Network data analysis

MDT - Master Digital Transformation

M.Francesca Marino

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università di Firenze

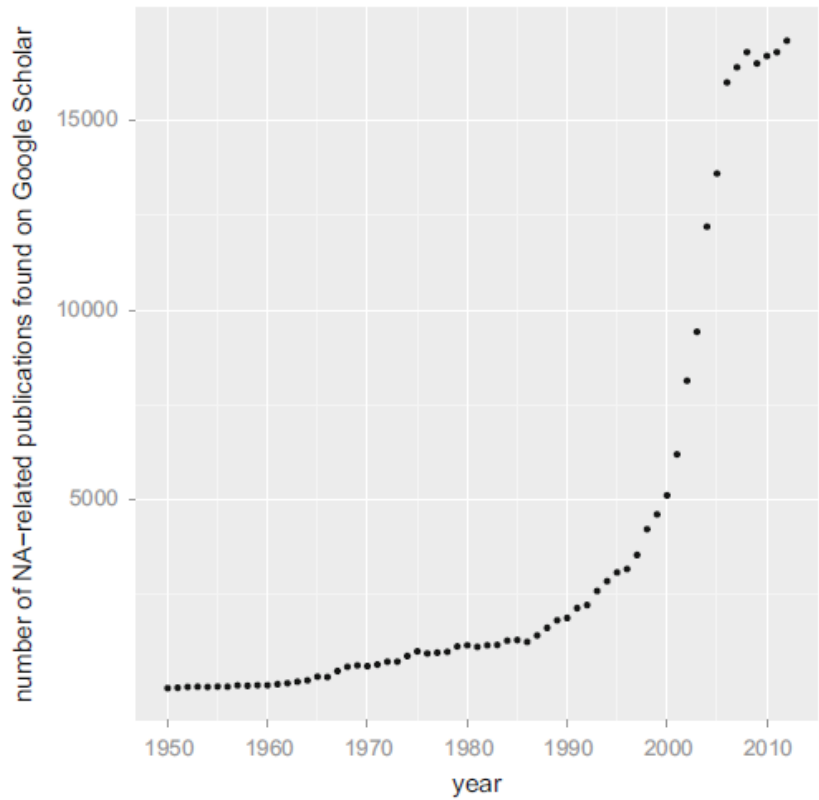
Table of contents

Intro

Why a module on Network Analysis?

- Networks are prominent in a number of social, business, economic, political, etc.., areas
- Everyone is entangled e.g. in various friendship and working relationships which can be naturally represented as networks
- The prospect of understanding the complex and erratic system behind such networks is an exciting one
- Similarly, trying to identify the main entities/actors in the complex system at hand, is surely interesting for scientists
- Network analysis seems to be one of the most promising frameworks within which these aspects can be combined, analyzed, and maybe even be understood

- This vast applicability of network analysis has led to a tremendous interest in methods provided by such a discipline
- There has been a dramatic increase of the number of articles with the keywords “network analysis” or “complex networks” as found by Google Scholar



Number of articles published in the given year containing the exact phrases “network analysis” and “complex network” as given by Google Scholar on the 12th of October, 2013

- Starting from about 100 articles (as found on Google scholar) in the 1950s, in 2013 more than 15,000 articles appeared
- And such a number has increased even more in the last 7 years
- This also happened thanks to the vast availability of data and computational power to treat them

The data

The data

- The first question you might have is: *what kind of data can actually be meaningfully represented as networks?*
- A first answer is: almost any kind of data
- The basic requirement is that there is a distinct set of *entities*, e.g., humans, employees, parties, organizations, computers, books, routes, etc..
- The second requirement is that there is a known *relationship* between these entities

- The information of whether any two entities are in the given relationship or not needs to be known for a large part of them
- Otherwise, any kind of analysis will be quite misleading
- Some obvious relationships between people are: friendship, kinship, employee-employer-relationships, etc..
- Another interesting type of relationship is membership
- In this case, it entails two different kinds of entities, e.g. people and institutions, actors and films, employees and companies, etc..

- Relationships between non-human entities are equally abundant
- Some examples include metabolic networks, protein-protein-interactions, neural networks, street networks, train connections, computer networks, company networks, countries exchange networks, etc..
- All of these examples might be considered ‘natural networks’
- Are there more abstract relationships that can also be represented as complex networks?

- Mathematicians have a more general understanding about what is a relationship and what is not. They talk about *relations*
- In a mathematical sense, two books can be defined to be "related" because their cover was created by the same designer
- This "relatedness" does not mean that they are necessarily related in any colloquial sense: their content can be very different!
- In detail, mathematically a *relation* R on a given set of entities O is just an arbitrary choice of pairs of these entities

$$R \subseteq O \times O$$

- On one side, this concept is much more general than the day-to-day notion of a relationship but, on the other hand much less intuitive
- A *relation* does not necessarily imply a real-world *relationship*
- It can even represent a relationship that would not be seen as meaningful in the real world
- For example, all humans with the same first name or all humans which share the same last digit of their ID-card number can be represented by a relation
- Mathematical relations can be also (meaningfully) derived from other relations
- One can build a second network based on the connection structure of another network by, for example, connecting two people if they share at least 8 friends in a friendship network

- So, a relationship is something that can be observed in the real world
- On the other hand, a relation is the mathematical structure which possibly represents a relationship
- Not all relations need to be necessary associated with any relationship
- The same relation, i.e., the same subset of pairs of a given set of entities O, can even represent different relationships

Formalizing network data

How are relations turned into network data?

