Università degli Studi di Napoli
Federico II

Network data in the Partial Least Squares framework

Stefania Spina

Doctoral Thesis
in Statistics

XXVI Cycle

Network data in the Partial Least Squares framework

Contents

List of Tables

List of Figures

Introduction

Network data can describe social contexts and a social context can influence the emerging of relationships [47]. The measurement of attitudes, behaviors, human features and structural characteristic of a context are retained in classical statistical variables (attributes data). A number of statistical models have been developed for network structure analyses. The most part of them arise from geographical connection models developed by Cliff and Ord [26] and assume that for a given actor in a network the value of an outcome is directly influenced by the values of the actors' neighbors outcomes.

The Network Effects Model (NEM) by Doreian [42] introduces this idea in the framework of auto-regressive models. This class of models allows accounting for dependence among actors in classical regression models, i.e. by assuming that all the variables are directly observed. This call for an extensive approach that will exploit the capability of the NEM to model the dependence structure among the units and the possibility to manage observed variables as well as latent constructs. Recently, Doreian *et al.* [48] have presented a method extending the Structural Equation Modeling (SEM) [7] to network data.

In Structural Equation Modeling (SEM) framework, real complex phenom-

ena can be studied taking into account causal relationships among a number of latent concepts (i.e. the Latent Variables - LV) each measured by several observed indicators defined as Manifest Variables (MV).
The inclusion of the relational data structure (adjacency matrix) in a SEM will offer the opportunity to model classical issues of Social Sciences, Economics and Marketing such as Social Influence and Homophily, for example. The advantage can be envisaged in the different specification of the path diagram.
This model specification offers higher flexibility, it allows to consider i) separate or joint effect of intrinsic opinions of the social actors, ii) the extent to which they are influenced by their alters, and iii) how people with similar characteristics are more likely to form ties.
Two complimentary methods emerged in the field of SEM: the so-called covariance-based SEM (also referred as LISREL models) and the more recent component-based SEM.
The goal of covariance-based approach is to reproduce the sample covariance matrix of the manifest variables by means of the model parameters.
By contrast, the aim of the component-based SEM is to obtain score values of latent variables as linear combinations of their associated manifest variables.
The component-based approach is also referred to as a prediction-oriented SEM compared to the classical SEM that is a confirmatory approach.
Among the component-based approaches the PLS Path Modeling (Wold [161] ; Tenenhaus *et al.* [149]) is the most widely used.
The PLS Path Modeling (PLS-PM) has the advantage that involves no assumptions regarding the population and the scale of measurement [60], so it works without distributional assumptions.

This kind of modeling is known as soft-modeling [161] in contrast with hard-modeling (i.e. maximum-likelihood estimation procedures) typical of covariance based SEM.

For this reason that a component-based approach to network data through Partial Least Squares path model algorithms is proposed in this thesis.

Thesis outline

This work is divided into four chapters.

In **Chapter 1**, the basic elements about networks and their representation, measurement and characterization, useful for the statistical modeling in the context of social networks research, are described. Then a brief review about the development of several statistical models for network data is given.

Chapter 2 is dedicated to presenting an overview of the historical development path modeling, the conceptual background and foundations of Partial Least Squares Path Modeling. Then a description of PLS-PM algorithm step by step is proposed.

In **Chapter 3**, we present a statistical soft-modeling framework to network data. The PLS-PM is extended to the analysis of network data by introducing the adjacency matrix in the model. A simulation study comparing this proposal to the classical network effects model is presented.

In **Chapter 4**, a feasible substantive interpretation in the scope of Social Science of this new approach is discussed. The sociological foundations of the social relations that provide a basis for the alteration of an attitude or behaviour by one network actor in response to another, labelled social influence or contagion in literature, are described. Then mathematical models of influence processes involving networks and related statistical models used in data analysis are reviewed.

Chapter 1

The analysis of social networks

Social network analysis has been used since the mid-1930s to advance research in the social and behavioral sciences, progressing slowly and linearly, until the end of century.
In 1990s, interest in social network analysis and use of the wide-ranging collection of social network methodology began to grow at a much more rapid rate.
The problem of how single individuals can combine in order to create enduring and functional societies, i.e. the problem of social order of Plato, has been solved through social network theory.
One of the most important notions in the social sciences is that individuals are embedded in thick webs of social relations and interactions.
Social network analysis offers the methodology to study structures of relationships linking individuals or other types of social units, such as organizations and countries. It is assumed that social ties are important because

they transmit behavior, attitudes, information, or goods.
The focus of the social sciences is to understand the social structure conceptualized as a network of social ties, considering two elements: i) social structure, i.e. a system of social relations tying distinct social entities to one another; ii) interest in understanding how social structure form and evolve.
The starting point has been the Durkheimian vision that dependence must be seen as central element to the idea of sociality and it has to be used to reconstruct the idea of social space.
In this space the units are not individuals but the ties that connect them. The variety of ties that enter into the construction of these social spaces can been modelized through dependence models.
Two distinct types of network models are common respectively: individual and relational-level models.
In the **individual level models**, the analysis focuses on **outcome of a single actor** and the network data are used to define explanatory variables.
By contrast, the **relational level models** use the **relationships among individuals in a network**, treating it as a multivariate dependent variable with individual linkages (or ties) as its elements [127].
Thus, the first type of models makes inference about attributes of the individuals, while the second ones about the ties linking the individuals.
In both models, a big problem is accounting for a complex correlation structure among outcomes due to the network.
If there are n individuals in a data set, this is of order $n \times n$ in an individual-level analysis and of order $n^2 \times n^2$ in a relational-level analysis [127].
The next section reviews some foundations of social network analysis: i)

the numerical and visual representation of network data sets; ii) the introduction of some basic network statistics; iii) the detection of subgroups within networks.
These measures are very important for network typology, because they can be used to develop models.
The last part of this chapter introduces a brief history of the statistical models for network data.

1.1 Definition of social networks

A social network consists of one or more sets of units - also known as nodes, actors, or vertices - together with the relationships or social ties among them [127].
The key concepts of network analysis are: nodes, relational tie, dyad, triad, sub-group and group.
The **units or nodes** that can be objects of study are people, groups, or organizations but also texts, artifacts, or concepts.
These elements, which form the network, are distinct from one another, can be uniquely identified, and are finite in number.
The most common representation of relationships employed in network ties can be: i) evaluation of one person with another (i.e. friendship, liking or respect); ii) transfers of material resources (i.e. business transaction); iii) behavioral interaction (i.e. talking together or sending messages); iv) biological relationship (kinship or descent [155]).
Most social network studies also include attribute data describing the actors, the relationships, or both.

The measurement of actor attributes (i.e. gender, race, socioeconomic status, revenues, purpose of business, etc...) is verified through composition variables.
Some sub-networks are interesting in this field as: dyad, triad, star or egocentric networks.
A **dyad** consists of a pair of actors and the tie between them, while a **triad** is a subset of three actors and the tie or ties among them.
A **star** consists of an actor and all relationships incident to it.
An **egocentric network** consists of an actor, i.e. ego, and the other actors in its immediate neighborhood, i.e. alters, and the relationships among them.
Another important type of variables that can be included in a network data set are structural variables, that are measured on pairs of actors.
Most social applications are centered on relationships that link elements within a single set of actors.
In this case we have a network that is known as **one-mode network**. There are also networks that may involve two sets of actors or one set of actors and one set of events. These types of networks are defined **two-mode networks** or **affiliation networks**, that will be described in another section.

1.1.1 Network study designs

There are many ways in which social network data can be gathered [155]. For example there are: i) questionnaires; ii) personal interviews; iii) direct observations [65]; iv) archival records [15] (e.g. based on administrative records or computer-mediated communication systems [114]); v) experiments (e.g. [151]; [72]); vi) other tecniques (e.g. small world [151] and diaries [35]).

When archival measures do not exist or do not include information about the relationships of interest, or when other methods of collecting network data are not feasible [115], surveys are required. There are two types of network surveys depending on type of network: 'whole' networks or 'egocentric' networks.

Whole-network studies seek to assemble data on ties linking all actors within some bounded social collective, where boundaries or rules of inclusion for actors are specified [104].

In this case, surveys can collect either **one-mode** network data, based on relationships among elements of a single set of actors, or **two-mode** network data, based on relationships among actors in two distinct sets.

By contrast, **egocentric network** studies have the objective of describing local social environments, limiting the measurement of the relationships in the vicinity of one or more focal units or actors [115].

In many studies, it is possible to integrate network data with other information, about attributes of actors or dyadic ties, or about group-level attributes in studies with two or more groups.

Among the most common instruments used for whole - network data, are:

- **The sociometric test** [123], where each actor identifies, within a network, the other people, i.e. alters, with whom he has a relationships;
- **Cognitive social structure task** [100], where respondents are used as informants about social ties between alters and their own relationships;
- **Socio cognitive mapping** [19], where respondents are asked to report sets of people who 'hang around together a lot' via free recall;

- **Pile sorts** [66], where each respondent is asked to sort a deck of cards, containing the names of the actors in a network, into mutually exclusive piles considering subsets of actors who are close to one another or who interact frequently.

Among instruments for egocentric networks, the most important are name generator instruments. In surveys, by using these instruments, e.g. telephone interviews [99] mail questionnaires [112] and web-based instruments [153], respondents are treated as informants to whom is asked information about:

- Attributes of particular alters in a network (e.g. race, age or ethnicity);
- Properties of alter-ego ties (e.g. emotional closeness or frequency of contact);
- Relationships among the alters themselves, in order to measure many different aspects of social network structure [113].

In these cases the alters are not surveyed or interviewed.

1.2 Representations of networks

The mathematical origins of network analysis permit to manipulate, calculate and visualize social networks.
Network data can be represented in a number of ways, depending on the type of the application.
In order to describe social network data mathematically, specific notations deriving from graph theory, as graphs, and algebraic notations, as for example adjacency matrices, are used [64] .

1.2.1 Representing networks with graphs

Networks are often represented by using graphs.
A **graph** is a relational structure consisting of two elements: a set of entities, called vertices or nodes, and a set of entity pairs indicating ties, called edges.
Formally, we represent such an object as $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the vertex set and $\mathcal{E}$ is the line set [17].
A **line** can be directed or undirected.
A **directed** line is called an **arc** , whereas an **undirected** line is an **edge**.
Specific types of graphs may be identified via the constraints satisfied by $\mathcal{E}$.

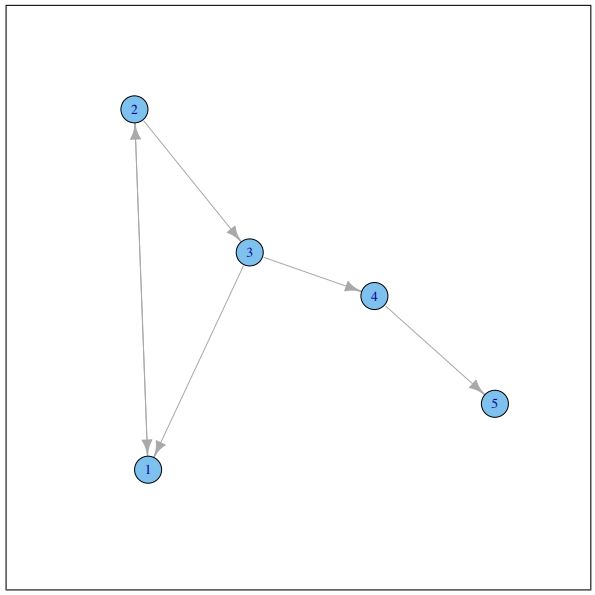

Figure 1.1: An example of directed graph

Graphs that represent dyadic relations, where there is no distinction between the 'sender' and the 'receiver' of the relation, are said **undirected** (or non-directed), and have edge sets which consist of unordered pairs of vertices.

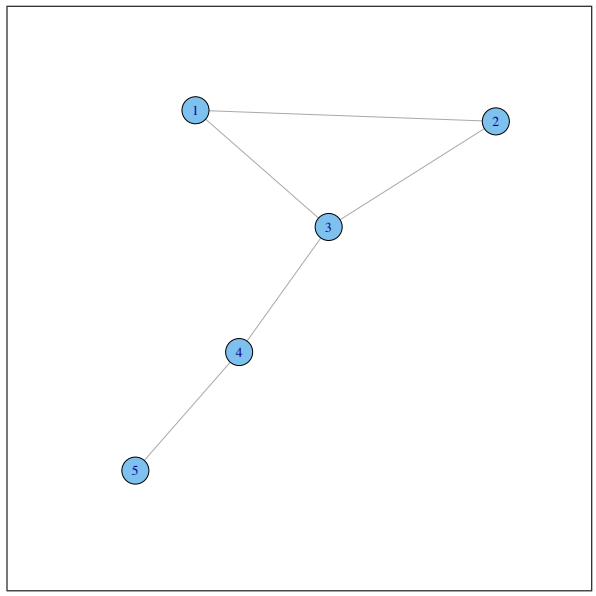

Figure 1.2: An example of indirected graph

For these relations, this principle can be defined formally as $\{n, n'\} \in \mathcal{E}$ iff vertex n is tied (or adjacent) to vertex n' where n , $n' \in \mathcal{N}$.
On the other hand, **graphs**, which represent relationships where 'sender' and 'receiver' roles are distinct and they have edges that are ordered multi-sets, are said **directed graphs (or digraphs)**. Formally, the requirement is that $(n, n') \in \mathcal{E}$ iff n sends a tie to n'.
It is possible to use arrow notation to denote ties, such that $n \to n'$ should

be read as n sends a tie to n', which is close to visualization of social networks.

A graph, independently if ordered or not, is said simple if it has no **loops**, i.e. when it has an edge going from a vertex to itself, if there is no edge having multiplicity greater than one.

