety share more information?
6. Do countries with strong trade relationships increase the likelihood of cultural assimilation?
7. What is the effect of friendships on mental health?

	Antecedents of networks (Networks = DV)	Consequences of networks (Networks = IV)
Group level	4, 5	1
Dyadic level	2, 5	6
Nodal level	3	7

What are the nodal units (vertices)?

What are the network relations (edges)?

What is the boundary

Challenges of statistical inference in SNA

1. Usually, network analysis are made on one network (N = 1)
    - Trying to extract general conclusions about social interactions or behaviours from N = 1 is difficult.
    - Sample ≈ Population

Challenges of statistical inference in SNA

1. Usually, network analysis are made on one network (N = 1)

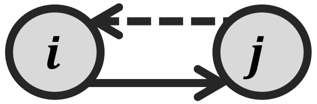
2. Data collection for network research is not random!
   - Inferential questions require us to explicitly formulate **hypotheses** and **test statistics** that operationalise them
   - Usually, the statistics used to **reject null hypotheses** (e.g., *p*-value) **depend on the sampling process**
   - Typical assumptions about sample distribution (e.g., **normality of the sample mean under CLT**) cannot be made

Challenges of statistical inference in SNA

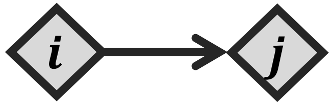
1. Usually, network analysis are made on one network (N = 1)

2. Data collection for network research is not random!

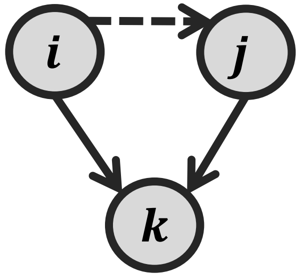
3. Observations in networks are not independent!
   - Networks exhibit many level of dyadic and triadic dependence
   - This might violate key assumption for identification (i.e., iid erros in classical regression framework)
   - We should be able to separate, identify, and control for them!



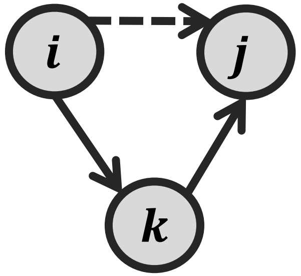
Reciprocity



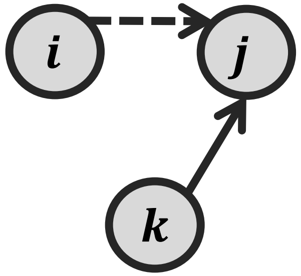
Homophily



Structural equivalence



Transitivity



Degree differentials

Choosing the right model

Antecedents of networks (Networks as *Dependent variable*)

Consequences of networks (Networks as *Inependent variable*)

Group level

Group characteristics

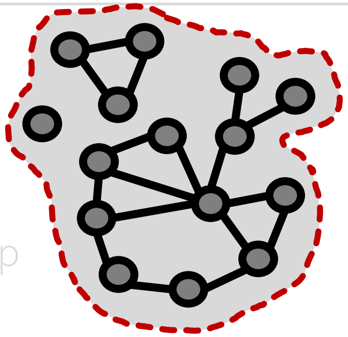
- Group size
- Composition
- Formal structure

Density or centralisation

Number of cliques

Level of "core-peripheriness"

Level of average homophily in a group



Group outcomes

- Performance
- Group culture
- Intervention effect

Node level

Nodal (individual) characteristics

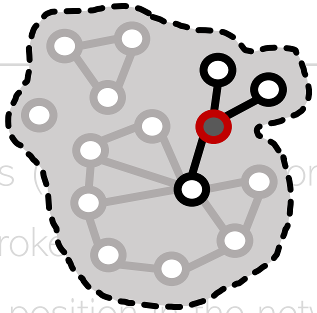
- Age
- Values
- Reputation

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (ons)

How much a node fills a structural hole (broke

How do node characteristics determine its position in the network



Nodal (individual) outcomes

- Performance
- Beliefs
- Well-being

(Agneessens, 2020; 2023)

3. Inference at the group level

Modelling at the group level

Antecedents of networks
(Networks as *Dependent variable*)

Group characteristics
- Group size
- Composition
- Formal structure

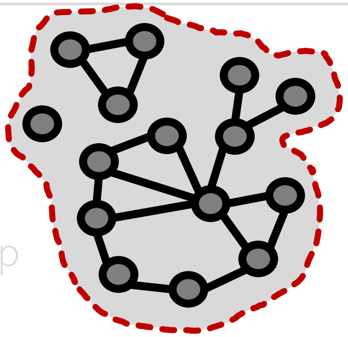
Group level

Density or centralisation

Number of cliques

Level of "core-peripheriness"

Level of average homophily in a group



Consequences of networks
(Networks as *Inependent variable*)

Group outcomes
- Performance
- Group culture
- Intervention effect

Examples:

- Does team communication enhances performance?
- Does demographic diversity increase group cohesion?
- Does group fragmentation hinder collective action?
- Does a more centralised friendship network improve the effects of health interventions in a classroom?

(Agneessens, 2020; 2023)

Modelling at the group level

Would classical statistical analyses (e.g., regressions) work here?

Examples:

- Does team communication enhances performance?
- Does demographic diversity increase group cohesion?
- Does group fragmentation hinder collective action?
- Does a more centralised friendship network improve the effects of health interventions in a classroom?

Modelling at the group level

Would classical statistical analyses (e.g., regressions) work here?

If we have a large enough (quasi)random sample of groups, and can demonstrate that "individuals" are not members of multiples groups: yes!

