Relational Phenomena

Mark S. Handcock

handcock@ucla.edu

Statistical Analysis of Networks
Based on Notes by Peter D. Hoff
September 26, 2024

Defining characteristics of network data

Relational data are data that include

- a set of objects, and
- measurements between pairs of objects.

Example (symmetric social network):

The canonical example of a relational dataset is the symmetric social network:

- $\{1, \ldots, n\}$ = labels of n individuals;
- $\{y_{i,j} : 1 \leq i < j \leq n\}$ = binary indicators of friendship between individuals

$$y_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are friends} \\ 0 & \text{if } i \text{ and } j \text{ are not friends} \end{cases}$$

$y_{ij} \equiv y_{ji}$

Components of network data

More generally, a network dataset consists of

- a set or multiple sets of objects:
  - **nodes** (objects, actors, egos, individuals)
- variables measured on nodes or pairs of nodes:
  - **dyadic variables**: measured on pairs of nodes (dyads)
  - **nodal variables**: measured on nodes

Standard data: nodes and nodal variables.

Relational data: nodes, nodal variables and dyadic variables.

The presence of dyadic variables is the defining feature of relational data.

Note:

One could expand the definition to triadic variables, but such data are rare.

Types of relations

Relations can be characterized as **undirected** or **directed**.

An **undirected** (or **symmetric**) relation has only one value per pair:

- $y_{i,j}$ measures the same thing as $y_{j,i}$;
- $y_{i,j}$ is equal to $y_{j,i}$ by design.

A **directed** relation has two values per pair: one value representing the perspective of each pair member.

- $y_{i,j}$ measures something different from $y_{j,i}$;
- $y_{i,j}$ may or may not be equal to $y_{j,i}$.

Types of relations

Undirected relations

- observer-reported friendships
- military conflict
- amount of time individuals spend together
- geographic distances between cities

Directed relations

- self-reported friendships
- military intervention of one country within another
- number of emails sent between individuals
- flight times between cities

Types of relations

Relations can be characterized as **binary** or **valued**:

- A **binary** (or **dichotomous**) relation takes only two values.
- A **valued** relation takes more than two values.
  - A valued relation whose possible values have an order is called **ordinal**.
  - A valued relation whose possible values lack an order is called **categorical**.

Most valued relations we will encounter are ordinal.

Binary relations

- presence of a friendship
- presence of a treaty between countries
- presence of connections between objects

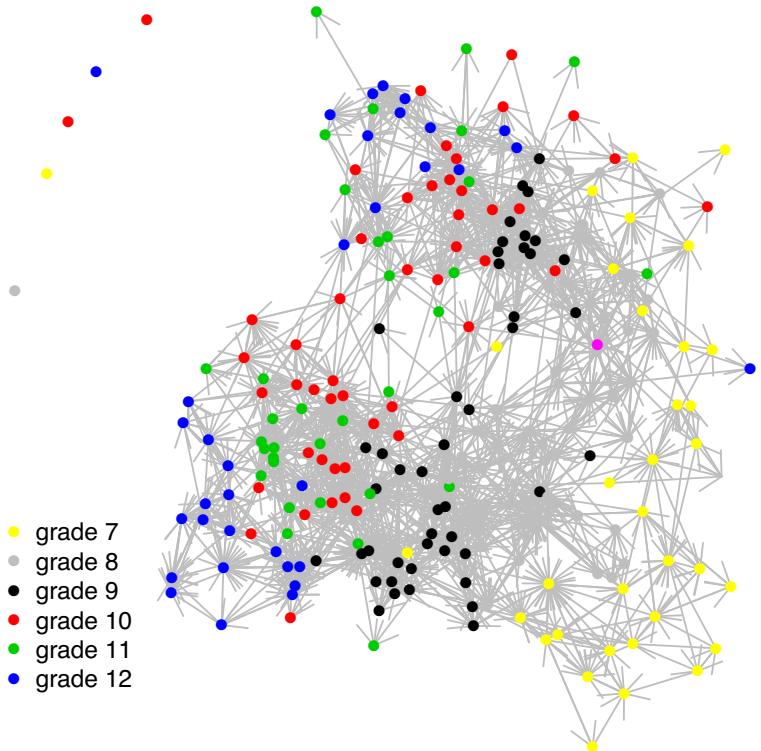
Valued relations

- amount of time spent together
- number of conflicts between countries
- economic transactions between entities