Finally, when in a **graph** there is a value associated with each line or each arc, it can be called a **valued graph**.

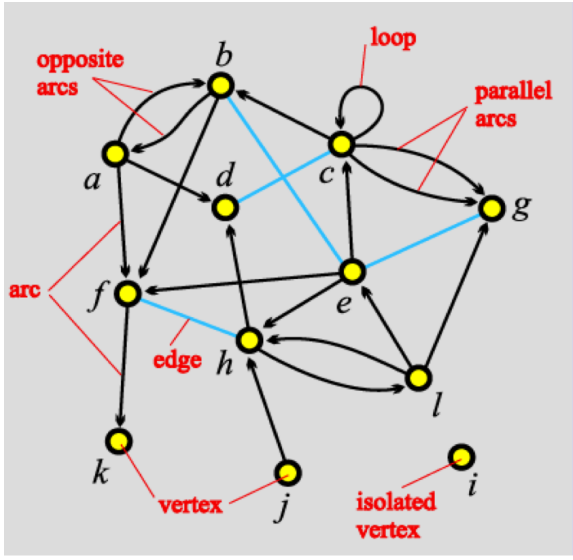

Figure 1.3: Several elements of graph. Source: Batagelj, 2 - 4 May 2012, Naples.

1.2.2 Path, cycle, walk and geodesic distance

There are also other properties describing graphs.
Actors in networks are connected to one another indirectly via intermediaries as well as directly.
An example is if a vertex can reach another by traversing a series of edges within the network.
It is called **path** a sequence of distinct but adjacent vertices n,..., n' together with their included edges.
This implies that the two vertices n and n' are connected, in an **undirected graph**, when there exists some n , n' path in $\mathcal{G}$.
In **directed graphs**, by contrast, there are some conditions, such as the existence of either a directed path from n to n' or a path from n' to n.
This requires a sequence of vertices that can be traversed, in order to get from one end of the path to the other, but this condition is not required to hold in both directions [17].
Another most stringent condition is that these paths cross the same intermediate vertices.
Vertices pairs satisfying this reciprocal condition are said to be recursively connected.
Among path-related concepts, there is an important network-based measure of the social distance separating actors: **geodesic distance**.
A **geodesic path** is a minimal length path between a given pair of actors and geodesic distance is its length.
A particular type of path, when the start and end-points are the same, is called **cycle**.
Both the path and the cycle are special cases of the **walk**, which is simply

a sequence of serially adjacent vertices together with their included edges. Unlike a path, a walk can be of any length, while the length of a path is $n - 1$.

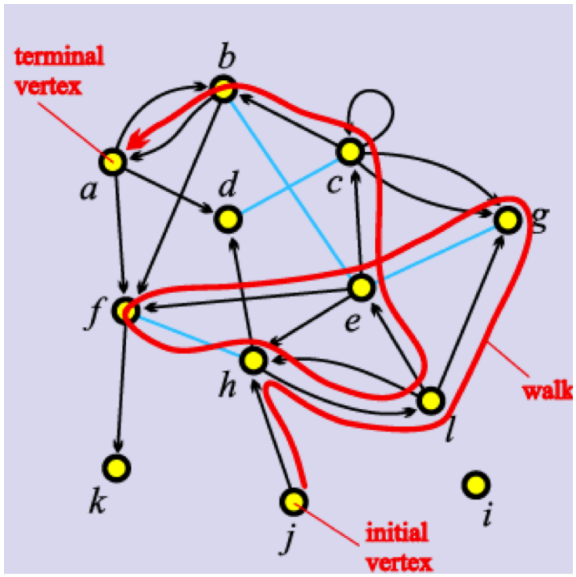

Figure 1.4: A walk in a graph. Source: Batagelj, 2 - 4 May 2012, Naples.

1.2.3 Representing networks through matrices

It is possible to record who is connected to whom on a given social relation via matrix too.
The information in a graph $\mathcal{G}$ may also be expressed in a matrix form. There are two such matrices that are especially useful:

- the affiliation matrix;
- the adjacency matrix.

The **affiliation matrix** is a rectangular matrix $I = n \times E$, that has **nodes** in rows and **events** in columns, such that $i_{ij} = 1$ if i is an end-point of edge j and 0 otherwise.

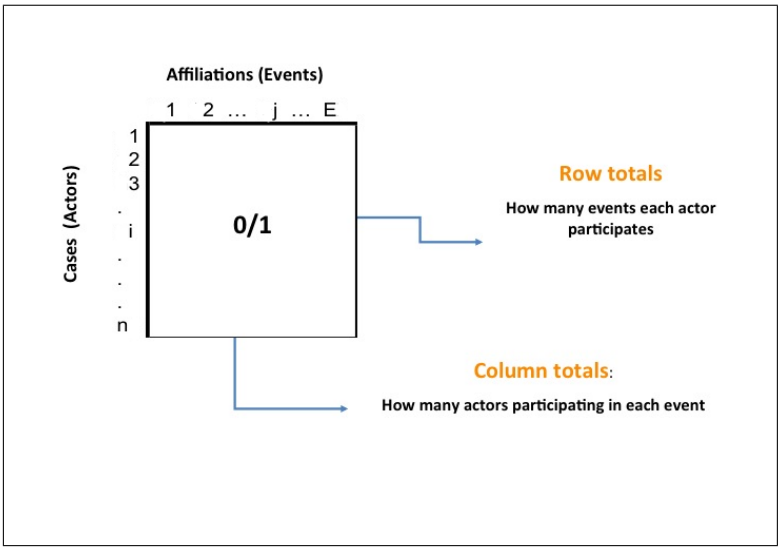

Figure 1.5: Affiliation matrix

Affiliation matrices are not often used in network research but they can be used in order to represent **hypergraphs**, i.e. graphs for affiliation network data, and **two-mode data**, i.e. networks about relations between two disjoint types of entities.
It is possible to obtain from an affiliation matrix an **adjacency matrix**. To get the one-mode representation of ties between rows (i.e. actors), it is necessary to multiply the affiliation matrix by its transpose.
By contrast, to get the one-mode matrix formed by the column entities (i.e. the number of people partecipating in each event), it is necessary to

pre-multiply the affiliation matrix by its transpose.

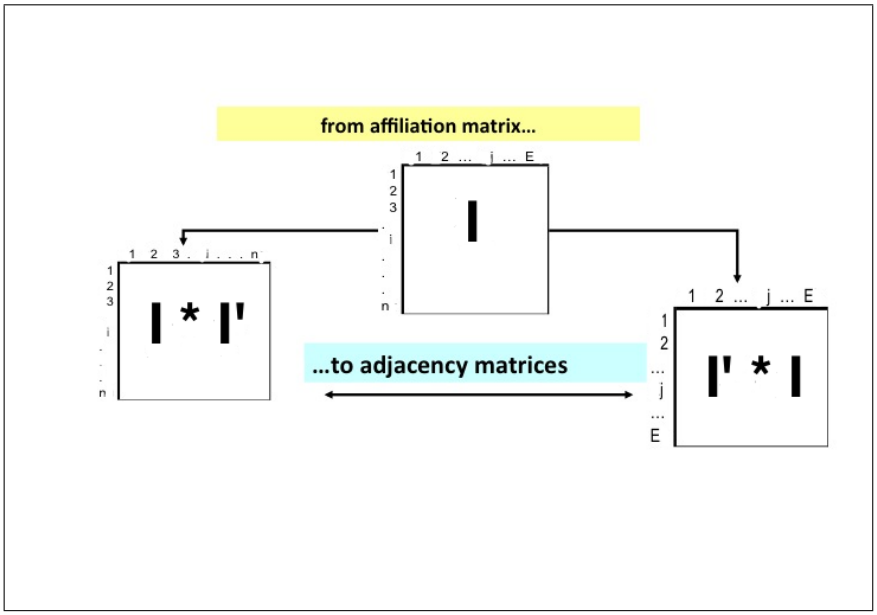

Figure 1.6: Matrices for social networks

If we dichotomize the matrix, obtained from one of the trasformations described above, with all elements of the diagonal are 0, we can obtain an another type of adjacency matrix .

The most part of social studies rely on the actor by actor adjacency matrix **A** a binary-valued $n \times n$ matrix, where its rows and columns refer to the same set of entities : a single mode.

For node set $\mathcal{N}$:

$$a_{ij} = \begin{cases} 1 & \text{if there is a } \textbf{tie} \text{ from node } i \text{ to node } j, \\ 0 & \text{if there is } \textbf{no tie} \text{ from } i \text{ to } j. \end{cases} \tag{1.1}$$

The adjacency matrix **A** is **symmetric**, where $a_{ij} = a_{ji}$, and it represents an **undirected graph** $\mathcal{G}$.
This is not true if $\mathcal{G}$ is a **digraph**.
When all elements of the diagonal of **A** are 0, $\mathcal{G}$ is a simple graph.
Otherwise, $a_{ii} = 1$ iff vertex i has a loop, both for directed and undirected graphs.
In presence of valued edges, the same above representation is used with the only difference that a_{ij} is the value of the (i, j) edge. It will be 0 when no edge is present.

1.3 Descriptive properties of networks

This paragraph reviews the most important network descriptive statistics and measures.
Upon obtaining network data, it is necessary to extract interpretable and useful information from a large and complex social structure.
Visualization of network data can be useful, but it is not sufficient for scientific work.
It is necessary to examine particular structural properties, quantifying them and comparing them against some baseline models or null hypothesis.
The structural index approach is a paradigm, whose basis is the development of descriptive indices, which quantify the presence or absence of particular structural features.
Properties of social networks can be defined at different levels of aggregation:

- local measures for individual nodes or small subsets;
- global measures requiring simultaneous information about the entire

graph.

We use **centrality** to refer to positions of individual vertices within the network, whereas we use **centralization** to characterize an entire network. A network is highly centralized if there is a clear boundary between the center and the periphery.
In this network, information spreads easily but the center is indispensable for the transmission of information.

1.3.1 Local and global measures

The very simplest property of a network is its number of actors **N**, known as its **order**.
For binary-valued networks, the corresponding relation-level statistic is the number of ties, known as **size**:

$$L = \sum_{i,j} a_{ij} \tag{1.2}$$

Properties of nodes and graphs can be defined using the concepts of adjacency and affiliation for the nodes and lines in a graph.
The three most widely used **centrality measures** are grounded in graph theory [63] and are: i) degree; ii) closeness; iii) betweenness.
Another important index, not belonging to the three Freeman's classic measures, is the **eigenvector centrality** [10].

While local measures describe local structure of a particular vertex, **global measures** quantify structural properties of the network as a whole. Such measures are useful when comparing networks and determining the large-scale structural context in which behaviour occurs.

The **centralization** of any network is a measure of how central its most central node is in relation to how central all the other nodes are.
The index of centralization has the property that the larger it is, the more likely it is that a single actor is quite central, with the remaining actors considerably less central [155]. The less central actors reside in the periphery of a centralized system.
In 1979 Freeman [63] has proposed the general mathematical definition of centralization for non-weighted networks.
Recall that $C_A(n_i)$ is the centrality index of actor i.
Define $C_A(n^*)$ as the largest value of the particular index that occurs across the n actors in the network; that is, $C_A(n^*) = \max_i C_A(n_i)$.

$$\sum_{i=1}^{n}[C_A(n^*) - C_A(n_i)] \tag{1.3}$$

is the sum of the differences between the most central node in a network and all other nodes, while

$$\max \sum_{i=1}^{n}[C_A(n^*) - C_A(n_i)] \tag{1.4}$$

is the theoretical maximum possible sum of differences in actor centrality, where the differences are taken pairwise between actors.

$$C_A = \frac{\sum_{i=1}^{n}[C_A(n^*) - C_A(n_i)]}{[\max \sum_{i=1}^{n}[C_A(n^*) - C_A(n_i)]} \tag{1.5}$$

This index is dimensionless, and varies between 0, in the case of a graph in which all vertices have the same centrality scores, that being $C_A(n^*)$ and 1, in the case of a graph of maximum concentration.

There is an index of network centralization for each measure of centrality but some centralization measures need special networks: degree centralization is applicable only to networks without multiple lines and loops, and closeness centralization requires a (strongly) connected network.

1.3.2 Density of a network

In the case of considering a network as a whole, the simplest structural characteristic is density that involves the number (and the proportion) of the edges in the whole graph.
The **density** is defined as size relative to the number of possible ties and equal to $L/(n(n-1))$ for **directed networks**.
Let L be the number of edges present in a graph which can take on any integer value from 0 to $n(n-1)$. In the case of **undirected networks**, we define the density of a graph Δ as the proportion of the number of edges present, L, to the maximum possible number of edges in a graph.
It can be calculated as:

$$\Delta = 2L/n(n-1) \tag{1.6}$$

The density of a network keeps track of the relative fraction of links that are present in a network, and because the average degree equals $2L/n$ the density is simply the average degree divided by $(n-1)$.

This **average degree**, divided by $n-1$, is exactly the density of the graph:

$$\sum C_D(n_i)/n(n-1) = \sum C'_D(n_i)/n = \Delta \tag{1.7}$$

Thus, mathematically, the density is also the average standardized degree.