Examples:

- Does team communication enhances performance?
- Does demographic diversity increase group cohesion?
- Does group fragmentation hinder collective action?
- Does a more centralised friendship network improve the effects of health interventions in a classroom?

Modelling at the group level

Would classical statistical analyses (e.g., regressions) work here?

If we have a large enough (quasi)random sample of groups, and can demonstrate that "individuals" are not members of multiples groups: yes!

And what if we don't?

Examples:

- Does team communication enhances performance?
- Does demographic diversity increase group cohesion?
- Does group fragmentation hinder collective action?
- Does a more centralised friendship network improve the effects of health interventions in a classroom?

Modelling at the group level

Running example: Does a more centralised friendship network improve the effects of health interventions in a classroom?

Let’s say we run a health campaign to prevent smoking in 20 school classrooms:

- In a classical *Potential Outcomes framework*, we would compare the “treated” classrooms vs similar classrooms that were not treated (controlling for potential spillovers)
- But what if we also want ***to analyse if the structure of the friendship network in a classroom is a moderating factor of the intervention*** (e.g., if the more “popular” central nodes changed their behaviour, it is likely that other classmates did too)
- We also collected data on the friendship networks of the “treated” classrooms

Modelling at the group level

Running example: Does a more centralised friendship network improve the effects of health interventions in a classroom?

What prevents us from calculating a correlation coefficient or regressing the "intervention outcomes" on the level of network centralisation of each classroom (e.g., the highest degree in the network)?

Modelling at the group level

Running example: Does a more centralised friendship network improve the effects of health interventions in a classroom?

What prevents us from calculating a correlation coefficient or regressing the “intervention outcomes” on the level of network centralisation of each classroom (e.g., the highest degree in the network)?

- Sample size?
- No random sampling!
  - No clear distribution from which we can calculate the test statistic

Modelling at the group level

On popular option to model this: **permutation (or randomisation) test.**

Basic idea:

- We want to determine whether the **observed correlation is stronger than what would be expected by random change.**

Procedure:

- Compute the actual correlation between the two variables
- Randomly shuffle (permute) one of the variables while keeping the other (usually the network variable) fixed
- Recalculate the correlation for each permutation (e.g., 1000 or 10,000 times)
- Compare the observed correlation to the distribution of permuted correlations

Modelling at the group level

Procedure:

- Compute the actual correlation between the two variables
- Randomly shuffle (permute) one of the variables while keeping the network variable fixed
- Recalculate the correlation for each permutation (e.g., 1000 or 10,000 times)
- Compare the observed correlation to the distribution of permuted correlations

Significance testing: the p-value is *the proportion of permuted correlations that are as extreme as or more extreme than the observed correlation.*

Interpretation: If the observed correlation is rare under randomisation, it suggests a meaningful relationship rather than a spurious association.

Modelling at the group level

The empirical p-value is computed as:

$$p = \frac{\sum_{b=1}^{B} I\left(\left|r^{(b)}\right| \geq |r_{obs}|\right)}{B}$$

where $I(\cdot)$ is the indicator function that counts the number of permutations where the permuted correlation, $r^{(b)}$, is as extreme as or more extreme than the observed correlation.

- If p is small (e.g., $p < 0{,}05$), we reject the null hypothesis that the observed correlation is due to chance.
- Otherwise, we fail to reject the null, suggesting no strong evidence of a relationship.

4. Inference at the node level

Modelling at the node level

Examples:

- Does leadership influence the position of an employee in the hierarchy?
- Do countries with more trading partners increase their GDP?
- Do municipalities with less personnel contract-out more?

<u>Nodal (individual) characteristics</u>

- Age
- Values
- Reputation

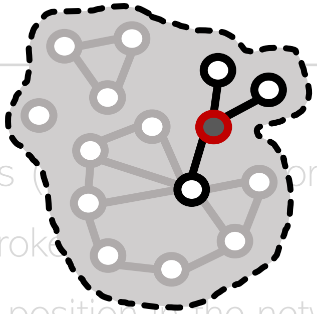
Node level

Degree, betweenness, closeness centrality

Level at which a node is similar to its alters (ons)

How much a node fills a structural hole (broke

How do node characteristics determine its position in the network



<u>Nodal (individual) outcomes</u>

- Performance
- Beliefs
- Well-being

Modelling at the node level

Examples:

- Does leadership influence the position of an employee in the hierarchy?
- Do countries with more trading partners increase their GDP?
- Do municipalities with less personnel contract-out more?

Would classical statistical analyses (e.g., regressions) work here?

Can we use our permutation-based method to derive consistent test statistics?

Modelling at the node level

Running example: Do countries with more trading partners have higher GDP?

Modelling at the node level

Running example: Do countries with more trading partners have higher GDP?

The problem now is that we might have **network autocorrelation**.

- It is likely that the GDP of trading partners or even number of trading partners my partner has can impact how much my GDP grows

In these cases we need to explicitly model the non-independence of the cases. For this there's a number of **autoregressive models** (CAR, SAR, VAR, etc.).

Basic idea:

- We add a new term (a weighted matrix) to our regression that accounts for the effect of the other node's values on the main dependent variable.

Modelling at the node level

We define our regression equation as:

$$Y = \ \beta_0 + \rho WY + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Where the term ρWY captures the other node's effect on each observations main outcome Y.

- W represents a weights matrix that captures the strength of the relationship between the nodes
- Y is the matrix with the outcomes for the rest of the observations.
- ρ captures the size of the effect

Modelling at the node level

We define our regression equation as:

$$Y = \beta_0 + \rho WY + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Where the term ρWY captures the other node's effect on each observations main outcome Y.

Running example: Do countries with more trading partners have higher GDP?

- In this case, ρWY captures the effect of the GDP of the trading partners.

5. Coding practice