Example: Friendship nominations

AddHealth study: Survey from $n = 254$ students giving for each participant

- a rank-ordering their top-five male friends and top-five female friends;
- sex, ethnicity and grade-level;
- participation in extracurricular activities.

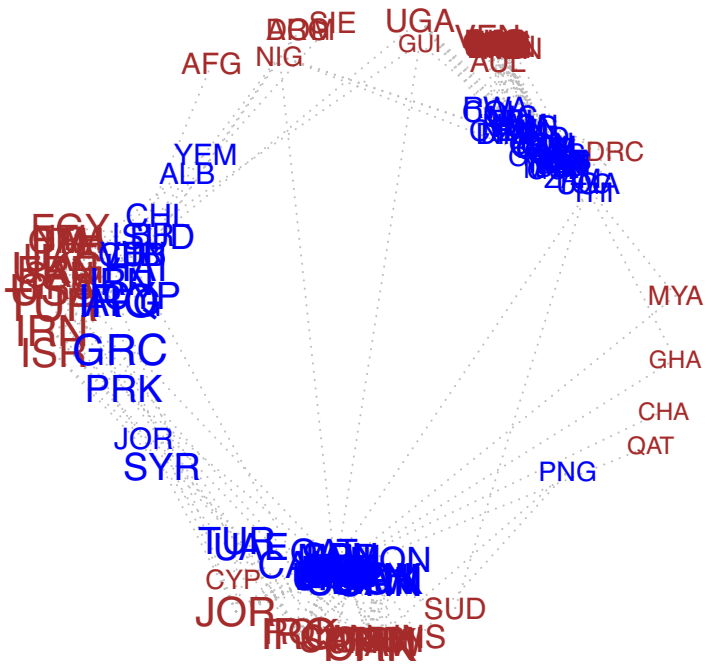


Exercise: Identify nodal variables, dyadic variables and type of relation.

Example (International relations):

The `conflict90s` dataset includes country data on

- population, GDP, and polity
- between-country trade, geographic distance, number of common IGOs, and the number of conflicts initiated by each node towards each other.



red = exports
blue = imports

IGO=Intergovernmental Organization

Exercise: Identify nodal variables, dyadic variables and types of relations.

Example (International relations):

Valued/dichotomous:

- valued: `imports`, `distance`, `shared_igos`, `conflicts`.
- dichotomous: none.

Directed/undirected:

- directed: `imports`, `conflicts`.
- undirected: `distance`, `shared_igos`.

Caution: `distance` and `shared_igos` are derived from nodal variables.

Such derived "relations" often have constraints that dyadic relations do not.

Matrix representation

Consider again the simplest case: a binary, undirected relation.

Suppose for a symmetric friendship relation among 5 individuals,

- person 1 is friends with persons 3, 4 and 5;
- person 2 is friends with person 4;
- person 3 is friends with person 5.

Then

- $y_{1,3}$, $y_{1,4}$, $y_{1,5}$, $y_{2,4}$, $y_{3,5}$ are all 1 (as are $y_{3,1}$, $y_{4,1}$, $y_{5,1}$, $y_{4,2}$, $y_{5,3}$);
- $y_{1,2}, y_{2,3}$ $y_{2,5}$, $y_{3,4}$ are all 0 (as are $y_{2,1}, y_{3,2}$ $y_{5,2}$, $y_{4,3}$).