1.3.3 Sub-structures of networks

One of the most common interests of structural analysts is in the **sub-structures** that may be present in a network.
Divisions of actors into groups and sub-structures can be a very important aspect of social structure.
As above, the unit of analysis in network analysis is an entity consisting of a collection of individuals and the linkages among them : i) individual actor level of analysis; ii) dyads level of analysis (two actors and their ties); iii) triad level of analysis (three actors and their ties); iv) subgroup level of analysis (e.g. clique ([111]; [83]; [155]) or components); v) global level of analysis.
When two actors have a tie, they form a group.
One approach of thinking about the group structure of a network begins with this most basic group, and seeks to see how far this kind of close relationship can be extended.
This is a useful way of thinking, because sometimes more complex social structures evolve, or emerge, from very simple ones.
There are three types of dyadic relationships in directed networks:

- mutual dyads, in which a tie from i to j is accompanied by one from j to i;
- asymmetric dyads in which there is a relationship between i and j in one direction, but not the other;
- null dyads in which there is no tie in either direction.

If all ties in a binary network are either mutual or null, the network is said to be symmetric, so the adjacency matrix $\mathbf{A}$ and its transpose $\mathbf{A}'$ are identical.
In undirected and binary networks, triads may include $0, 1, 2, or\ 3$ relationships.
In the case of three relationships, **triads** are said to be **closed or transitive**.
In the first case, each pair of actors are linked by a direct tie.
By contrast, each pair of actors are linked by an indirect path through the third actor.

1.4 A brief history of statistical network models

In this paragraph an overview of the historical development of statistical network modeling is presented.
The subsequent discussion focuses on a number of prominent static and dynamic network models and their interconnections.
The development of statistical methods for social networks has been verified in the last eighty years.
Three generations of research about statistical models in social network analysis can be distinguished.
This difference among models depends on the substantive research questions and the nature of the available data.

1.4.1 The first generation

Beginning in the late 1930's, the first generation of research dealt with the distributions of various network statistics.
A hundred years before Moreno, the French sociologist Durkheim has argued that human societies are more than simply a sum of various parts and, like biological systems, they are made up of interrelated components. In particular, he argues that any social phenomenon can only be understood in relation to others and to wider social context.
As such, the reasons for social regularities are to be found not in the intentions of individuals but in the structure of the social environments in which they were embedded.
In his famous study on suicide (1897), he states that one social phenomenon, i.e. suicide, can only be understood by looking at how individuals are embedded within a larger social system.
This abstract social structure becomes tangible, through Moreno and Jenning's sociometry, presenting one of the first statistical analysis of social network.
It has been *the mathematical study of psychological properties of populations, the experimental technique and the results obtained by application of quantitative methods* [123].
The approach also has made use of sociograms - diagrams of points and lines used to represent relations among people - which are visual depictions of individuals and their relationships to others in a group, representing a precursor to the graph representation for networks.
They have simulated a random network process by randomly assigning process to individual actors, obtaining the first simulation of a random digraph

distribution [155].

Later, many researchers have experienced the difficulty in using sociograms, once the network in question reached beyond a certain size.

So the introduction of using matrices has been necessary for structuring and analysing network data (e.g. [61]).

In the 1940s and 1950s, research in social networks has advanced along several directions.

One of the latter has been the development of a mathematical structure with Moreno's sociograms using matrix algebra and graph theory, in order to make it possible to discover emergent groups in network data [111].

Another direction has been de Sola Pool and Kochen's work [35] which have analysed the "small world" problem, i.e short paths of connections linking most people in social spheres, developing a program of laboratory experimentation on networks.

Twenty years later, Stanley Milgram had tested their propositions empirically, leading to the now popular notion of "six degrees of separation", i.e. the shortest path between any two people for completed chains has a median length of around six [35].

By the 1960s, the anthropologists begin to see societies as a *pattern or network (or system) of relationships obtaining among actors in their capacity of playing roles relative to one another* [124].

In the 1970s, the center of gravity of network search has shifted to sociology with White and other researchers, which focused attention to the importance of multiple relations, moving the term social networks from a metaphor to an analytical concept.

The anthropologists' attention is directed to ego networks highlighting issues of multiple relations and how such relations enabled or constrained

individuals.
By contrast, the attention of White and other researchers is directed to complete networks enabling the analysis of individuals within the context of the overall social network, increasing the range of analytical possibilities.
They have placed their work alongside the balance theory, because they have used the position balance theory as one of the structural hypotheses that block modelling could test for.
The goal of blockmodelling is to uncover a number of structural features of networks, by using matrices and matrix algebra.
In particular, the matrices are restructured in such a way so that actors who share a similar set of incoming or outgoing ties to others were grouped together into one block within a matrix, in which the nodes represented structural positions rather than individuals. This similarity is defined as structural equivalence.
Another key contribution is the influential strength of weak ties theory developed by Granovetter [78]. Granovetter states that social ties could be distinguished according to their strength, i.e. in weak and strong ties. Tie strength is a combination of an amount of time, the emotional intensity, intimacy and the reciprocal services that characterize the tie [78].

1.4.2 The second generation

The social science network research community, i.e. the second generation, begins in the 1970's and continues into the 1980's.
It is built upon earlier efforts, in particular the Erdös Rényi-Gilbert model, engendering the field of random graph theory.
Two mathematicians, Pál Erdös and Alfréd Rényi, have played an impor-

tant role in understanding the properties of random networks, merging probability theory and combinatorics with graph theory, establishing **random graph theory**, a new branch of mathematics [9].
They have introduced the independence digraph [53], sometimes called a Poisson random graph or Binomial random graph, in which all edges are independent and identically distributed but the graph is conditioned to have a specific degree sequence.
This model is typically denoted $\mathcal{G}$ (n, p), where n represents the number of vertices and p the probability that an edge (i, j) exists, for all i, j.
The presence of each possible tie is independent with $Y_{ij} \sim Bernoulli(p_{ij})$ where $\mu_{ij} = log(p_{ij})$ denotes the logarithm of the probability of a tie from i to j.
Enforcing a homogeneity assumption $\mu_{ij} = \mu$ for all i and j, this is simplified to a single-parameter model, under which the probability distribution of possible networks:

$$Pr(Y = y; \mu) = \exp(\mu t_1(y))(1 - exp(\mu))^{n(n-1)t_1(y)} \tag{1.8}$$

depends only on the network statistic $t_1(y) = \sum_{i;j} y_{ij}$, that represents the total number of ties [127].

The random network model was introduced by Gilbert [76] the same year in which Erdös and Rényi have published their first paper on the subject.
In those years it had been possible to see an increasing interest in developing statistical models for the analysis of social network data. The most famous of these was the family of models, referred to as exponential random graph models (ERGMs) that include the $p1$, $p2$ and p^* models.
The social network, in these models, is treated as the dependent variable

and thus the analyst wants to explain the network structure.
The exponential random graph family, given the issue of interdependence among network actors, addresses it by making use of an exponential function of a linear set of parameters, because it is impossible to make use of theoretical distributions (e.g. normal curve).
The $p1$ model is created by Holland and Leinhardt. It is an extention of the Erdös-Rényi - Gilbert model to permit looking at the role of sender and receiver effects, including reciprocity.
They make the assumption that all pairs of actors are independent of one another, in order to allow easy computation of maximum likelihood estimates using a contingency table formulation of the model [58].

This model includes parameters for tie density, the propensity for reciprocity of ties, and individuals' tendencies to express and receive ties.
They introduces the $p1$ probability density by including the homogeneity conditions $\mu_{ij} = \mu$ and $\rho_{ij} = \rho$ for all i and j, and treating the sets of parameters α_i and Υ_j as fixed effects:

$$p_1(y) = Pr(Y = y) = \exp\{(\mu t_1(y) + \sum_{i}^{n} \alpha_i t_{2i}(y) + \sum_{j}^{n} \Upsilon_j t_{3j}(y) + \rho t_4(y)\}/\kappa(\theta) \tag{1.9}$$

where network statistics $t_{2i}(y)$, $t_{3i}(y)$, and $t_4(y)$ respectively refer to the outdegree of actor i, the indegree of actor j, and the number of mutual dyads and $\kappa(\theta)$ is a normalizing constant.
It is restrictive because they consider only network statistics corresponding to configurations of one or two actors, when in reality, other types of rela-

tions exist, such as multiple dyads, i.e. transitivity or closure.
In addition, in the mid 1970's there is a big growth of the triadic analyses in order to study structural balance (for more details see [85]; [125]) and transitivity theory (or more details see [90]).
The first social network methodology is represented by the researches of Davis, Holland and Leinhard, that provided strong statistical evidence about transitivity that is a very important structural tendency in social network ([91]; [32]).
Davis shows that a basic feature of many social networks is the tendency towards transitivity (friends of my friends are my friends), so that Holland, Leinhardt and Johnsen substain that it is very important to test it by examining triads and the triples that they contain.
This model also allows various generalizations to multidimensional network structures [152] and stochastic blockmodels([92]; [154]).
The $p2$ model [152] can be seen as an extension of the well-known p, taking into account the dependent nature of the data and the relation with explanatory variables.
This extension to a Generalized Linear Mixed Model (GLMM) allows the inclusion of covariates, and models the remaining variability by random effects.
In specific, it is a random effects model with covariates for the analysis of binary dyadic data that represent a social network or directed graph, using nodal and/or dyadic attributes as covariates.
Like the $p1$ model, the $p2$ model does not explain the network structure very much, i.e. it is very limited. The model tries to explain how the number of ties found in a network and how the actor and dyadic covariates can explain outdegree, indegree and reciprocity.

It is controlled by checking the differences between actors' degree scores and reciprocity.
One social network at a time is analyzed in this model, but analyzing multiple networks simultaneously, can provide greater generalizability of research results compared to analyses of single-network data.
The *multilevel p2* model estimates the parameters more efficiently respect to $p2$ model and quantifies the differences between networks by modeling the variability of parameters over networks [166].
In the $p2$ model [152], the tie variables are regressed on explanatory variables, while the dependence of ties from and to the same actor are modeled using random effects.
The *multilevel p2* model defines an identically specified $p2$ model with varying parameters for multiple independent social networks.
By the 1980s, social network analysis has become an established field within the social sciences, with a professional organization (INSNA, International Network for Social Network Analysis), an annual conference (Sunbelt), specialized software (e.g., UCINET), and its own journal (Social Networks).

1.4.3 The third generation

The third generation begins with Exponential Random Graph Models (ERGMs), so called because they have an exponential term on the right hand side. They are also commonly called the p^* class of models ([62];[130]; [134];[156]).
The distinctive feature of ERGMs is that the assumption of independence among network tie variables may be relaxed, allowing different assumptions on the dependencies among network variables to be incorporated [30].
Frank and Strauss [62] introduce the Markov dependence, in which a possible tie from i to j is assumed to be contingent on any other possible ties

involving i or j, even if, the status of all other ties in the network is known. Markov dependence can be characterized as the assumption that two possible network ties are conditionally dependent when they have a common actor.

In the specific, a random undirected graph is a Markov graph iff its probability distribution can be written as:

$$Pr(Y = y; \theta) = \kappa(\theta)^{-1} \exp\Big(\sum_{(k-1)}^{(n-1)} \theta_k S_{3:k}(y) + \tau S_4(y)\Big) \tag{1.10}$$

where $S_{3:k}(y)$ are the number of k-stars and $S_4(y)$ is the number of triangles.

Using appropriate network statistics recognizing directionality, this model is generalized to directed networks [127].

This vision is more restrictive because the effects of a given configuration do not depend on only network isomorphism but it can vary with characteristics of actors.

The Markov dependence has inspired Wasserman and Pattison [156], which further have extended this class of models, describing them as p^* models.

They make a general formula that can allow a wider array of network statistics in the equation, than those specified by Frank and Strauss [62], such as reciprocity, the presence of edges, the number of closed triads and the number of two-paths or two-stars.

They have showed how a Markov parametric assumption provides just one of many possible sets of parameters.

Most initial investigations focus on undirected and directed single, dichotomous relations.

Robins *et al.* [134] Koehly and Pattison [98] propose the generalization of

p^* models to valued relations and to more than one relation.
Frank and Strauss' approach is more restrictive because the effects of a given configuration do not depend on only isomorphism within the network but it could vary with the characteristics of actors.
This type of dependence is called Markov attribute dependence [136] where a configuration's effect may depend only on attributes of those actors involved in it, so that, the parameter for the density configuration y_{ij} might depend on attributes of actors i and j, but not on those of actors $k \neq i$, j.
The effect of any network configuration may depend on actor attributes, but applications focus on the density effect [127].
Markov dependence seems unrealistic for large networks, where individual actors may not even be aware of each other, and have no means to come into contact, yet their possible tie is still taken to influence other possible ties.
When in an ERGM there is at least one pair of ties that do not share an actor, this model becomes non - Markovian.
This is verified when there are configurations involving four or more actors, for example a k-path (indirect path of length k) and k-cycle ($k > 3$), in which a sequence of k ties involving k distinct actors begins and ends with the same actor.
Or, in according to Pattison and Robins [131], there are non-Markov dependencies among ties that do not share an actor but may be interdependent through third party links.
For instance, Y_{ij} may be conditionally dependent on Y_{rs} for four distinct actors if there is an observed tie between i or j or r or s.
These realization-dependent models can be developed through what Pattison and Robins (2002) describe as partial dependence structures. These

models also permit the introduction of more complex configurations involving attribute effects.

An other important approach is Social Influence Model Approach ([135]; [68]).

This approach generalizes the p^* class of models for social network data to predict individual-level attributes from network ties.

The p^* model for social networks permits the modeling of social relationships in terms of particular local relational or network configurations.

Through these models, attribute variables are included in a directed dependence graph and the Hammersley-Clifford theorem is employed to derive probability models whose parameters can be estimated using maximum pseudo-likelihood.

In the 1990s, the development of statistical methods for social networks make consistent progress towards the last decades.

This development depends on the substantive research questions and the nature of the available data.

For instance, in the last decade there has been a big increase of more complex statistical models of network data, considering, for example, the dynamic nature of social networks or the relationship between networks data, attribute data and behavioral or attitude data.

Several works can witness the paradigm change.