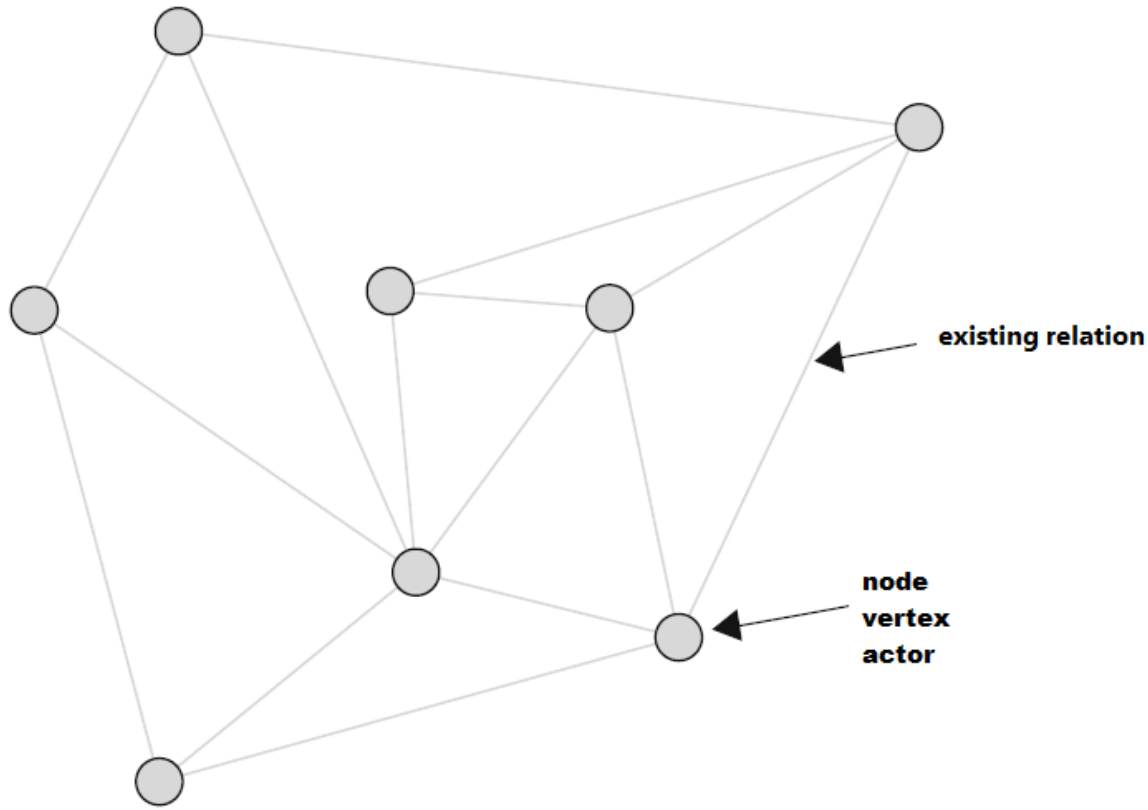
- When the entities and a relationship of interest have been identified, a range of decisions are required to turn the concept of relationship into a procedure that decides, for each pair of entities, whether they are mathematically related or not
- In most cases, when data are turned into a network representation, several decisions have to be made
    - Does the relationship of interest contain a direction?
    - Is it necessary to include this information in the mathematical relation?
    - Are there different levels of intensity of a given relationship?
    - Is it necessary to differentiate between them, by assigning weights to the pairs in the relation?
- Each of these decisions changes the set of available tool of analysis and the interpretation of the corresponding results

Formally, network data consist of

- *Nodes (Entities)* $\rightarrow \{1, \ldots, n\}$
  - Represent the object of the analysis
  - Also known as actors, vertices, units, or individuals
- *Dyadic (Tie) variables* $\rightarrow Y_{ij},\ i, j = 1, \ldots, n$
  - Measure the existing relation between nodes i and j
  - Identify existing ties (or edges, or links, or arcs) between nodes
- *Attributes* $\rightarrow \boldsymbol{x}_i, \boldsymbol{x}_{ij}, i = j = 1, \ldots, n$
  - Attributes characterizing node $i \rightarrow$ *nodal attributes*
  - Attributes characterizing the tie variable $Y_{ij} \rightarrow$ *relational attributes*

While **nodes and dyadic variables are essential components** for defining a network, **attributes are not**

Network visualization



NB. Representing interactions becomes progressively cumbersome as the number of nodes and/or relations increase

Types of relations: **directed** and **undirected**

- A relation is said to be *undirected* if a connection from node i to j *directly implies* a connection from j to i (symmetric connections only)

$$Y_{ij} = Y_{ji}$$

that is, Y_{ij} and Y_{ji} measure the same thing and are equal by design

- A relation is said to be *directed* if a connection from node i to j *does not directly imply* a connection from j to i (symmetric and asymmetric connections)

$$Y_{ij} = Y_{ji} \quad \text{or} \quad Y_{ij} \neq Y_{ji}$$

that is, Y_{ij} and Y_{ji} measure different things and may or may not be equal

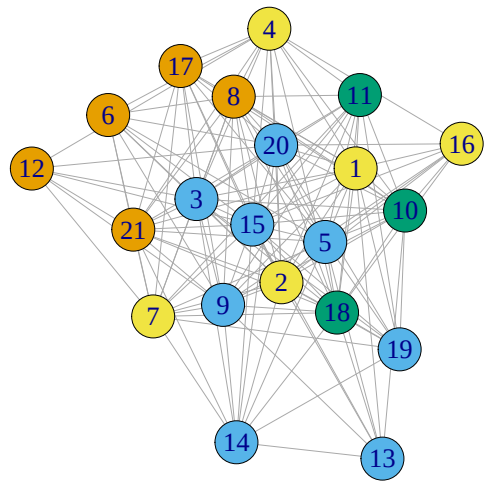
Some examples

- Undirected relations
  - Are *Mario* and *Francesco* friends on Facebook?
  - Are computer A and B connected?
  - How many conflicts have occurred between Germany and USA?
- Directed relations
  - Is *Mario* following *Francesco* on Instagram?
  - Which is the amount of goods Italy imports from Germany?
  - How many emails has *Mario* sent to *Francesco*?

A real-data example: advice network

21 employees from a high-tech company

- Did node *i* and *j* exchange advice to *each other* in the last week?
- Department and age of each employee

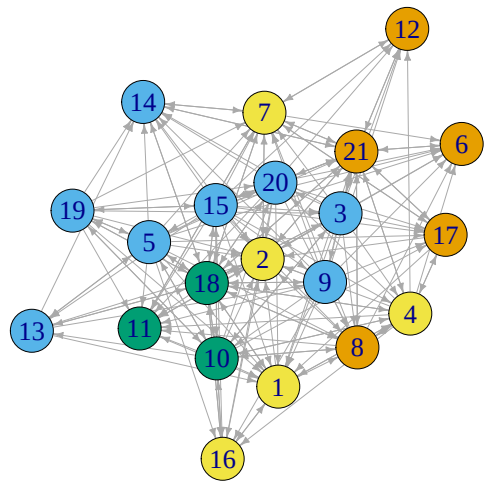


Exercise: identify tie variables, nodal variables, and the type of the relation

A real-data example: advice network

21 employees from a high-tech company

- Did node *i* ask advice *to* node *j* in the last week?
- Department and age of each employee