$$\mathbf{Y} = \begin{pmatrix} na & y_{1,2} & y_{1,3} & y_{1,4} & y_{1,5} \\ y_{2,1} & na & y_{2,3} & y_{2,4} & y_{2,5} \\ y_{3,1} & y_{3,2} & na & y_{3,4} & y_{3,5} \\ y_{4,1} & y_{4,2} & y_{4,3} & na & y_{4,5} \\ y_{5,1} & y_{5,2} & y_{5,3} & y_{5,4} & na \end{pmatrix} = \begin{pmatrix} na & 0 & 1 & 1 & 1 \\ 0 & na & 0 & 1 & 0 \\ 1 & 0 & na & 0 & 1 \\ 1 & 1 & 0 & na & 0 \\ 1 & 0 & 1 & 0 & na \end{pmatrix}$$

Sociomatrices

More generally, any relational variable measured on a nodeset can be represented by a **sociomatrix**:

Sociomatrix: An square matrix with undefined diagonal entries.

A sociomatrix can represent a wide variety of relational data:

$$\mathbf{Y}_1 = \begin{pmatrix} na & 0 & 1 & 1 & 1 \\ 0 & na & 0 & 1 & 0 \\ 1 & 0 & na & 0 & 1 \\ 1 & 1 & 0 & na & 0 \\ 1 & 0 & 1 & 0 & na \end{pmatrix} \quad \mathbf{Y}_2 = \begin{pmatrix} na & 2.1 & na & 0.0 & 0.1 \\ 0.0 & na & 4.1 & 0.0 & na \\ 2.1 & 2.9 & na & 0.0 & 1.2 \\ 0.0 & 0.0 & na & na & 5.4 \\ na & 2.1 & 4.1 & 0.0 & na \end{pmatrix}$$

$\mathbf{Y}_1$ may be the sociomatrix of a binary, undirected relation.

$\mathbf{Y}_2$ is the sociomatrix of a valued, directed relation with missing values.

Sociomatrices

Dichotomous/valued

- A **dichotomous** relation is represented with a **binary** sociomatrix.
- A **valued** relation is represented with a **valued** sociomatrix.

Undirected/directed

- An **undirected** relation is represented with a **symmetric** sociomatrix.
- A **directed** relation is represented with a possibly **asymmetric** sociomatrix.

relation	sociomatrix
undirected dichotomous	symmetric binary
undirected valued	symmetric valued
directed dichotomous	asymmetric binary
directed valued	asymmetric valued

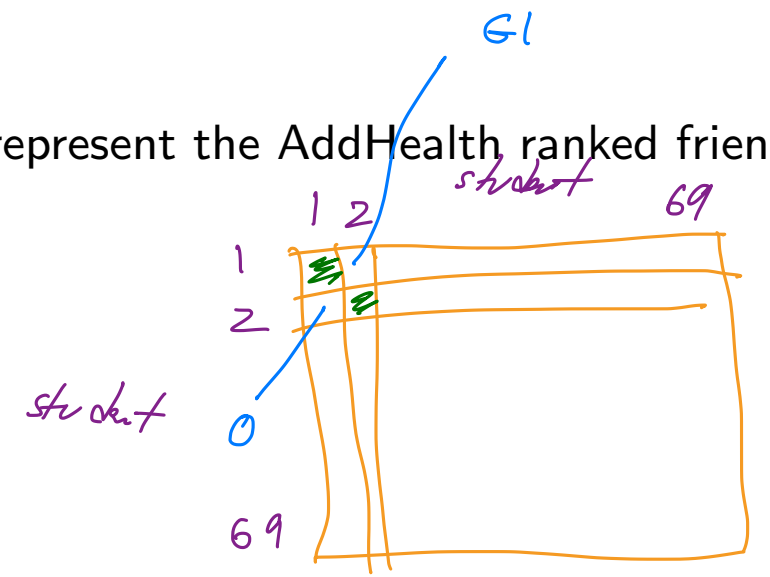
Missing data

Completely/partially observed: A relation is

- **completely observed** on a nodeset if only the diagonal is missing;
- **partially observed** on a nodeset if other entries are missing.