Stochastic actor-based models for network dynamics ([93]; [139]; [143]) are developed in this framework to analyze longitudinal data on social networks while changing actors' attribute data.

For example, Snjiders *et al.* [142] have studied the co-evolution of network dynamics of friendship and delinquency using such models while Steglich *et al.* [144] have studied the co-evolution of friendship networks and smoking

behavior.
When a social network emerges in presence of a hierarchical structure, i.e. when the units are nested in clusters, a proper modelization can be represented by the Multilevel Modeling approach.
Multilevel models ([141]; [34]) are very useful for investigating the nature of connections in ego networks when there is the assumption that one's alters do not overlap with alters of an other ego.
A clear evolution of this paradigm can be envisaged through the collection of essays edited by Patrick Doreian and Frans Stokman ([146], [147], [49]) devoted to network dynamics.
An important distinction among the different models is the possibility to use (or not) inferential procedures to evaluate statistical significant effects.

1.4.4 A focus on Network Autocorrelation Models

A general representation of dependencies in data is the latent space model of Lazarsfeld and Henry [105], where the goal is to parsimoniously represent dependencies between multiple variables within individuals.
In this category, we can find the stochastic block-model by Holland *et al.* [92], Snijders and Nowicki [140], Nowicki and Snijders [126] and Daudin *et al.* [31].
Other types of models are the latent class clustering models (Hoff *et al.*, [88]) in which the probability of a link between pairs of actors depends on the distance between them and on their observed characteristics.
Extensions are in Hoff [87] and Handcock *et al.* [82].
Considering our peculiar research interest, we focus on Network Autocorrelation Models ([40]; Leenders [108]), developed from the geographical

connections models by Cliff and Ord and other authors ([26]; [5]; [28]).
These models can be extended to settings in which the autocorrelation stems from social, not physical, proximity [158].
Autocorrelation across units is not a nuisance to be removed but a substantive effect to be measured and tested [41].
There are two types of Network Autocorrelation Models: the Network Effects Model [42] and the Network Disturbances Model [44].
The **Network Effects Model** assumes that the dependent variable, for a given individual, is a function of the exogenous variables and the values of the same dependent variable observed on other individuals.
In particular, Doreian [43] defines the Network Effects Model where social dependence is incorporated through the addition of a lagged dependent variable on the right-hand side of the regression equation.
The outcomes for actors are not statistically independent as assumed by many regression models, leading to a complex correlation structure.
These models use a $n \times n$ network matrix of interdependencies, that is an adjacency matrix, to model this correlation structure [127].
Furthermore, attributes data may induce similarities among units and can be used as explanatory variables in the regression model as well.
Thus, it seems that autocorrelation models of contextual effects are best suited to theories of specific network processes [50].
This is the model:

$$y = \rho Ay + X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I) \tag{1.11}$$

where ϵ denotes a vector of stochastic errors, under usual assumptions, ρ is a scalar that measures the magnitude of the network effects, and β is a vector of regression coefficients ([4]; [41]; [79]; [80]; [121]; [159]).

In this type of model the actor outcome can depend directly on the outcomes of its own alters.
The vector $\mathbf{A}y$ containes, for each focal actor, the value of the outcome sum for all its alters, as such, alters' outcome contributes to y in proportion to the influence on ego.
Thus, $\mathbf{A}y$ is a **network effects dependent variable**.

In the case of Network Disturbances Model [44] the specification of the linear equation is

$$y = X\beta + \epsilon \tag{1.12}$$

but the disturbance is specified as:

$$\epsilon = \rho A\epsilon + \nu \tag{1.13}$$

with $\nu \sim N(0, \sigma^2 I)$ representing the white noise disturbance terms and parameter ρ measures the strength of the network autocorrelation.
In this case, the errors ϵ, rather than the outcomes y themselves, may be interdependent and the network autocorrelation can be modeled via inclusion of a term $\bar{\epsilon}_A = A\epsilon$ in the specification of the distribution of the error term.
The vector $\bar{\epsilon}_A$ containes, for each focal actor, the (weighted) average stochastic errors for the alters.
Under the common assumption that the errors ϵ are stochastically independent by the explanatory variables X, the network autocorrelation term $\bar{\epsilon}_A$ is likewise independent by X.
The implied mean vector and covariance matrix of ϵ are respectively 0 and $var(\nu)\{(I–\rho A')(I–\rho A)\}^{-1}$.

The model, in this case, may be rewritten as:

$$y = \rho Ay + X\beta – \rho AX\beta + \nu \tag{1.14}$$

This differes from precedents only by the addition of the network lagged covariate term $\rho AX\beta + \nu$, which measures the effect of other actors' covariates on the outcome for an actor.

A natural generalization combining Network Effects and Network Disturbances Models can be constructed as well as models with multiple autocorrelation regimes [45], i.e. models may also be specified using both $\mathbf{A}y$ and $\mathbf{A}\epsilon$.

The following regression model containes both autoregressive outcomes and network autocorrelation [5]; [16], allowing for different adjacency matrices for the two:

$$y = \rho_1 A_1 y + X\beta + \epsilon \quad \epsilon = \rho_2 A_2 \epsilon + \nu \tag{1.15}$$

where A_1 and A_2 are the adjacency matrices for the network effects and network autocorrelation effects, respectively.

This model includes two sources of correlation in y and in $X\beta$.

The substantive choice between modeling "contagion" through either autocorrelating the dependent term or the disturbance term reflectes a theoretical difference of how contagion was supposed to take place.

Furthermore, social dependence is analyzed in the structural equations modeling framework.

This alternative approach, introduced by Folmer and Oud [129], considers the presence of latent variables.

This approach illustrates social dependence through the network lagged variables as latent variables in the structural model, while relationships

among network lagged variables and observed attribute data are represented in the measurement model [109].
A recent approach is present in Doreian *et al.* [48] considering the covariance based SEM method.
In this thesis we deal with network data and structural equation models looking at the component based SEM method, i.e. Partial Least Squares-Path Modeling.

Chapter 2

Partial Least Squares Path Modeling: Notations and Definitions

2.1 A brief review of PLS - Path Modeling

The first work of the Partial Least Squares (PLS) approach to path models with latent variables (LVs) was published by Wold in 1979.
It was proposed as a component-based estimation procedure different from the classical covariance-based LISREL approach (SEM-ML).
Thus, Herman Wold opposes SEM-ML [96] "hard modeling" to PLS "soft modeling" [149].
Partial Least Squares-Path Modeling (PLS-PM) is considered as a soft modeling approach, where it *involves no assumptions about the population or scale of measurement* ([60]; [54]) and consequently works without distributional assumptions and with nominal, ordinal, and interval scaled variables, but some assumptions must be fulfilled, i.e. Gaussian classical linear

ordinary least squares and predictor specification [25]. This specification required that the systematic part of the linear regression must be equal to the conditional expectation of the dependent variable. PLS Path Modeling aims to estimate the relationships among $Q(q = 1, \ldots, Q)$ blocks of manifest variables (MVs) which are indicators of unobservable constructs, usually called LVs.

Let be P variables $(p = 1, \ldots, P)$ observed on n units.

The resulting data are collected in a partitioned data table X:

$$X = [X_1, \ldots, X_q, \ldots, X_Q] \tag{2.1}$$

where X_q is the generic q-th block made of P_q variables.

These types of variables are also known as indicators or items that assume as manifest variables containing information, reflecting one aspect of the construct; hence, we use the information contained in the indicators to obtain an approximate representation of the latent variable.

In PLS - Path Modeling an iterative procedure permits to estimate the outer weights (w) and the latent variable scores $(\hat{\xi})$ solved through alternating single and multiple linear regressions.

At a later stage the path coefficients (β) are estimated by means of a regular regression between the estimated latent variable scores in accordance with specific structural relations.

The PLS-Path Modeling follows the SEM notations and symbols, including the use of a path-diagram to picture the relations among the latent variables and between each manifest variable and the corresponding latent variable. Namely, the p manifest variables are pictured by rectangles or squares, while circles represent the q latent variables.

Arrows define the relations among latent and/or manifest variables.

As in SEM, even in the PLS-PM, the overall relations among manifest and latent variables are modeled through a system of equations.
The goal of PLS-PM is not the reproduction of the sample covariance matrix, unlike the classical covariance-based approach.
For this reason, PLS - Path Modeling is considered as an exploratory approach more than as a confirmatory one [54].
This involves that the classical parametric inferential framework is replaced by resampling methods such as jackknife and bootstrap through empirical confidence intervals and hypothesis testing procedures ([24]; [149]; [54]).
There is a lack of important statistical properties for the estimates, i.e. coefficients are known to be biased but consistent at large ([22];[23]; [54]).

2.2 The PLS path model

Two sub-models compose a Structural Equation Model:

1. The **measurement model or outer model**;
2. The **structural model or inner model**.

The first one analyses the relationships between each latent variable and its manifest variables, while the structural model analyses the relationships among the latent variables.
A latent variable (LV) is called **endogenous**, if it is supposed to depend on other LVs and **exogenous** one otherwise.
Another important element is represented by the **weight relations**, that are used to estimate case values for the latent variables [25].
The crucial part of a PLS-PM is the estimation of the weight relations by using a two-step algorithm to determine them: an outside and inside

approximation.
The description of both these two steps is presented in the next section.

2.2.1 Structural Model

In the PLS Path Modeling framework, the **structural model** can be written as:

$$\xi_j = \sum_{(q:\xi_q \to \xi_j)} \beta_{qj}\xi_q + \zeta_j \tag{2.2}$$

where ξ_j is an endogenous latent variable, β_{qj} is the path coefficient linking the exogenous q-th latent variable to the j-th endogenous one and ζ_j is the error in the inner relation.

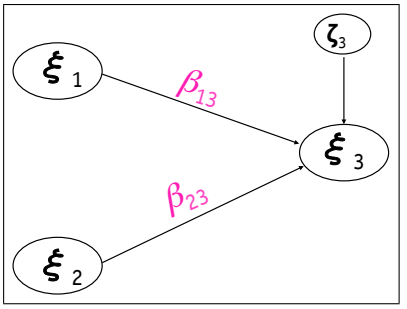

Figure 2.1: Structural model in a path diagram

The only hypothesis of this model is what Wold named *predictor specification hypothesis* [162]:

$$E(\xi_j|\xi_q) = \sum_{(q:\xi_q \to \xi_j)} (\beta_{qj}\xi_q) \tag{2.3}$$

which implies that $cov(\xi_q, \zeta_j) = 0$ and $E(\zeta_j) = 0$

2.2.2 Measurement Model

The **measurement model** formulation depends on the direction of the relationships between the latent variables and the corresponding manifest variables ([60]; [54]).

There are three types of measurement model that relate the MVs to their LVs:

1. Reflective model (or outwards directed model);
2. Formative model (or inwards directed model);
3. MIMIC model (a mixture of the two previous models).

- *Reflective model*

In the reflective model, each manifest variable reflects the corresponding latent variable.

In this case, it assumes that the block of manifest variables related to a latent variable measures a unique underlying concept and the indicators linked to the same latent variable should covary: changes in one indicator imply changes in the others [54].

For this reason the internal consistency has to be checked, i.e. each block has to be homogeneous and unidimensional, in order to reflect a unique latent construct [54].

The measurement model reproduces the factor analysis model so in the reflective model each variable is a function of the underlying factor.

In a reflective model, using formal terms, each relation between each manifest variable (MV) $x_{pq}(p = 1, \ldots, P_q)$ and the corresponding LV is generally modeled as a simple regression model, i.e. :

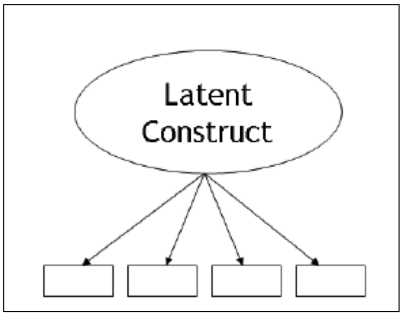

Figure 2.2: Reflective model in a path diagram

$$x_{pq} = \lambda_{pq}\xi_q + \epsilon_{pq} \tag{2.4}$$

where λ_{pq} is the loading associated to the p-th manifest variable in the q-th block and ϵ_{pq} represents the imprecision in the measurement process with the predictor specification hypothesis:

$$E(x_{pq}|\xi_q) = \lambda_{pq}\xi_q \tag{2.5}$$

In addition to these two equations in PLS-PM each latent variable is defined as a linear combination of the corresponding manifest variables. Thus, each latent variable is obtained as:

$$\xi_q = \sum_p w_{pq}x_{pq} \tag{2.6}$$

where w_{pq} is the outer weight associated to the generic manifest variable x_{pq}. This equation is referred as weight relation [55].

There are several tools for checking homogeneity and unidimensionality of a block:

1. Cronbach's α;
2. Dillon-Goldstein's (or Joreskög's) ρ ;
3. Principal component analysis of a block.

Cronbach's $\boldsymbol{\alpha}$ is a measure of internal consistency that quantifies unidimensionality of a block of variables. This index can be expressed as:

$$\alpha = \frac{\sum_{(p \neq p')} cor(x_{pq}, x_{p'q})}{P_q + \sum_{(p \neq p')} cor(x_{pq}, x_{p'q})} \times \frac{P_q}{P_q - 1} \tag{2.7}$$

where P_q is the number of manifest variables in the q-th block.
The larger is the $\sum_{(p \neq p')} cor(x_{pq}, x_{p'q})$ the more the block is unidimensional [149]. A block is considered unidimensional if this index is larger than 0.7 for confirmatory studies.