Exercise: identify tie variables, nodal variables, and the type of the relation

Types of relations: **binary** or **valued**

- A relation is said to be *binary* if tie variables Y_{ij} take only two values
  - These values correspond to the presence or the absence of a connection between pairs of nodes
  - Networks measuring binary relations are also called *unweighted networks*
- A relation is said to be *valued* if tie variables Y_{ij} can take more than two values
  - These values correspond to interaction strength, frequency of interconnections, capacity of interaction, etc...
  - Networks measuring valued relations are also called *weighted networks*

Some examples

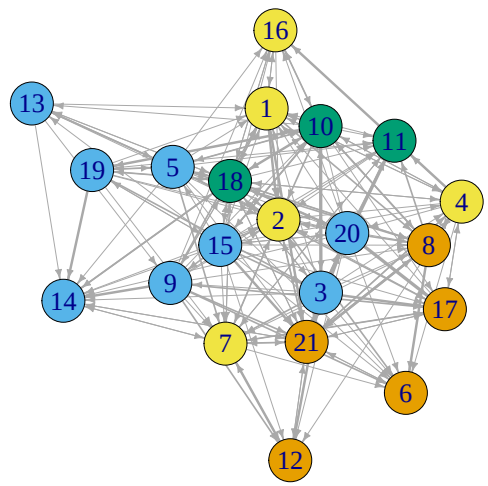
- Binary relations
  - Are *Mario* and *Francesco* friends?
  - Is *Mario* following *Francesco* on Instagram?
  - Are computer A and B connected?

- Valued relations
  - Which is the amount of goods Italy imports from Germany?
  - How many conflicts have occurred between Germany and USA?
  - How many emails has *Mario* sent to *Francesco*?

A real-data example: advice network

21 employees from a high-tech company

- How many times did node i ask advice *to* node j in the last week?
- Department and age of each employee



Exercise: identify dyadic variables, nodal variables, and the type of the relation

Formalize relational or network data

To *formalize* a **binary relation** from a mathematical point of view, we may distinguish two different approaches

- a *graph* representation
- a *matrix* representation

A *graph* $G = (V, E)$ is a *mathematical structure* consisting of

- a set V of *vertices* (nodes) with labels

$$\{v_1, v_2, \ldots, v_n\}$$

- a set E of *edges* (or links or ties) with elements

$$\{e_1, e_2, \ldots, e_m\}$$

Each element $e \in E$ is a pair of vertices $\rightarrow e = (i, j)$ related each other

Undirected graphs

- Undirected graphs allow us to formalize *undirected relations*
- In this case, **edges** e in the edge set E **have no** direction and the edge (i, j) **is equal** to the edge (j, i)

$$(i, j) = (j, i)$$

- That is, each edge e denotes an *unordered pair of nodes*

Directed graphs (Digraphs)

- Directed graphs allow us to formalize *directed relations*
- In this case, *edges* e in the edge set E **have** a direction and the edge (i, j) **is not equal** to the edge (j, i)

$$(i, j) \neq (j, i)$$

- That is, each edge e denotes an *ordered* pair of nodes

Storing graphs: edge list

- Graphs are often stored on a computer in terms of their edge set E. This is typically known as *edge list*
- The edge list completely represents the graph, unless there are *isolated nodes*, that is nodes that are not connected with other nodes
- In this case, the binary network is completely represented by an edge list and a list of isolated nodes

Undirected binary relations

An easier way to represent network data is via the *adjacency matrix*

$$\boldsymbol{Y} = \begin{pmatrix} \textbf{na} & 1 & 1 & 1 & 1 \\ 1 & \textbf{na} & 0 & 0 & 0 \\ 1 & 0 & \textbf{na} & 0 & 1 \\ 1 & 0 & 0 & \textbf{na} & 0 \\ 1 & 0 & 1 & 0 & \textbf{na} \end{pmatrix}$$

- $\boldsymbol{Y}$ is an $n \times n$ matrix with elements

$$Y_{ij} = \begin{cases} 1 & \text{if there is an relation } \textit{between} \text{ node i and j} \\ 0 & \text{otherwise} \end{cases}$$

- Diagonal elements represent *self-relations* – typically not considered
- Off-diagonal elements are equal to 1 if there is a relation **between** nodes i and j
- Due to symmetry, $Y_{ij} = Y_{ji}$, that is $\boldsymbol{Y}$ is a *symmetric* matrix

Directed binary relations

The adjacency matrix $\boldsymbol{Y}$ representing such relations is an $n \times n$ matrix with elements

$$Y_{ij} = \begin{cases} 1 & \text{if there is an edge } \textit{from} \text{ node i } \textit{to} \text{ j} \\ 0 & \text{otherwise} \end{cases}$$

- As before, diagonal elements in $\boldsymbol{Y}$ represent self-relations – not considered (na by default)
- Off-diagonal elements are equal to 1 if there is a relation **from** node i **to** j
- Due to *asymmetry*, $Y_{ij} \neq Y_{ji}$, that is $\boldsymbol{Y}$ is an *asymmetric* matrix

Valued relations

Sometimes we need to deal with valued relations between pairs of nodes

- *Count variables* $\rightarrow$ number of email exchanges between employers, number of joint works between two authors, etc...

$$Y_{ij} = 0, 1, 2, \ldots$$

- *Ordinal variables* $\rightarrow$ friendship intensity, military relationship between countries, etc...

$$Y_{ij} = 0, 1, \ldots, k$$

- *Real valued variables* $\rightarrow$ import/export between countries

$$Y_{ij} \in \mathcal{R}$$

How to formalize such relations?