Exercise:

How would you represent the AddHealth ranked friendship data in terms of a sociomatrix?



Related non-relational data

Several data types share some features with relational data:

- Affiliation networks
- Dyadic data
- Correlation networks/graphical models

Some network researchers contort these into what seems like relational data.

However,

- these data types are structurally different from relational data, and
- generally require different methods for analysis.

Affiliation networks

International relations data: The number of shared IGOs is a dyadic variable, but it is derived from a set of binary attribute variables:

	org 1	org 2	org 3	$\cdots$
country 1	0	1	0	0
country 2	0	1	1	0
country 3	1	0	1	0
$\vdots$				

To many, such data would simply be **multivariate binary data**.

To people who see everything in terms of a network, such data is sometimes called an **affiliation network**.

Exercise:

Compute the number of shared affiliations between nodes from an affiliation matrix.

Affiliation networks

Affiliation network: A relational network between two sets of nodes for which

- there is no overlap between the two node sets;
- there are no relationships within a node set.

Examples:

- Movie ratings: relations between people and movies
- Consumer data: relations between people and products
- Animal behavior: presence at various locations and/or points in time

Terminology:

- **modes**: (generally) non-overlapping groups of objects
- **bipartite relation**: a relationship defined on pairs consisting of one member from each of two modes.

Affiliation networks are sometimes referred to as **two-mode bipartite** networks.

(The term "bipartite" is borrowed from graph theory.)

Affiliation networks

Affiliation networks may seem like relational data:

- they may involve relations between people;
- they may consist of "dyadic" measurements on pairs of objects.

They differ from relational data as previously defined:

- within-mode ties are structurally impossible;
- many features of one-mode networks are meaningless for affiliation networks (reciprocity, transitivity);
- they can be treated simply as matrix-valued or multivariate data, for which there is a large and separate statistical literature.

For these reasons, we will not spend much time on bipartite networks.

Dyadic data

For this course, a better term than relational data would be dyadic data: Recall,

Defining feature of relational data:
One or more variables measured on dyads, that is, **dyadic data**.

Unfortunately, this term is already in use: According to one definition, dyadic data arises when

> "each person is linked to one and only one other person in the sample and both persons are measured on the same variables."
> (http://davidakenny.net/dyad.htm) .

Examples include

- GPA measured on pairs of roommates;
- educational attainment of mother-daughter pairs;
- twin studies.

Exercise: Consider the relation "has ever been married to". Discuss situations where data on this relation may be dyadic, bipartite, or relational.

Correlation networks

Suppose a set of several variables are measured on a population of subjects:

subject	v_1	v_2	v_3
1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
4	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$
$\vdots$		$\vdots$	

Such data are typically referred to as **multivariate data**.

From such data, people often form a correlation matrix:

$$\mathbf{C} = \begin{pmatrix} 1 & c_{1,2} & c_{1,3} \\ c_{2,1} & 1 & c_{2,3} \\ c_{3,1} & c_{3,2} & 1 \end{pmatrix}$$

From this, one may compute the partial correlation matrix:

$$\mathbf{P} = \begin{pmatrix} - & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & - & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & - \end{pmatrix}$$

Here, $\rho_{i,j}$ is the correlation between i and j "controlling" for other variables.

Correlation networks

Suppose a set of several variables are measured on a population of subjects:

subject	v_1	v_2	v_3
1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
4	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$
$\vdots$		$\vdots$	

Such data are typically referred to as **multivariate data**.

From such data, people often form a correlation matrix:

$$\mathbf{C} = \begin{pmatrix} 1 & c_{1,2} & c_{1,3} \\ c_{2,1} & 1 & c_{2,3} \\ c_{3,1} & c_{3,2} & 1 \end{pmatrix}$$

From this, one may compute the partial correlation matrix:

$$\mathbf{P} = \begin{pmatrix} - & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & - & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & - \end{pmatrix}$$

Here, $\rho_{i,j}$ is the correlation between i and j "controlling" for other variables.