Dillon-Goldstein's (or Joreskog's) $\boldsymbol{\rho}$ [157] is better known as composite reliability.
It supposes that the correlation between each MV x_q and its LV ξ_q is positive.
For this reason the block is considered as homogenous as $\sum_{(p=1)}^{P_q} \lambda_{pq}$ is large.
The Goldstein- Dillon's ρ is defined by:

$$\rho = \frac{(\sum_{(p=1)}^{P_q} \lambda_{pq})^2}{\sum_{(p=1)}^{P_q} \lambda_{pq})^2 + (\sum_{(p=1)}^{P_q} (1 - \lambda_{pq}^2)} \tag{2.8}$$

A block is considered unidimensional when the Dillon-Goldstein's ρ is larger than 0.7. According to Chin [24], this statistic is considered to be a better indicator of the unidimensionality of a block than the Cronbach's α.

The first statistic assumes that each manifest variable is equally important in defining the latent variable.
In Dillon-Goldstein's ρ, by contrast, this assumption does not hold because it is based on the loadings of the model rather than the correlations observed between the manifest variables in the dataset. With a **Principal component analysis of a block**, a block may be considered unidimensional if, according Kaiser's rule, the first eigenvalue of its correlation matrix is higher than 1, and the second one smaller than 1, or at least very far from the first one [149]. In order to assess whether the eigenvalue structure is significant or rather is due to sampling fluctuations a bootstrap procedure can be implemented. In case the hypothesis of unidimensionality is rejected, it is possible to identify some groups of unidimensional sub-blocks by considering the variable-factor correlations displayed on the loading plots. PLS path modeling is a mixture of a priori knowledge and data analysis. In the reflective way, the a priori knowledge concerns the unidimensionality of the block and the signs of the loadings and the data have to fit this model. If they do not, they can be modified by removing some MVs that are far from the model. Another solution is to change the model and use the formative way.

- *Formative Model*

In the formative model, the LV is supposed to be generated by its own MVs, i.e each manifest variable or every set of manifest variables represents a different level of the underlying latent concept. This model does not assume homogeneity nor unidimensionality of the block, for this reason the block of MVs can be multidimensional and the indicators need not to co-vary.

Thus the measurement model could be expressed as:

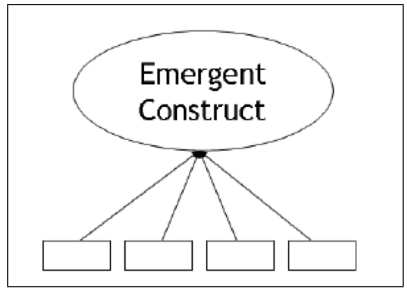

Figure 2.3: Formative model in a path diagram

$$\xi_q = \sum_{p=1}^{P_q} w_{pq} x_{pq} + \delta_q \tag{2.9}$$

where w_{pq} is the coefficient linking each manifest variable to the corresponding latent variable and δ_q is the error that represents the part of the latent variable not explained by the block of manifest variables. The assumption behind this model is the following predictor specification:

$$E(\xi_q | x_{pq}) = \sum_{p=1}^{P_q} w_{pq} x_{pq} \tag{2.10}$$

which implies that residual vector $E(\delta_q) = 0$ and is uncorrelated with the MVs.

- *MIMIC Model*

The MIMIC model is a mixture of the reflective and formative models. The scores of the standardized latent variable $\hat{\xi}_q$ associated to the q-th latent variable ξ_q are computed as a linear combination of its own block of manifest variables by means of the weight relation defined as:

$$\hat{\xi}_q = \sum_{p=1}^{P_q} w_{pq} x_{pq} \tag{2.11}$$

where the variables x_{pq} are centred and w_{pq} are the **outer weights**.

2.3 The Partial Least Squares Algorithm

The PLS algorithm includes the following three stages:

1. iterative approximation of latent variable scores;
2. estimation of latent variable scores;
3. estimation of path coefficients.

The debate of convergence of the PLS algorithm focuses on the core of the PLS algorithm, i.e. the first stage.

2.3.1 The first stage: Iterative process

The **first stage** of the PLS path modeling algorithm consists of four steps [149]:

- Step 0: Initial arbitrary outer weights;
- Step 1: Compute the external approximation of latent variables;
- Step 2: Obtain inner weights;
- Step 3: Compute the internal approximation of latent variables;

- Step 4: Calculate new outer weights;

- Repeat step 1 to step 4 until convergence of outer weights.

Step 0: Initial arbitrary outer weights

We start the iterative process by assigning any arbitrary non-trivial linear combination of indicators can serve as an outer proxy of a latent variable [86].

Step 1: Outer approximation of the latent variable scores

Outer proxies of the latent variables are estimated as a linear combination of its own manifest variables.

$$\nu_q \propto \pm \sum_{p=1}^{P_q} w_{pq} x_{pq} = \pm X_q w_q \tag{2.12}$$

where ν_q is the standardized outer estimate of the q-th latent variable ξ_q; the x_{pq} are centred MVs, the symbol $\propto$ means that the left side of the equation corresponds to the standardized right side and the "$\pm$" sign shows the sign ambiguity.
This ambiguity is solved by choosing the sign making ν_q positively correlated to a majority of x_{pq}.

Step 2: Estimation of the inner weights

Two LVs are adjacent if exists a link between the two variables: an arrow goes from one variable to the other in the causality path-diagram.

Inner weights are calculated for each latent variable in order to reflect how strongly the other latent variables are connected to it, considering the existing links with the other Q' adjacent latent variables:

$$z_q \propto \sum_{q'=1}^{Q} d_{qq'} e_{qq'} \nu_{q'} \tag{2.13}$$

Let $d_{qq'}$ be the generic element of the square matrix D of order Q, where $d_{qq'} = 1$ if the LV ξ_q is connected to $\xi_{q'}$ in the path diagram and $d_{qq'} = 0$ otherwise.
The inner weights $e_{qq'}$ can be determined through one of three different schemes available:

1. the **centroid scheme**, i.e. Wold original scheme, where

$$e_{qq'} = sign\ cor(\nu_q, \nu_{q'}) \tag{2.14}$$

2. the **factorial scheme**, i.e. the Lohmöller scheme, where

$$e_{qq'} = cor(\nu_q, \nu_{q'}) \tag{2.15}$$

3. the **path weighting scheme** or structural scheme, where LVs connected to ξ_q are divided into two groups:

$$e_{qq'} = cor(\nu_q, \nu_{q'}) \ \ if\ \nu_{q'}\ predicts\ \nu_q \tag{2.16}$$

$$e_{qq'} = the\ regression\ coefficient\ \ if\ \nu_{q'}\ is\ predicted\ by\ \nu_q \tag{2.17}$$

They are described in details in the next section.

Step 3: Inner approximation of the latent variable scores

Inner proxies of the latent variables are calculated as linear combinations of the outer proxies of their respective adjacent latent variables, using the inner weights previously determined.

Step 4: Estimation of the outer weights

The calculation of outer weights depends on the type of relation existing between a block of MVs and the underlying LV. Estimation of the outer weights w_{pq} depends on the chosen model.

In the **reflective model**, they are calculated as the covariances between the inner proxy of each latent variable and its indicators. In the outer estimate of the LV, it is the regression coefficient of the simple linear regression of each MV on the inner estimate of the corresponding LV:

$$w_{pq} = cov(x_{pq}; z_q) \tag{2.18}$$

taking into account that z_q is standardized.
This outer estimation is named **Mode A** in PLS-PM literature [149].
The regression coefficient reduces to the covariance between each manifest variable and the corresponding inner estimate of the latent variable.
In case the manifest variables have been also standardized, such a covariance becomes a correlation.

In the **formative model**, they are the regression coefficients in the multiple regression of the inner estimate z_q on its MVs X_q, i.e., the elements of the vector:

$$w_q = (X_q' X_q)^{-1} X_q' z_q \tag{2.19}$$

where X_q comprises the P_q manifest variables x_{pq} previously centred and scaled by $\sqrt{(\frac{1}{N})}$. This scheme is called **Mode B**.
These four steps are repeated until the change in outer weights between two iterations drops below a predefined limit, i.e. the convergence is measured in terms of stability of the numerical values over two successive iterations.

Convergence algorithm

The convergence of the iterative PLS Path Modeling algorithm is verified according to a stopping rule, most often defined as

$$\max(|w_{pq}^{(s)} - w_{pq}^{(s-1)}|) < 10^{-5} \tag{2.20}$$

where s refers to the s-th iteration. It is stated that "convergence is always verified in practice" [81] and is "guaranteed only for the two-block case, but practically always encountered in practice even with more than two blocks" [149]. Even though the PLS path modeling algorithm may converge in practice, there is concern about the missing proof of its convergence [95], inspiring researchers to search for this proof [86].

2.3.2 The second stage: Computation of the latent variable scores

Upon convergence, the estimates of the latent variable scores are obtained as:

$$\hat{\xi}_q \propto X_q w_q \tag{2.21}$$

2.3.3 The third stage: Computation of path coefficients

In the last stage of the PLS-PM algorithm path coefficients are estimated through OLS multiple regressions among the scores of estimated latent

variable:

$$\hat{\beta}_j = (\hat{\Xi}'_{\to j}\ \hat{\Xi}_{\to j})^{-1}\ \hat{\Xi}'_{\to j}\ \hat{\xi}_q \tag{2.22}$$

where ξ_j is the generic endogenous LV score vector and $\hat{\Xi}_{\to j}$ is the matrix of the corresponding latent variable scores.

2.4 Model validation

Through validation process of the PLS-PM it is possible to calculate suitable indexes to measure its predictivity performances and fitting. According to PLS-PM structure, a path model can be validated at three levels [149]:

1. the quality of the measurement model;
2. the quality of the structural model;
3. each structural regression equation.

That is why, PLS Path Modeling provides **three** different **fit indices**:

1. the communality index;
2. the redundancy index;
3. the Goodness of Fit (GoF) index.

The **communality index** measures the quality of the measurement model for each block. It is defined, for block q, as:

$$Com_q = \frac{1}{P_q}\sum_{p=1}^{P_q} cor^2(x_{pq}, \hat{\xi}_q) \forall q : P_q > 1 \tag{2.23}$$

The **average communality** is the average of all the $cor^2(x_{pq}, \hat{\xi}_q)$ where P_q is total number of MVs in all blocks.
The **redundancy index** measures the quality of the structural model for each endogenous block, taking into account the measurement model.
It is defined, for an endogenous block , as:

$$\overline{Com} = \frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} cor^2(x_{pq}, \hat{\xi}_q) \tag{2.24}$$

where P_q is total number of MVs in all blocks.
The **redundancy index** measures the quality of the structural model for each endogenous block, taking into account the measurement model. It is defined, for an endogenous block, as:

$$Red_j = Com_j \times R^2(\hat{\xi}_j, \hat{\xi}_{q:\xi_q \to \xi_j}) \tag{2.25}$$

The **average redundancy** for all endogenous blocks can also be computed.

$$\overline{Red} = \frac{1}{J} \sum_{j=1}^{J} Red_j \tag{2.26}$$

A global criterion of **goodness-of-fit (GoF)** can be proposed [3] as the geometric mean of the average communality and the average R^2 :

$$GoF = \sqrt{\overline{Com} \times \overline{R^2}} \tag{2.27}$$

where the average R^2 value is obtained as:

$$\overline{R^2} = \frac{1}{J} R^2(\hat{\xi}_j, \hat{\xi}_{q:\xi_q \to \xi_j}) \tag{2.28}$$

As a matter of fact, differently from SEM-ML, PLS path modeling does not optimize any global scalar function so that it naturally lacks of an index that can provide the user with a global validation of the model (as it is instead the case with R^2 and related measures in SEM-ML).
The GoF represents an operational solution to this problem as it may be meant as an index for validating the PLS model globally. Since PLS Path Modeling has not distributional assumptions, its inferential tools are usually based on resampling techniques, i.e. cross-validation methods like jack-knife and bootstrap [52]. It is possible to build a cross-validated version of three fit indices by means of a blindfolding procedure ([24]; [110]; for more details see [149]).

2.5 New Approaches of PLS-PM

2.5.1 Mode PLS

In the last years, PLS Path Modeling (PLS-PM) has been reinterpreted. In the classical approach, in order to estimate the outer weights, which are very important for building the latent variable scores, we use two different estimation procedures based on OLS regression, i.e. Mode A and Mode B. These two modes assume a unique latent variable behind each block of manifest variables. An intermediate mode between Mode A and Mode B, i.e. **Mode PLS**, has been proposed in order to:

- estimate multidimensional latent variables;
- overcome multicollinearity problems that may lead to nonsignificant regression coefficients;

- have interpretable weights because of the difference in sign between the regression coefficient of an MV and its correlation with the LV.

It verifies when running a PLS regression and retaining a number for each block of significant PLS components [54].
PLS regression can nicely replace OLS regression for estimating path coefficients [54] whenever one or more of the following problems occur:

1. missing latent variable scores;
2. strongly correlated latent variables;
3. a limited number of units as compared to the number of predictors in the most complex structural equation.

With **Mode PLS** we search for m orthogonal PLS-R components, $t_{kq}(k = 1, \ldots, m)$ which are as correlated as possible to z_q and also explanatory of their own block X_q.
The number m of retained orthogonal components is either chosen by the cross-validation methods or defined by the user.
In the specific, two new modes have been proposed, which integrate a PLS Regression as an estimation technique, in order to estimate outer weights in PLS-PM:

1. the PLScore;
2. the PLScow.

In both procedures PLS Regression replaces OLS regression but there are some differences.
The **PLScore Mode** is oriented to maximizing correlations among latent

variables (LVs) and the PLS Regression is run under the classical PLS-PM constraint of unitary variance for the latent variable score.