- **Matrix representation** $\rightarrow$ *adjacency matrix*

$$\boldsymbol{Y} = \begin{pmatrix} na & 3 & 8 & 0 & 1 \\ 1 & na & 0 & 2 & 0 \\ 1 & 0 & na & 0 & 1 \\ 7 & 0 & 0 & na & 0 \\ 10 & 6 & 1 & 0 & na \end{pmatrix}$$

- **Graph representation** $\rightarrow$ *weighted edge list*

$$E = \{e_1, e_2, \ldots, e_m\},$$

with $e = (i, j, Y_{ij})$ and Y_{ij} being the value associated to the pair (i, j)

Matrix representation

- 👍 It allows to deal with missing information
- 👎 Memory required to store the matrix grows *quadratically* with n
  - This disadvantage is particularly evident in the case of *sparse* networks
  - In this case, few relations between nodes are observed, so that the matrix contains many zeros
  - Besides the presence of few information, we still need a lot of memory to store the matrix

Graph representation

- It does not allow to deal with missing information
- Memory required to store the matrix grows *linearly* with n
  - This advantage is particularly evident in the case of *sparse* networks
  - Denoting by m the number of observed relations between the n nodes, we only need to store an object of size $m \times 2$
  - The size of the object becomes $m \times 3$ in the case of weighted edge list

Affiliation networks

They entail relations between nodes belonging to *non-overlapping sets*

		Set 2		
		a	···	r
Set 1	1			
	⋮			
	n			

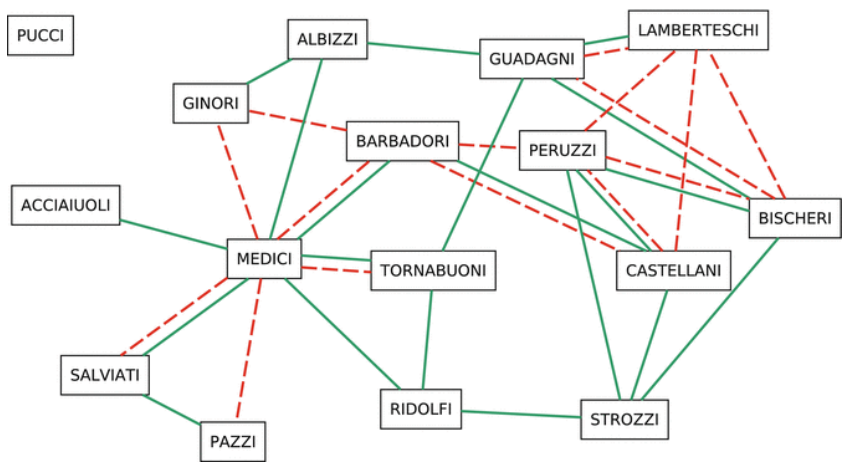
Some examples

- Relations between actors and movies – Set 1: actors – Set 2: movies
- Relations between buyers and products – Set 1: buyers – Set 2: products

Affiliation networks are also known in the literature as *bipartite networks* or *two-mode networks*

Multiplex networks

- They entail different types of relationships between nodes in a single graph
- The number of nodes and relations between them might vary
- As an example, consider the network of marriage and business ties observed between Renaissance Florentine families



Describing the network

Describing the network

Let us consider the matrix representation of a given binary network

$$\boldsymbol{Y} = \begin{pmatrix} na & 1 & 0 & 1 \\ 0 & na & 0 & 1 \\ 1 & 1 & na & 1 \\ 0 & 0 & 1 & na \end{pmatrix}$$

- Even with few nodes, $\boldsymbol{Y}$ can be a complicated object
- **AIM**: exploit proper *network statistics* to describe general features of the network
- In network analysis language, a **statistic** may be defined as any function of the adjacency matrix $\boldsymbol{Y}$

$$t_1(\boldsymbol{Y}), \ldots, t_r(\boldsymbol{Y})$$

Network statistics (II)

Network statistics are frequently used to obtain information about the *network cohesion*

- Questions of interest we aim to answer can be
  - How connected are the nodes a network?
  - Do *friends* of a given node in a network tend to be *friends* as well?
  - Are there nodes that tend to be more connected than others?
  - Do pages in the web tend to be linked with respect to a similar type of contents?

Question: how connected is the network?

To answer, let us consider the **density** *of the network* ρ

It is obtained as the number of observed relations divided by the total number of possible relations

$$\rho = \frac{\#\text{ of observed ties}}{\#\text{ possible ties in the network}}$$

The denominator

- In the case of directed networks, the denominator is $n \times (n-1)$
- In the case of undirected networks, the denominator is $n \times (n-1)/2$

The numerator

We can start from the adjacency matrix $\boldsymbol{Y}$ and compute it as

$$\sum_i \sum_{j \neq i} y_{ij} \quad \text{or} \quad \sum_i \sum_{j < i} y_{ij}$$

$$\rho = \frac{\#\text{ of observed ties}}{\#\text{ possible ties in the network}}$$

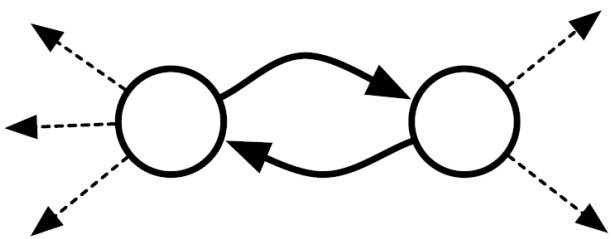
- The density ρ strictly lies in the range $0 \leq \rho \leq 1$
- A network for which ρ *tends to a constant* when the number of nodes increases is said to be **dense**
- A network for which ρ *tends to zero* when the number of nodes increases is said to be **sparse**
- Based on the above definition it is clear that

$$\rho \simeq \Pr(Y_{ij} = 1)$$

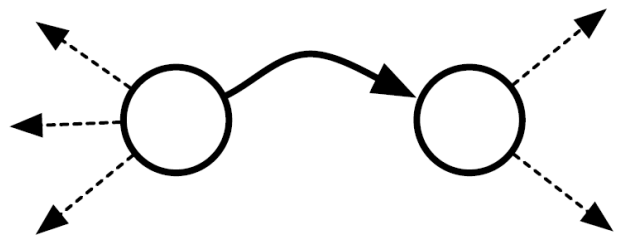
that is, ρ is an estimate of the probability of observing a tie between randomly sampled nodes

Question: how strong is the tendency to return a tie in a network?