The first PLS component t_{1q}, if x_{pq} are standardized variables, is defined with this formula (Esposito Vinzi et al., 2010b):

$$t_{1q} = X_q w_{1q} = \frac{1}{\sqrt{(\sum_p cor^2(z_q, x_{pq})}} \times \sum_p cor(z_q, x_{pq}) \tag{2.29}$$

When x_{pq} are not standardized variables, instead, the correlation is substituted by the covariance, the vector w_{1q} is normalized and a regression of z_q on t_{1q} is run. So the residuals z_{q1} and X_{q1} of the regressions of z_q on t_{1q} are calculated as:

$$z_{q1} = z_q - c_{1q} t_{1q} \tag{2.30}$$

and

$$X_{q1} = X_q - t_{1q} p'_{1q} \tag{2.31}$$

where c_{1q} is the regression coefficient from the regression of z_q on t_{1q} and p_{1q} is the vector of regression coefficients from the regression of the variables in X_q on t_{1q}. It is possible to define the second component as:

$$t_{2q} = X_{q2} w_{2q} = X_q w^*_{2q} \tag{2.32}$$

where w^*_{2q} is different from w_{2q} because the former refers to the original variables in X_q, while the latter refers to the residuals and would be very difficult to interpret. The next orthogonal components are defined by iterating the procedure described above on residuals from the previous component.

The **PLScow Mode**, by contrast, is oriented to maximizing covariances between LVs and the outer weights with the constraint of unitary norm according to classical normalization constraints of PLS Regression.
We have the same solution of the New Mode A (described in the next section) if we normalize the outer weights to unitary variance at each step of the algorithm PLScore Mode and we use a one-component PLS regression as the outer estimation mode.
If more components are considered, keeping the normalization constraint on the outer weights, PLScow Mode gives solutions between New Mode A (one PLS component) and a New Mode B (as many PLS components as there are MVs in a block).
These new modes are linked to the standard Mode A and Mode B outer estimates in PLS-PM as well as to the **New Mode A** proposed in a criterion-based approach by Tenenhaus and Tenenhaus [150].

2.5.2 The optimization criteria

Recent works by Hanafi [81], Kramer [100] and Tenenhaus and Tenenhaus [150] prove that the PLS-PM iterative algorithm optimizes different statistical criteria according to the different options chosen for the computation of the outer and inner proxies of the latent variables.

The outer proxies of the latent variables can be traditionally obtained through the choice of two different scheme: the Mode A (also referred as reflective model) and the Mode B (also referred as formative model) [149].

In particular, Hanafi [81] proved that the outer weights obtained through the PLS-PM algorithm maximize the following criteria when the mode B option is chosen for outer proxy computation:

$$\arg\max_{w_q} \left\{ \sum_{q \neq q'} c_{qq'} \cdot g\left(Corr(X_q w_q;\ X_{q'} w_{q'})\right) \right\} \quad st\ \ \|X_q w_q\| = 1 \qquad (2.33)$$

In 2007 Kramer showed that the PLS-PM algorithm was not based on a stationary equation related to the optimization of a twice differentiable function when Mode A was used for all the blocks in the model.
In the same work, Kramer proposed a slight modified version of the classical Mode A outer scheme in which a normalization constraint is put on outer weights rather than latent variable scores.
If this new scheme (also referred as New Mode A) was used for all the blocks in the model, Kramer [100] proved that the PLS-PM iterative algorithm was monotonically convergent to the criterion:

$$\arg\max_{\forall w_q} \left\{ \sum_{q} c_{qq'} \cdot g\left(Cov(X_q w_q;\ X_{q'} w_{q'})\right) \right\} \quad st\ \ \|w_q\| = 1 \qquad (2.34)$$

These recent works are very interesting from a theoretical point of view because they reduce the cases where the PLS-PM algorithm seems to be an heuristic approach at the case when the inner estimation takes explicitly into account the direction of the path weighting scheme [56]. Hence, PLS-PM algorithm seems to be an heuristic approach only when the path weighting scheme is used [55].

2.5.3 New Mode A Model

Tenenhaus and Tenenhaus [150] have slightly adjusted Mode A in which a normalization constraint is put on outer weights rather than on LV scores.

They showed that **Wold's procedure**, with the use of the **new Mode A** in all the blocks and the **centroid scheme** for the inner estimation of the LVs, monotonically converges to the criterion:

$$\underset{\|w_q\|=1}{\arg\max} \sum_{(q \neq q')} |cov(X_q w_q, X_{q'} w_{q'})| \tag{2.35}$$

Instead, when the **factorial scheme** is used for the inner estimation of the LVs, **Wold's procedure** converges to the criterion:

$$\underset{\|w_q\|=1}{\arg\max} \sum_{q \neq q'} c_{qq'} cov^2(X_q w_q, X_{q'} w_{q'}) \tag{2.36}$$

In the classical mode A the outer weights are computed with the formula $w_q = \frac{(X_q' z_q)}{\|X_q' z_q\|}$, but normalized so that the outer component is standardized. The new mode A shrinks the intra-block covariance matrix to the identity. This shrinkage is probably too strong, but is useful for very high-dimensional data because it avoids the inversion of the intra-block covariance matrix. When we use in the same model both new Mode A and Mode B and the centroid scheme, Wold's procedure is shown to converge to the criterion:

$$\underset{\|w_q\|=1}{\arg\max} \sum_{(q \neq q')} c_{qq'} |cor(X_q w_q, X_{q'} w_{q'}) \times \sqrt{var(X_{q'} w_{q'})^{\tau_{q'}}} \sqrt{var(X_q w_q)^{\tau_q}}| \tag{2.37}$$

While, when the factorial scheme is used, it converges to the criterion:

$$\underset{\|w_q\|=1}{\arg\max} \sum_{(q \neq q')} c_{qq'}\ cor^2(X_q w_q, X_{q'} w_{q'}) \times var(X_{q'} w_{q'})^{\tau_{q'}} var(X_q w_q)^{\tau_q} \tag{2.38}$$

In these equations we have:

- $\tau_q = 1$, i.e. using outer weights with unitary variance, when the block q is estimated by new Mode A, leading to criteria based on maximizing covariances among adjacent LVs;

- $\tau_q = 0$, i.e. using standardized LV scores, when the block q is estimated by Mode B, leading to criteria based on maximizing correlations among adjacent LVs.

This new estimation mode has the major advantage, as compared to classical Mode A, to maximize a known criterion.

Due to the good proprieties of the **New Mode A**, it is used in developing the component-based approach to network data through Partial Least Squares algorithms.
In the rest of this work, when referring to PLS-PM algorithm it always refers to the solutions obtained using the New Mode A for the outer proxies computation and the centroid scheme for the inner proxies computation.

Chapter 3

Modelling Network Data through Partial Least Squares Methodology

3.1 Theoretical background

Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual.
Anyone who either cannot lead the common life or is so self-sufficient as not to need to, and therefore does not partake of society, is either a beast or a god (Aristotle, Politics 350 B.C.).
Most people, usually, live in groups because it is rare that the lone individual has no connection to other men and women. Virtually all the activities of their lives - working, learning, worshiping, relaxing, playing, and even sleeping - occur in groups rather than isolated from others.
For this reason, actors adapt their behaviour, attitude, or belief, to the

behaviours, attitudes, or beliefs of other actors with whom are in contact [108].
The framework underlying this work is constitued by the relations in which individuals socially interact and like each other, involving a need for shared mutual understanding, which is expressed in similarities between two people ([18]; [78]; [94]; [101]; [164]).
Social researchers sought to understand which are the causes of similarity in these relations among people.
Laumann, in 1979, took to be the *hallmark of a network analysis...to explain, at least in part, the behaviour of network elements...by appeal to specific features of the interconnections among the elements* [103] .
It is necessary to consider that it is not possible to de-contextualize the social relationships because context can influence the emerging of relationships [47] .
This is verified because the **social context** of a social network is made up of the human and symbolic features that are intrinsic to situations where social network data are collected.
The focus here is on examining the impact of social context on social network structure.
The measurement of attitudes, behaviors, human features and structural characteristic of a context are retained in classical statistical variables (attributes data).
Roughly speaking, social network analysis often does not take into account actor's attributes. It wants to examine contemporarily both the social network and attributes of actors.

In social and behavioral sciences (e.g., psychology, sociology, economy), there are many concepts of theoretical nature, i.e. any variable that does

not correspond directly to anything observable must be considered as unobservable [39] or **latent** and that can not be obtained by means of a real-world sampling experiment [117].
For example, sociologists refer in these terms to social structure, social stratification and social status.
When we work with these two types of variables (i.e., developing theories and models) we tend to conceive expected causal relationships on them.
An alternative used by researchers is Structural Equation Modelling (SEM) ([7]; [97]; [12]).
The basic idea is that complexity inside a system can be studied taking into account a whole of causal relationships among latent concepts, called Latent Variables (LV), each measured by several observed indicators usually defined as Manifest Variables (MV).
Structural equation models typically do not take into account the network effects that appear as an important determinant of individuals' actions.
Some researches have dealt with methods of estimating the causal impact of network effects, once such a network has been made a part of a causal model. A mathematical formalization of the effects of social network on behaviors is given by the Network Effects Model (see par. 1.4.4.).
From an empirical point of view, these models are far from being directly observable. It still remains the possibility of measuring them as latent factors depending from multidimensional constructs. Putting all together, we propose a component-based approach to network data through Partial Least Squares-path model algorithms.
A simulation study is presented in order to compare the Network Effects Model with the proposed approach by examining coefficients estimated from the two methods while controlling for specific attributes: network size, net-

work density, autocorrelation coefficients and standard deviation of disturbances. Results will be discussed.

3.2 The effects of social networks on outcomes

In general, a social network represents any pattern of relationships between actors. Some examples can be friendship among adolescents, coauthorship among scientists, trade between countries and so on.
Recently, in many disciplines, a new theoretical vision tries to understand how social networks can influence outcomes.
We assume that the network is binary and observed at one point in time. In this case we consider an adjacency matrix **A** (see par. 1.2.3.). On the other hand, willing to assume the absence of transitivity in the network, e.g. in a friendship network, some friends of i's friends are not i's friends, their attributes will affect i's outcome only through their effect on i's friends' outcomes.
The common feature linking all of these examples is that the **units of analysis are interdependent**.
If we want to analyze the dependence of a variable from one or more independent or explanatory variables, we can try to describe the relation existing among them. A statistical tool useful to describe this type of relation is represented by regression models.
Regression models are particularly vulnerable, when there are interdependent units embedded within social structures, i.e. network effects ([44]; [70]).
If the interdependencies can be represented in the form of a network autocorrelation model, it is then possible to incorporate them into regression

type analyses.
In 1996, Doreian [44] proposed the re-examination of structural equation models whenever network autocorrelation is present.
Oud and Folmer [129] reclaimed this approach in 2008, within geographical framework, in order to represent spatial dependence.
One of the main issue of this thesis is to deal with dependence when **A** represents social distance and network data can be enter in a SEM model through an autocorrelation term.
The autocorrelation effect is represented through a scalar, i.e. ρ, estimating the extent to which an actor's outcome is affected by the behavior of those to whom is socially close.
In this approach, latent variables representing social dependence, have as indicators the observed values of the neighbouring social units.
A component-based approach to network data through Partial Least Squares-path model algorithms is proposed.
This model specification offers higher flexibility, it allows to consider i) separate or joint effect of intrinsic opinions of the social actors, ii) the extent to which they are influenced by their alters, and iii) how people with similar characteristics are more likely to form ties. This can be envisaged by different specifications of the path diagram.

3.3 Model specifications

The PLS Path Modeling will be now specified to include the Network Effects Model, defined in matrix notation as:

$$Y = \rho AY + X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I) \tag{3.1}$$

Let's assume that we have t variables (X and Y) measured on n observations and that - in the more general form - these variables can be divided in different blocks each associated with a well defined latent concept.
We will use the following notation in the paragraph:

- $\mathbf{C}_{(n,t)}$ is the data set containing n observations and t variables about attributes and outcomes;
- $\mathbf{A}_{(n,n)}$ is an adjacency matrix;
- $\mathbf{C}$ can be divided in Q (mutually exclusive) blocks $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_Q$;
- Each block $\mathbf{C}_q$ ($q = 1, \ldots, Q$) has p_q variables;
- In the $\mathbf{C}_Q$ block there are the p_Q outcome variables ($\mathbf{Y}$);
- Each block $\mathbf{C}_q$ is associated with a different latent variable ξ_q.

Following Doreian *et al.* [48] we include **the network-lagged variables** $\mathbf{A}Y$ as indicator of a latent variable in the measurement model, representing a bridge between statistical analysis and social theories.
Then we use this latent variable as an exogenous latent variable in the structural model.
In particular, we transform each outcome in $\mathbf{C}_Q$ through $\tilde{\mathbf{Y}} = \mathbf{A}\mathbf{C}_Q$, and we denote $\tilde{\mathbf{C}}_Q$ the block of the dataset $\mathbf{C}$ that contains the network effects dependent variables.

The **measurement model** related to a generic network effects dependent variable $\tilde{Y}_h$ ($h = 1, \ldots, p_Q$) can be written in matrix notation as:

$$\tilde{Y}_h = \lambda_h \xi_Q + \epsilon_h \tag{3.2}$$

where:

- λ_h is the loading associated to the generic network effects dependent variables;
- ξ_Q is the latent variable measuring the network effects, i.e. the latent variable associated to the network effects dependent variables, including in the Q - th block;
- ϵ_h is the error term that represents the imprecision in the measurement process.

The model in eq. 3.1 is modified in order to include the effects of the network effects dependent variables.
We refer to this new formulation of the structural model as the *Network Effects Structural Model*:

$$\xi_j = \sum_{q:\ \xi_q \to \xi_j} \beta_{qj} \xi_q + \rho_Q \xi_Q + \zeta_j \tag{3.3}$$

where:

- ξ_j is the endogenous latent variable;
- $\beta_{q\in(1:Q-1)}$ are the path coefficients linking the exogenous latent variables, associated to the $1, \ldots, (Q-1)$ blocks, to the j endogenous one;

- ρ_Q is the path coefficient linking the exogenous latent variable, associated to the block Q (i.e. to the network effects dependent variables $\tilde{\mathbf{Y}}$), to the endogenous latent variable ξ_j ;
- ζ_j is the error term in the inner relation.

We can notice that in the standard network effects model (3.1) only the average effect $\rho\mathbf{A}Y$ shows up. By contrast, in this new specification the social dependence is captured by two kinds of parameters:

- ρ_Q in the structural model;
- λ_h in the case of reflective mode of the measurement model.

We may argue that the latent variable approach offers a much richer representation of the social network structure than the standard network effects approach since it allows to analyze the relationship between the observed variables and the corresponding latent construct and further obtaining the latent variables scores.

3.4 The Partial Least Squares algorithm with Network Effects

There are two main procedures [81] of the PLS path modeling algorithm: the original and less known procedure invented by Wold ([162], [163]), and a modified procedure developed by Lohmöller [110].
The **PLS Path Modeling algorithm by Lohmöller** is the best-known procedure because computes the latent variable scores of each latent variable $\xi_q^{(s+1)}$ $(q = 1, \ldots, Q)$ at iteration $s + 1$ as a function of all the latent variable scores $\xi_q^{(s)}$ obtained during the previous iteration s. The advantage

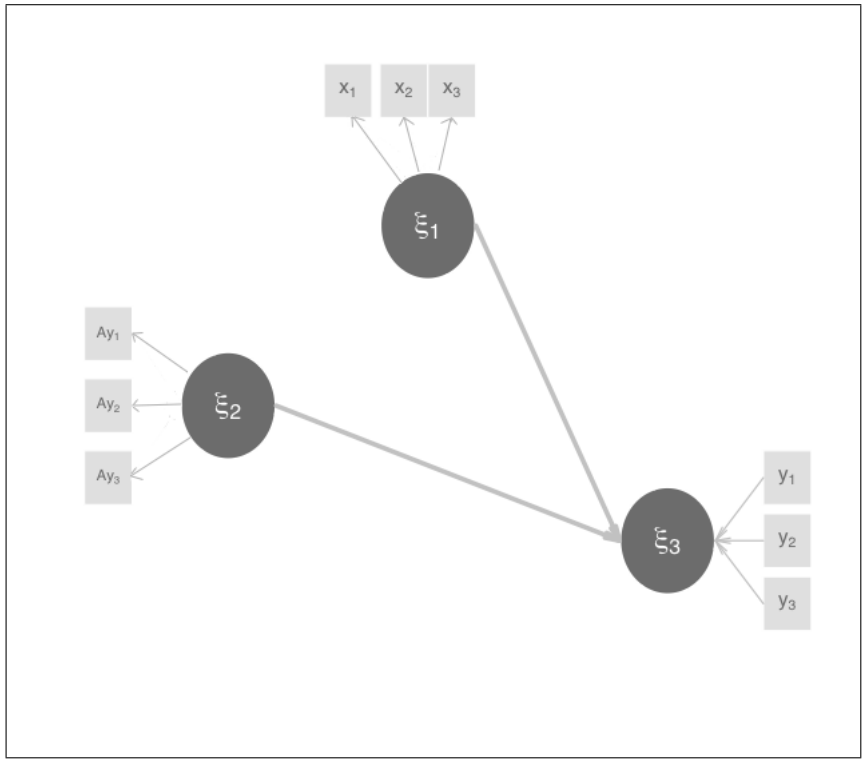

Figure 3.1: An example of a path diagram of PLS-PM with network data

of Lohmöller's procedure is that it can be calculated by means of matrix algebra and it is easier to implement.
By contrast, **Wold's procedure** relies always on the information available in the last iteration. As shown by Hanafi [81], the Wold's procedure seems to be more interesting for its monotony properties.
According to Kramer [102] and Tenenhaus and Tenenhaus [150] in this thesis we use the PLS path modelling algorithm, following rules hold:

1. Wold's procedure is used;
2. New Mode A is applied to all measurement models;
3. The centroid scheme is used as inner weighting scheme.

As described in the chapter 2, the PLS algorithm includes the following three stages:

1. iterative approximation of latent variable scores;
2. estimation of latent variable scores;
3. estimation of path coefficients.

The core of PLS algorithm is the **first stage**, consisting of four steps [149] that we describe considering the introduction of adjacency matrix in the model.

3.4.1 The first stage: Iterative process

Step 0: Initial arbitrary outer weights

We start the iterative process by assigning any arbitrary non-trivial linear combination of indicators can serve as an outer proxy of a latent variable [86].

Step 1 - Outer approximation of the latent variable scores

Outer proxies of the latent variables are estimated as a linear combination of their own manifest variables:

$$\nu_q \propto \pm C_q w_q \tag{3.4}$$

where:

- ν_q is the standardized outer estimate of the q - th latent variable ξ_q;
- C_q is the matrix that contains the manifest variables about actors' attributes and outcomes (i.e. our network effects dependent variables);
- the "$\pm$" sign shows the sign ambiguity that is solved by choosing the sign making ν_q positively correlated to a majority of manifest variables.

Step 2 - Estimation of the inner weights

Inner weights are calculated for each latent variable in order to reflect how strongly the other latent variables are connected to it, considering the existing links with the other Q' adjacent latent variables:

$$z_q \propto \sum_{q'=1}^{Q} d_{qq'} e_{qq'} \nu_{q'} \tag{3.5}$$

Let be:

- $d_{qq'}$ is the generic element of the indicator matrix D of order Q, where $d_{qq'} = 1$ if LV ξ_q is connected to $\xi_{q'}$ in the path diagram and $d_{qq'} = 0$ otherwise;

- the inner weights $e_{qq'}$ are determined through the Wold original scheme, i.e. the centroid scheme, where $e_{qq'}$ is equal to the sign of the correlation between v_q and $v_{q'}$.

Step 3 - Inner approximation of the latent variable scores

Inner proxies of the latent variables are calculated as linear combinations of the outer proxies of their respective adjacent latent variables, using the inner weights previously determined.

Step 4 - Estimation of the outer weights

The calculation of outer weights depends on the type of relation existing between a block of MVs and the underlying LV.
For the estimation of the outer weights w_{pq} we use a specific mode of the reflective model, i.e. the *New mode A*:

$$w_q = \frac{C_q' z_q}{\left\| C_q' z_q \right\|} \tag{3.6}$$

where w_q is a vector of outer weights $w_q = (w_{1q}, \ldots, w_{pq})$

3.4.2 The second stage: Computation of the latent variable scores

Upon convergence, the estimates of the latent variable scores are obtained as:

$$\hat{\xi}_q \propto C_q w_q \tag{3.7}$$

where the Q-th score is associated to the latent variable linking our network effects dependent variables.

3.4.3 The third stage: Computation of path coefficients

In the last stage of the PLS-PM algorithm path coefficients are estimated through OLS multiple regressions among the scores of estimated latent variable:

$$\hat{\beta}_j = (\hat{\Xi}'_{\to j}\ \hat{\Xi}_{\to j})^{-1}\ \hat{\Xi}'_{\to j}\ \hat{\xi}_j \tag{3.8}$$

where ξ_j is the generic endogenous LV score vector and $\hat{\Xi}_{\to j}$ is the matrix of the corresponding latent variable scores, containing the information derived from attribute data and network data.

3.5 A Simulation Study to compare NEM and PLS-PM coefficients

The specification of the Network Effects Structural Model will be now analyzed in order to highlight some properties of the derived estimates and then, to compare results obtained within the specification of the Network Effects Model.

The idea is to set-up a simulation scheme where population parameters are predefined and used in determining a *population* variance - covariance matrix.
The equations, within this matrix, show the variance and covariance of the observed variables as functions of the model parameters. They are sufficiently general in order to capture most SEMs with continuous variables [8].
Then, resampling from the population a large number of datasets coherent with the specified covariance structure but for some random disturbances, the two methods (PLS-PM and NEM) will be carried out and corresponding estimates will be compared.
In this way, we are able to analyze some statistical properties of the estimators in terms of bias and consistency.
The simulated data should be consistent both for the network effects and for the structural equation model.
At this aim we have implemented an ad-hoc procedure by using the `R` software and related packages.
Namely, we use well known theoretic results as the model-implied variance-covariance matrix definition for the SEM structure and the QR decomposition for generating the network effects outcomes and attribute data variables as defined in NEM.

3.5.1 The simulation scheme

In the simulation scheme, we consider the i) network density and ii) the sample size (n) as factors that can influence the network structure.
According to a simulation study of Mizruchi and Neuman [122], the esti-

mate of network effects autocorrelation parameter ρ, i.e. estimated $\hat{\rho}$ tends to be lower than the population ρ.
This tendency becomes more pronounced with higher density in random graphs also in well-known structures.
One argument, consistent with the work of Festinger et al. [57], suggests that the extent to which actors are affected by that of their peers will be especially pronounced in highly cohesive groups.
Furthermore, we control for the network autocorrelation coefficient ρ and for the standard deviations of disturbances (σ_ϵ).
These 4 factors take two level each, giving raise to 2^4 possible run schemes (see tab. 3.1).

Also, we need to fix (in step 0) all population parameters to be used in the specification of the structural model.
Indeed, we need a data structure made of exogenous and endogenous variables that produce corresponding latent variables coherent with SEM and with NEM.

3.5.2 The simulation procedure

The simulation procedure is organized as follows:

Step 0 - Initialization phase

Let be:

- P the number of exogenous variables $\mathbf{X}_p$;
- Q the number of endogenous variable $\mathbf{Y}_q$;

Run	$\rho^{(*)}$	Net Density	Net Size	$\sigma_\epsilon^{(*)}$
1	0.25	0.2	300	0.1
2	0.25	0.2	300	0.5
3	0.25	0.2	100	0.1
4	0.25	0.2	100	0.5
5	0.25	0.5	300	0.1
6	0.25	0.5	300	0.5
7	0.25	0.5	100	0.1
8	0.25	0.5	100	0.5
9	0.75	0.2	300	0.1
10	0.75	0.2	300	0.5
11	0.75	0.2	100	0.1
12	0.75	0.2	100	0.5
13	0.75	0.5	300	0.1
14	0.75	0.5	300	0.5
15	0.75	0.5	100	0.1
16	0.75	0.5	100	0.5

Table 3.1: The 16 Factor-levels combinations used in simulations. (*) The actual values of ρ and σ_ϵ will be slightly modified by the numerical procedure.

- Λ_x the loadings associated to manifest exogenous variables, by choosing and fixing the population parameters of a SEM;
- Λ_y the loadings associated to manifest endogenous variables, by choosing and fixing the population parameters of a SEM;
- Φ the variance-covariance matrix of the latent variables ;
- B path-coefficients of the endogenous variables;
- Γ path-coefficients among the endogenous and exogenous variables.

These population quantities are fixed and used to build the model-implied variance-covariance $\Sigma(\Omega)$ matrix according to the usual SEM definition.
It is used to write variance and covariance terms of manifest variables as a function of SEM coefficients (for more details see Bollen, 1989 [7]).

Let Σ be the matrix of variance - covariance of the population

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{3.9}$$

where:

- Σ_{xx} is the variance covariance matrix of manifest variables X;
- Σ_{yx} is the intercovariance matrix between the endogenous and exogenous manifest variables;
- Σ_{yy} is the variance covariance matrix of manifest variables Y.

The three matrices within matrix Σ can be rewritten in function of the parameters of the model. So the matrix of the exogenous manifest variables, in terms of the parameters of the model, is:

$$\Sigma_{xx}(\Omega) = E(xx') = E[(\Lambda_x\xi+\delta)(\Lambda_x\xi+\delta)'] = \Lambda_x E(\xi\xi')\Lambda_x'+\Lambda_x E(\delta\xi')\Lambda_x'+E(\delta\delta') \tag{3.10}$$

with the assumptions that $E(\xi\xi') = \Phi$ and $E(\delta\delta') = \Theta_\delta$, the new equation is:

$$\Sigma_{xx}(\Omega) = \Lambda_x\Phi\Lambda_x' + \Theta_\delta \tag{3.11}$$

The matrix of the endogenous manifest variables is:

$$\Sigma_{yy}(\Omega) = E(yy') = E[(\Lambda_y\eta+\epsilon)(\Lambda_y\eta+\epsilon)'] = \Lambda_y E(\eta\eta')\Lambda_y'+\Lambda_y E(\epsilon\eta')\Lambda_y'+E(\epsilon\epsilon') \tag{3.12}$$

with the assumption that $E(\epsilon\epsilon') = \Theta_\epsilon$ the new equation is:

$$\Sigma_{yy}(\Omega) = \Lambda_y E(\epsilon\epsilon')\Lambda_y' + \Theta_\epsilon \tag{3.13}$$

By considering the structural model, the endogenous variables can be expressed also as:

$$\eta = (I - B)^{-1}(\Gamma\xi + \zeta) \tag{3.14}$$

with the assumption that $E(\zeta\zeta') = \Psi$ the matrix Σ_{yy} becomes:

$$\Sigma_{yy}(\Omega) = \Lambda_y[(I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1\prime}]\Lambda_y' + \Theta_\epsilon \tag{3.15}$$

The matrix of inter-covariance can be written as:

$$\Sigma_{yx}(\Omega) = E(XY') = E[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \epsilon)'] = \tag{3.16}$$

$$= \Lambda_x E(\xi\eta')\Lambda_y' + \Lambda_x E(\xi\epsilon') + E(\delta\eta') + E(\delta\epsilon') \tag{3.17}$$

with the assumption that the errors are uncorrelated, the new equation is:

$$\Sigma_{yx} = \Sigma_{yx}' = \Lambda_y(I - B)^{-1}\Gamma\Phi'\Lambda_x' \tag{3.18}$$

By substituting these three new equation in the Σ we have the implied variance covariance matrix $C = \Sigma(\Omega)$ described as:

$$\Sigma(\Omega) = \begin{bmatrix} \Lambda_x\Phi\Lambda_x' + \Theta_\delta & \\ \Lambda_y(I - B)^{-1}\Gamma\Phi'\Lambda_x' & \Lambda_y[(I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1\prime}]\Lambda_y' + \Theta_\epsilon \end{bmatrix} \tag{3.19}$$

Step 1 - Data structure

In the first step of our simulation process, we generate one adjacency data matrix **A**, with given density and number of nodes (according to the simulation scheme).
We also derive X and Y as multinormal random variates with covariance structure assigned according to the population Σ by using the `mvrnorm` routine in `R`.
These data-sets are used to compute the population network coefficients,

i.e. ρ and the coefficients β in terms of the elements in Σ, so they are coherent both with the true SEM data structure and with the true NEM specification.