When dealing with *directed networks*, not all ties are reciprocated (bidirectional)



reciprocated



unreciprocated

- In social networks, reciprocated ties provide information about *social status*
  - is a friendship relation perceived as being equal?
  - Is a party considered stronger than the other?
- In transportation networks, reciprocated ties provide information about *reachability of a node*
- In citation networks, reciprocated ties provide information about the *influence of nodes*

Reciprocity (III)

Reciprocity R is defined as the fraction of reciprocated ties

$$R = \frac{\#\text{ of reciprocated ties}}{\#\text{ of observed ties}}$$

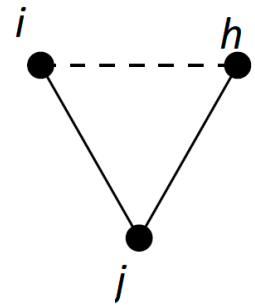
- The reciprocity coefficient lies in the range $0 \leq R \leq 1$
- $R = 0$ implies that all relations observed between nodes in the network **are not** reciprocated
- $R = 1$ implies that all relations observed between nodes in the network **are** reciprocated
- Clearly, in *undirected networks*, reciprocity is always 1.

Question: are friends of friends also friends?

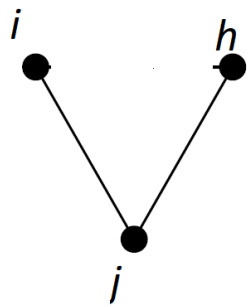
- Let us focus on an **undirected binary network**
- To answer this question, we need to introduce the concept of **transitivity**, which in turn depends on *triads*, that is subgraphs induced by 3 nodes

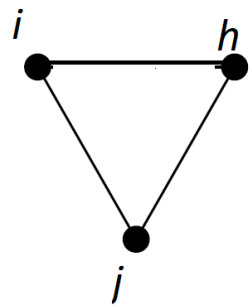
$$\boldsymbol{Y}[(i,j,h),(i,j,h)] = \begin{pmatrix} na & y_{ij} & y_{ih} \\ y_{ij} & na & y_{jh} \\ y_{ih} & y_{jh} & na \end{pmatrix}$$

Consider the node j which is connected to both nodes i and h



Which are the possible states for the above triad?





- In the first case, we have a **path of length 2**
- In the second case, we have a **closed path of length 2**

To measure the transitivity of the network, we may rely on the *clustering coefficient* or *transitivity coefficient*

$$C = \frac{\#\text{ of closed paths of length 2}}{\#\text{ paths of length 2}}$$

- The clustering coefficient lies in the range $0 \leq C \leq 1$
- $C = 1$ implies a **perfect transitivity** in the network
- $C = 0$ implies no closed triads
- C provides an estimate of the conditional probability of observing a relation between nodes sharing a common friend

$$C \simeq \Pr(Y_{ih} = 1 \mid Y_{ij} = 1, Y_{jh} = 1)$$

What about directed networks?

- Computation of the clustering coefficient in the case of **directed network**, frequently proceeds by ignoring the directed nature of the relations
- It is however possible to generalize transitivity to account for the direction of the ties

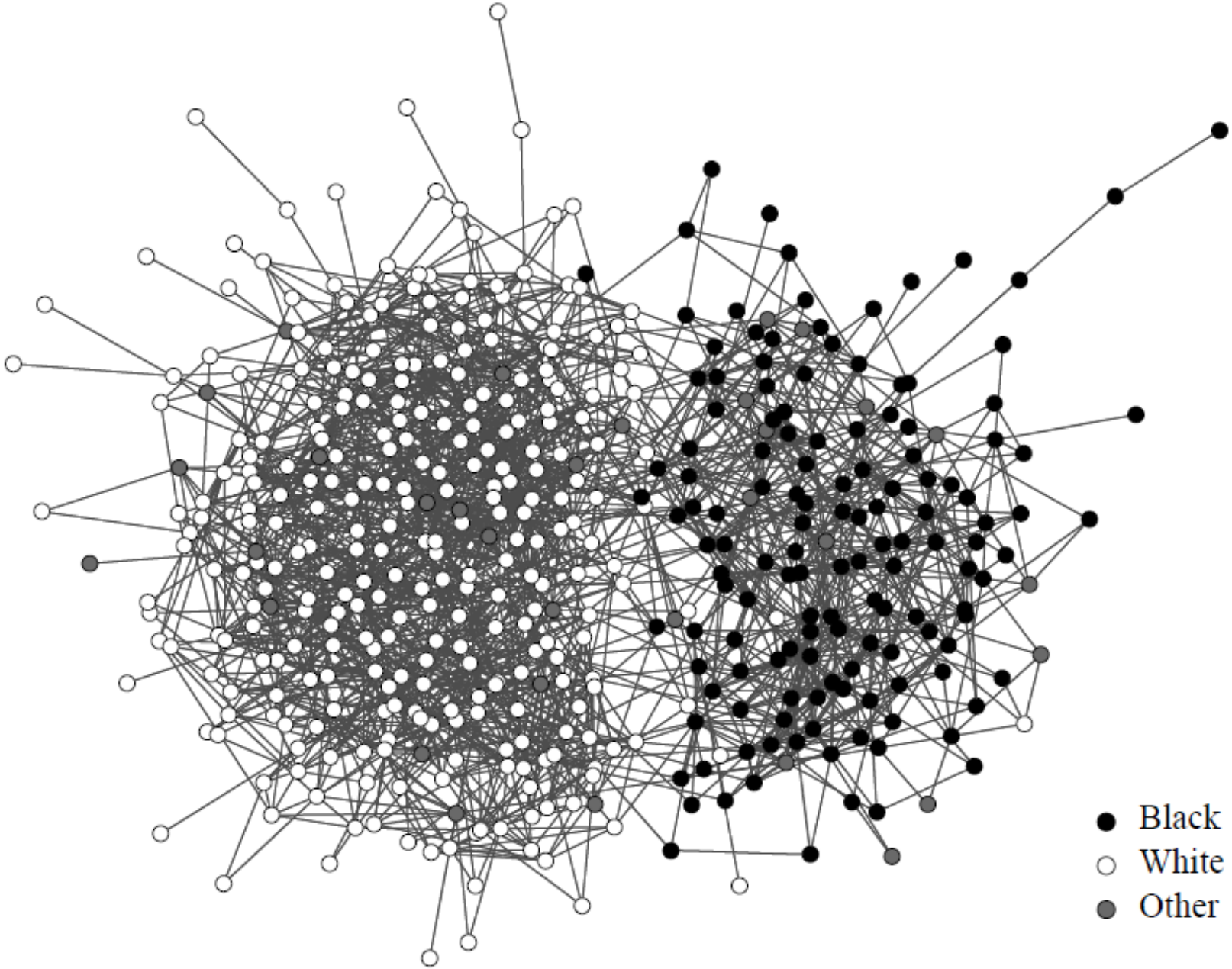
Question: are highly connected nodes similar each other?

Suppose we have a network in which nodes are *classified according to a given characteristic*

- For instance, the nodes could represent people and can be classified according to nationality, race, age, gender, political beliefs, socio-economic status, etc..
- Or they could be web pages classified by what language they are written, or by the importance they have in a given community
- Or they could be employees classified according to the role occupied in the company, the department they belong to, their wage, etc..

All these features represent attributes for the nodes

- In real-world networks, it is quite likely to observe ties between nodes which are similar to each other than between those who are not
- This property is known as *assortative mixing* or *homophily*
- In some networks, the opposite pattern is also seen, which is called *disassortative mixing* or *heterophily*
- In this case, it is more likely to observe ties between nodes having *dissimilar* attributes, e.g. dating is largely disassortative with respect to gender
- There are *two ways* to quantify the assortative mixing of a network which depend on the nature of the attribute we are interested in


Black
White
Other

Qualitative attributes

- Suppose we have an *undirected network* in which the nodes are classified according to a given characteristic that has a *finite set* of possible values
- The network is assortative if a significant fraction of the ties run between nodes of the same type
- A simple way to quantify assortativity would be to measure that fraction
- However, this is not a very good measure because, for instance, it is 1 if all nodes are of the same type
- All friends of a human being, for example, are also human beings, but this is not really an interesting statement
- A better measure would be large in non-trivial cases but small in trivial ones

- We may proceed by looking at
  - the fraction of ties running between nodes of the same type: A
  - the fraction of such ties we would expect to find if they were positioned at random without regard for node type: B
  - the difference between these two quantities: $Q = A - B$
- For the trivial case in which all nodes are of a single type, 100% of the ties run between nodes of the same type
- But this is also the expected figure, since there is nowhere else for the ties to fall
- In this case, $Q = 0$, telling us that there is no a non-trivial assortativity
- Only if A is significantly greater (respectively, lower) than B, we would expect our measure give a positive (respectively, negative) score
- This score is called the *modularity*

- It takes **positive values** when there are **more** ties between nodes of the same type than we would expect by chance: *assortative mixing*
- It takes **negative values** when there are **less** ties between nodes of the same type than we would expect by chance: *disassortative mixing*
- Q is strictly lower than 1, even if it is not bounded from below
- This is in some ways unsatisfactory: how is one to know when the network has strong assortative/disassortative mixing and when it doesn't?

- To rectify the problem, we can normalize Q by dividing by its value for the perfectly mixed network
- Then the normalized value of the modularity is given by

$$\frac{Q}{Q_{max}}$$

- This quantity is called *assortativity coefficient*
  - It ranges between $[-1, 1]$
  - It takes value 1 in the case of perfect mixing
  - It takes value -1 in the case of perfect disassortative mixing

Quantitative attributes

- Suppose we have an *undirected network*, in which nodal variables assume *quantitative values*
- In this case, we may say not only when two nodes have exactly the same value of the attribute, as in the previous case, but also when the attribute is approximately the same
- To measure the assortative mixing of the network, we could simply discretize the attribute, treat it as qualitative, and proceed as before
- A better approach is based on the use of a *covariance measure*

- When the attribute values tend to agree, the covariance will be positive, indicating assortative mixing
- When the opposite is true, the covariance will be negative
- The perfect mixing is obtained when all ties are observed between nodes having exactly the same value of the attribute
- As before, the covariance can be normalized, in order to obtain a coefficient ranging between $[-1, 1]$
- This can be considered as a generalization of the Pearson's correlation coefficient

Some extensions

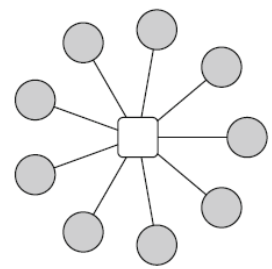
- The assortative coefficient introduced so far has been defined for *undirected networks*
- Directed or weighted versions may be defined as well based on in-going in and out-going ties
- It would also be possible, in principle, to have assortative (or disassortative) mixing according to a set of characteristics
- In this case, we will need to measure the distance between nodes by means of some appropriate metrics
- Formal treatment of vector assortative mixing, however, has not been much pursued in the network literature so far

Describing the nodes

Describing the nodes

- A large volume of research on networks has been devoted to the concept of *centrality*
- This research addresses the question "Which are the most important or central nodes in a network?"
- There are of course many possible definitions of importance and, correspondingly, many centrality measures for networks
- A starting point for defining a centrality measure is the following statement (Freeman, 1979)

A node in a star network is more central than any other node in any other position in any other type of network



- Node centrality is based on three different dimensions
  - its connectedness
  - its role as a mediator
  - its closeness to the others
- The first dimension measures potential communication activity
- The second dimension measures the potential for control
- The last one the potential independence or efficiency of a node (e.g., in passing messages to all other nodes)

Centrality measures

Let us consider an *undirected, unweighted, network*.

Common **centrality measures** are:

- degree centrality $\rightarrow$ it quantifies the absolute number of contacts a node has in a network, i.e., a quantification of connectedness
- closeness centrality $\rightarrow$ it quantifies how close a node is to others
- betweeness centrality: it quantifies the mediating position of a node
- eigenvector centrality: it jointly quantifies the number and the importance of contacts a node has in a network

Degree centrality

IDEA: a node is *central* for a given network if it is connected to many other nodes

- The *degree* or *degree centrality* of node i denotes the number of ties involving i
- In other words, it measures how many direct contacts a node has in the network
- This provides information on the direct influence of the node in the network and its access to first-hand information
- It ranges from 0 to $n-1$, and can thus be normalized dividing it by $n-1$
- The higher (respectively, smaller) the index, the more (respectively, less) central the node in the network

Closeness centrality

IDEA: a node is *central* for a given network if *it is close* to many other nodes

- To quantify how close two nodes are, we may rely on the *geodesic distance*
- The *geodesic distance* d_{ij} between two nodes i and j is the length of the shortest path existing between these two nodes
- To compute the index, we start from the *farness* of the node: the sum of its distances to other nodes

$$farness_i = d_{i1} + d_{i2} + \cdots + d_{in}$$

- The closeness centrality of node i is simply the inverse of its farness

$$closeness_i = \frac{1}{farness_i}$$

- The index is is higher (respectively, lower) for more (respectively, less) central nodes

Betweenness- centrality

IDEA: a node is *central* for a given network if *it is located in between* many other nodes

- The *betweenness centrality* of node i is defined as the fraction of geodesic (shortest) paths passing through it

/

Eigenvector centrality

IDEA: a node is *central* for a given network if it is *connected to other central nodes*

- The centrality of a node i is proportional to the weighted sum of the centralities of its neighbors
- In this respect, a node can be *central* because
  - it has many neighbors
  - it has *important* neighbors
  - both

Degree centrality

We need to distinguish between *in-degree* and *out-degree* centrality

- The *in-degree* is the number of *ingoing ties* of a given node and can be interpreted as the *popularity* of node i
- The *out-degree* is the number of *outgoing ties* of a given node and can be interpreted as the *expansiveness* of node i

Closeness centrality

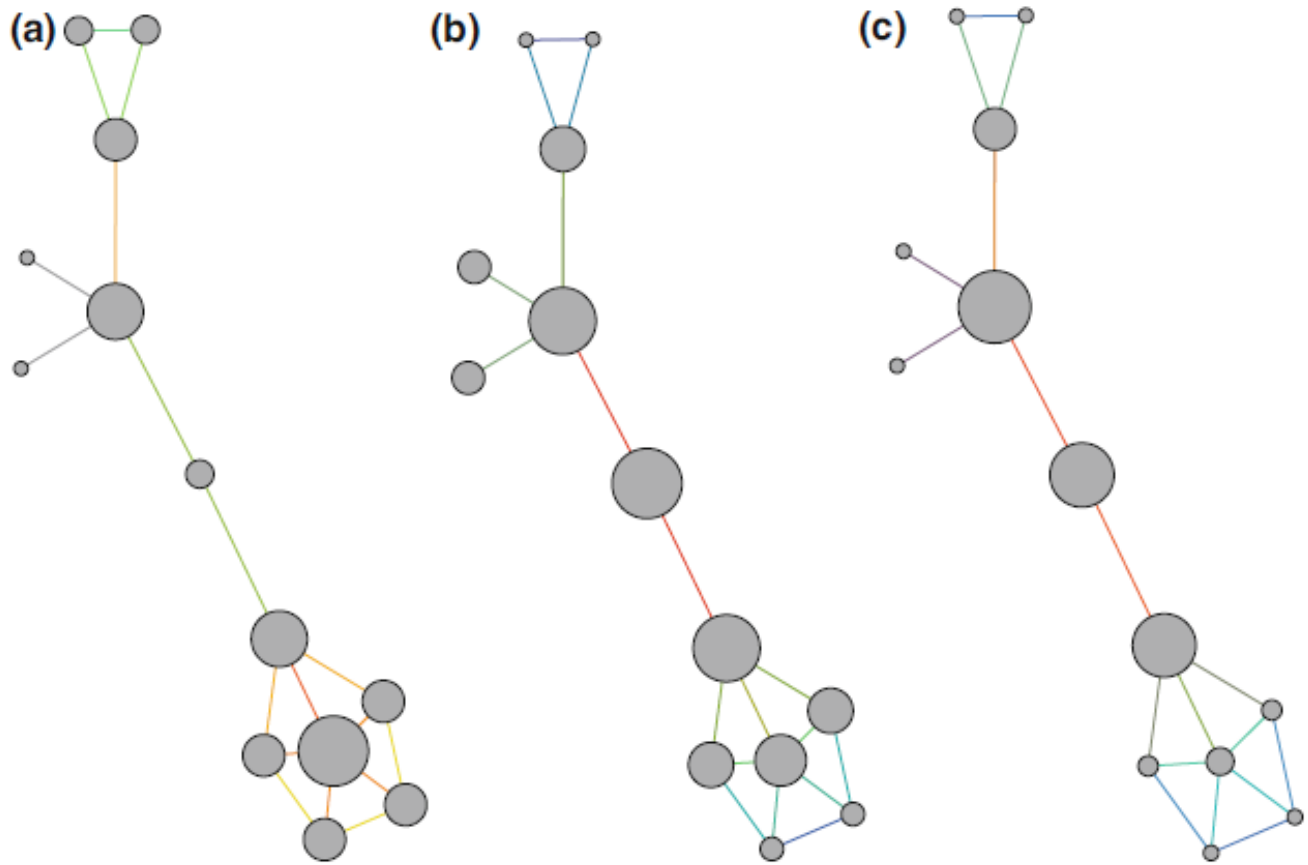
When looking the mean geodesic distance between i and all other nodes in the network, we need to account for the direction of the ties

Betweeness centrality

When looking at the number of geodesic paths connecting two nodes (passing or not through i), we need to account for the direction of the ties

Eigenvector centrality

Typically, the focus is on in-going ties that provide information on the popularity of the node



A graph in which the nodes' size is determined by their centrality according to different centrality indices. **a** Degree centrality, **b** closeness centrality, **c** betweenness centrality. It can be seen that different centrality indices identify different nodes as the *most central* ones

Centralization of a graph

- Sometimes centrality indexes are used to describe the structure of the whole graph, namely its *centralization*
- There are two different interpretations of this term
  - In the first version, it is interpreted as the degree to which all nodes are close to each other
  - In this way, it measures the compactness of the graph
  - In the second version, it is interpreted as the degree to which the most central nodes dominates the structure of the graph (Freeman, 1979)
  - In this way, it is based on the intuition that a star is the most centralized graph of a given size

- Let $C_{max} = \max_i C_i$ be the maximum centrality index observed in the network. A *network centralization index* as proposed by Freeman is defined as

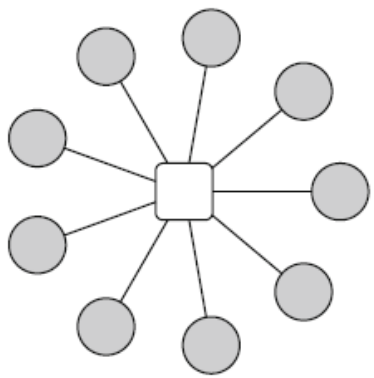
$$CI = \frac{\sum_{i=1}^{n}[C_{max} - C_i]}{\max_Y \sum_{i=1}^{n}[C_{max} - C_i]}$$

  where the maximum at the denominator is referred to *all possible network configurations*

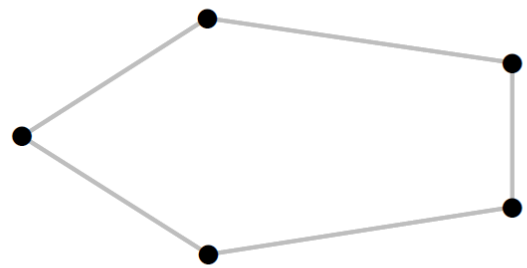
- That is, it is defined by the difference of the centrality values and the maximal centrality value in the graph, normalized by the maximally achievable sum of differences over all graphs of the same size
- The normalization term clearly varies wrt to the chosen centrality index
- $0 \leq CI \leq 1$
- $CI = 0$ when all nodes are equally central
- $CI = 1$ when one node is maximally central and the others are minimally central

Two extreme configurations

star network



circle network



What about directed networks?

Degree centralization

Two degree centralities are differentiated: the in- and the out-degree, so two different centralization indexes can be defined

Closeness and betweeness centralization

We need to account for the direction of the ties

Eigenvector centralization

Typically, the focus is on in-going ties providing information on the popularity of the nodes

Network data modeling

Network data modeling

- Network data are characterized by a number of dependencies which have been found empirically as well as theoretically
  - Reciprocation
  - Homophily
  - Transitivity
  - Degree variability
  - Etc..
- Typically, these make standard statistical models based on the independence assumptions between units not appropriate
- The literature contains however various ways to represent network dependencies in statistical models

Three broad approaches may be distinguished

- *Incorporating network structure through covariates*
  - A statistical model for independent data is considered
  - Network dependencies are represented by including in the model explanatory variables (e.g., Gulati and Gargiulo, 1999)
- *Controlling for network dependences*
  - Network dependencies are considered by specifying a covariance structure but are not explicitly modeled (e.g., Lindgren, 2010)
- *Conditioning on statistics that express network dependencies*
  - A potentially large number of parameters is included in the statistical model
  - These allow to represent the complex dependence existing between the tie variables Y_{ij}

- Under the latter approach, the observed network is considered as *an outcome of a random draw from the postulated model*
- It is natural or plausible to consider that the observed network data could also have been different
- The network could have been observed on a different moment, nodes could have been different, external influences could have been different, etc..
- In such a "population" of different networks, systematic patterns in the data captured by the parameters in model would be the same, while the observed outcome (the whole network) could have been different

- The aim of the statistical model is to represent the main features of the network via a small number of parameter estimates

- Expressing the uncertainty of those estimates gives an indication of how different estimates might be if indeed the researcher had observed a different network from the "population"

Modeling alternatives: an overview

Different statistical models have been developed in the literature to express (model) network dependencies

- Conditionally Uniform Models (Holland and Leinhardt, 1976)
- Latent Space Models (Lazarsfeld and Henry, 1968)
- Exponential Random Graph Models (Frank and Strauss, 1986)

Conditionally Uniform Models

- Network properties that researchers wish to control for are summarized by means of proper network statistics
- It is assumed that, conditional on these statistics, the distribution of the network is uniform
- That is, each network satisfying the constraints leading to the desired statistic has the same probability to occur; all others have 0 probability to occur
- This reflects the idea that statistics to which we are conditioning contain all relevant information for the phenomenon under investigation; all the rest is randomness
- ISSUE: conditionally uniform models become very complicated when more elaborated properties are identified and a richer set of conditioning statistics is considered
- Because of this, they are not currently used much

Latent space models

- They assume the existence of *latent (i.e., unobserved) variables*, such that the observed variables have a simple probability distribution given the latent variables
- The specification of the latent variables' distribution identifies the *structural model*
- The distribution of the observed variables, conditional on the latent ones, identifies the *measurement model*
- We may distinguish latent space models according to the type of structural model they rely on
  - *Discrete Latent Space Models* (Holland et al.,1983; Snijders and Nowicki, 1994; Nowicki and Snijders, 2001, Daudin et al., 2008; Airoldi et al., 2008)
  - *Distance Latent Space models* (Freeman, 1992; Hoff et al., 2002)

Stochastic blockmodels (e.g., Daudin et al., 2008)

- A node-level latent variable A_i, defined on a finite support, is introduced and provides a clustering of the nodes
- Conditional on A_i and A_j, tie variables Y_{ij} are independent
- The corresponding distribution only depends on A_i and A_j
- That is, latent variables capture all the dependence between observed tie variables

Mixed membership model (e.g., Airoldi et al., 2008)

- Modeling assumptions are similar as those of the SBM, even though each node can be a member of several classes
- This allows us to describe situations where the nodes may play multiple roles

Distance Latent Space models

A function $d : \mathcal{N} \to [0, \infty)$ denotes a *distance function* if

- $d(i, i) = 0 \forall i = 1, \ldots, n$
- $d(i, j) = d(j, i), \forall i, j = 1, \ldots, n$
- $d(i, j) \leq d(i, k) + d(k, j), \forall i, j, k = 1, \ldots, n$ (triangle inequality)

Distance Latent Space models assume that

- the nodes are points in some latent space $A = \{1, \ldots, n\}$ with distance function $d(i, j)$
- the probability of observing a tie between nodes is a decreasing function of the distance between such points
- That is, the closer the points, the larger the probability that they are tied

Exponential Random Graphs

- Rather than conditioning on latent variables for capturing dependence between tie variables, such a dependence is explicitly modeled
- The probability of observing the network is modeled as a function of given network statistics (sufficient statistics)
- Such statistics reflect are defined in order to capture dependencies between tie variables
- Conditional on them, tie variables are assumed to be independent

Final remarks

Final remarks

- During the last 20 years, tremendous developments have taken place in network modeling
- The number and size of network data have gone up since the beginning of network data analysis, from some tens of nodes at that time, to millions and billions today
- The first explosion of the data volume date back to the massive digitization of the data in the nineties
- An even stronger shock arose at the start of the twenty-first century
- The democratization of the Internet led to the emergence of many and varied large networks: physical infrastructures, the WWW, and a plethora of online social and sharing networks above all

- These recent revolutions created two veins in the network research field in the late nineties
- Analyzing these new large graph datasets within a reasonable amount of time became another challenging issue
- Interaction data grows very fast as function of the number of individuals and its analysis brings combinatorial issues
- Thus, new inference methods need to be efficient from a computational point of view