Step 2 - Resampling from population

In the second step we fine-tune the data to be best suited for a Network Effects Model (according to equation 3.1) and SEM specification (equation 3.3).
At this aim, we generate a random vector of disturbance ϵ with standard deviation specified according to the simulation scheme.
Then we use a solving strategy to find the solution of the inverse regression model as implemented in the `qr.solve` function in `R`, having fixed the NEM coefficients.
At the end of this second step, we have for each of the 16 schemes the following constants and variables.

Constants to be held in all iterations:

- Population parameters for SEM and NEM.
- Adjacency data $\mathbf{A}$.

Variables depending from random disturbance in each iteration:

- The $\mathbf{Y}$ outcome as a function of the NEM coefficients (β and ρ) but for a random disturbance with given variance.
- The $\mathbf{Y}$- network effects outcome, that is $\mathbf{A}Y$.

- The **X** attribute data (or indicators in the SEM terminology) to define the latent construct in SEM and the explicative variable in the NEM.

About PLS-PM algorithm, in this simulation study, we use:

1. **Wold's procedure**, summarized step by step below, for its monotony properties (major details see the precedent chapter);
2. **New Mode A** applied to all measurement models;
3. **The centroid scheme** as inner weighting scheme.

Reiterating the procedure for $S = 500$ times, having fixed the values according to the 16 factor-level combinations, we obtain random fluctuations in the data structures and can estimate the coefficients' sampling distribution by carrying out both NEM and PLS-PM on the data.

3.5.3 Results

For sake of comparison with a typical NEM, we set $Q = 1$ (one outcome), then we consider $P = 3$ (attribute data on the nodes).
The results are illustrated in the following.
In Figure 3.2 we compare the distributions of the 500 random replications of estimated coefficients obtained for the NEM and for the PLS-PM under the 16 controlling conditions of the simulation scheme.
The obtained results are very rich and allow to analyze the different effect, the factor may have on bias, efficiency and consistency of the two kinds of estimators.
In this work, we concentrate only on the coefficients that are common to the two methods, that is the exogenous coefficients $\beta_p (p = 1, \ldots, 3)$ and the ρ correlation coefficients.

The distribution are showed as paired box plots, for each of the 16 runs. Since, we find some degenerate solutions in the data generation process, we need to trimmer the empirical distributions and decided to show only non-trivial results.
This could lead to drop out some true outlier for both the two methods, but we expect they should randomly appears in both methods.
In the PLS-PM, there are some datasets with problems, i.e. they do not converge or converge to improper solutions. These type of datsets are called "imperfect".
When the object of interest is not represented by non-converged samples, they can be eliminated by the analysis, because they provide irrelevant information and threaten external validity [132].
In our simulations, in average, less than 3% of the runs in each diagram (about 10 - 15 samples) give outliers. In this case we have decided to eliminate them.
In a further study, we plan to check for such degeneracies.
In Figure 3.2, for all three coefficients, we show two-paired box-plots of the attribute coefficients' empirical distributions of the NEM (dark line) and of the PLS-PM (gray line) in the 16 schemes.
The paired labels aid to recover the 16 conditions of our experiment, according to Table 3.1.
Then, we show analog results for the ρ coefficients estimated with the two methods, where dark are NEM coefficients and gray are PLS-PM coefficients. In Figure 3.3 we may appreciate how systematic patterns appear when changing the factor-level combination under control.
Changes in the disturbances' variance seem to affect the relative efficiency of the NEM coefficients (runs 2, 4, 10, 12 in both Figure 3.2 and Figure 3.3),

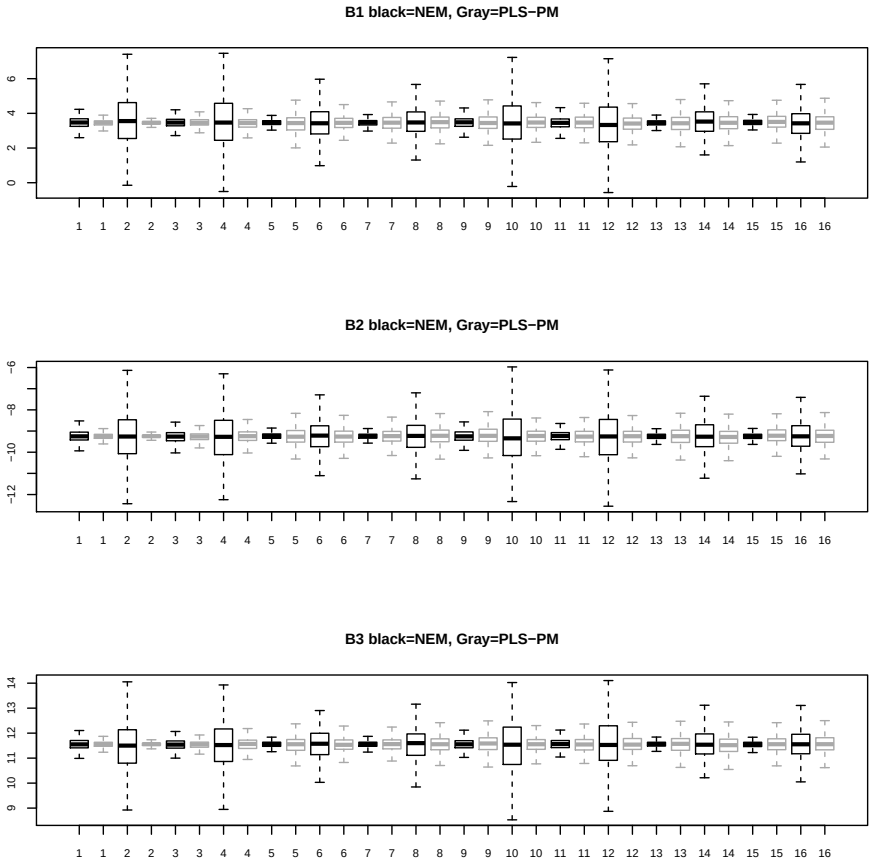

Figure 3.2: Boxplots of the attribute coefficients' empirical distributions in the 16 simulation schemes: dark are NEM coefficients, gray are PLS-PM coefficients.

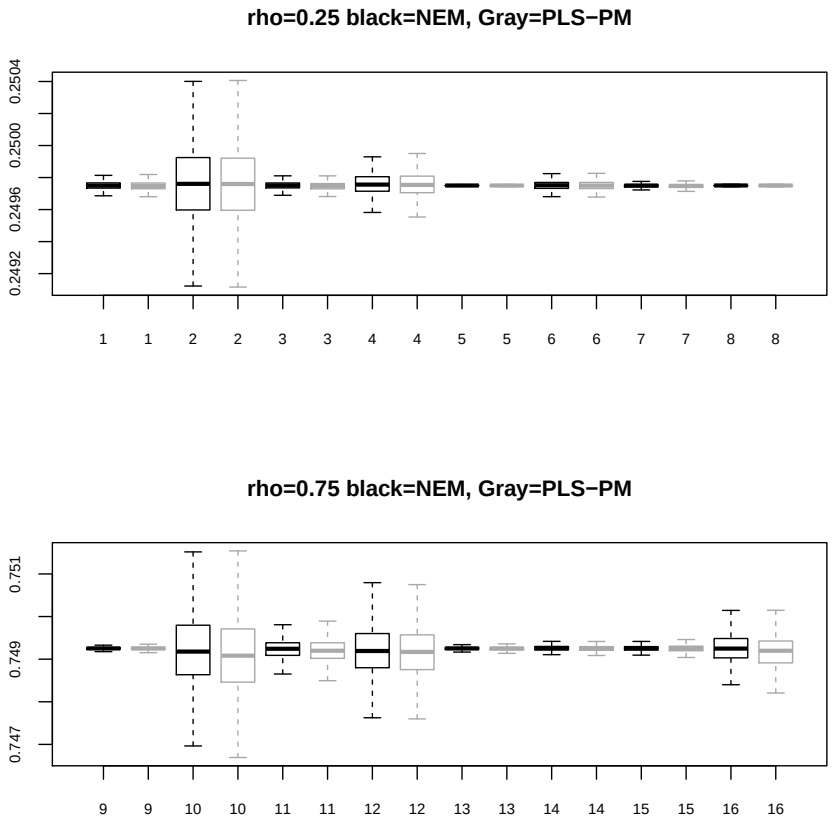

Figure 3.3: Boxplots of the ρ coefficients' empirical distributions in the 16 simulation schemes: dark are NEM coefficients, gray are PLS-PM coefficients.

while a proper significative changes when changing simulation conditions seem affect also network size.
Deepen analysis is provided to give reliable results, even if it appears that NEM coefficients' distributions show larger fluctuations around the expected values. We use the pooled mean squared errors of the β's empirical distributions as a response variable in an Anova model. The effects of the controlling factors on the simulation results are analyzed.

Table 3.2: Simulation Effects

	Net Size	Net Density	ρ	σ_ϵ
NEM	-0.27615	-0.00468	-0.01215	0.50416
	(0.257303)	(0.257303)	(0.257303)	(0.257303)
	0.3061	0.9858	0.9632	**0.0759**
PLS-PM	0.54803	-0.38818	0.89048	-0.09149
	(0.37899)	(0.37899)	(0.37899)	(0.37899)
	0.1760	0.3277	**0.0385**	0.08137

In Table 3.2 there are the simulation effects and their significance (p-value).
We may observe that the bias actually increases when variance of disturbances increases in the NEM model (although p-value is 0.0759).
The ρ coefficient appears to show significative effect only for the PLS-PM method; indeed, bias of estimates increases when the network autocorrelation coefficient is larger (p-value is 0.0385).
The size of the network in the considered range (100-300 nodes) and the network density do not seem to affect the mean squared error of estimates.

Chapter 4

An innovative contribution in the sociological field: Modelling Social Influence through Component based Models

When actors adapt their behaviour, attitude, or belief, to the behaviours, attitudes, or beliefs of other actors with whom they are in contact, this process can be defined as **social influence** or **contagion** ([106]; [107]; [108]). In the last years, different studies have analysed social behaviour and institutions by reference to relations among such concrete social entities as people, organizations and nations.

Relational analysis contrasts on the one hand with reductionist methodological individualism and on the other hand, with macro-level determinism, whether based on technology, material conditions, economic conflict,

adaptive evolution or functional imperatives.
In this more intellectually flexible structural middle ground, analysts situate actors and their relations in a variety of contexts.
In terms of quantitative analysis, conventional data sets are usually viewed in terms of units of analysis and variables that reflect a **variable-centered approach** [1]. This can be characterized as a merely statistics approach.
Based on the prospect of revolutionizing social science via networks, many special social network tools have been created to analyse network data [45]. This is considered a **network-only approach**.
Recently, statistical tools have been used to analyse network data for network concerns (e.g. Wasserman and Pattison, [156]) and some of the network tools have been used on conventional attribute data.
The study of social influence is a crucial field because it permits to work with both network and attribute data.

4.1 Theoretical background and hypotheses

Some authors (Simon[138]; Cartwright[20]) state that influence is a special instance of causality, that is, the modification of one person's responses by the actions of another.
The processes that underlie influence are different and include relations of authority, identification, expertise and competition.
Through the network approach to social influence one needs:

- elucidating the **substantive processes** that underlie that there should be structural effects in the attitudes and behaviours of actors;
- defining **interpersonal proximity** in a network, in an appropriate manner given these processes;

- assessing the **predictive success of the approach** using available mathematical and statistical models of social influence processes.

4.2 Social influence from a sociological point of view

In line with an early position of Durkheim [51], sociological theories of action tended to focus on the exterior constraining influences of individuals' social environments, including their stable definitions of situations and the antecedents of such definitions.
Hence, there was a heavy concentration of interest on social facts as race, gender, family background, population ecology and demography. The analysts so described the social differentiation of groups, communities and organizations through nominal classifications of actors (based on gender, race, ethnicity, religion or occupation among other variables).
Through the idea that networks of social relations define social differentiation, the positions of actors are revealed by their patterns of relations with other actors, and a differentiated social structure is defined by the existence of actors who occupy different positions in network of social relations.
According to Mead [120], the social control depends on the degree to which individuals in society are able to assume attitudes of others who are involved with them in common endeavors.
An element that creates a bridge between these two approaches is the **interpersonal influence**.
Interpersonal influence is a foundation of actors' effects to control their social environment by modifying the attitudes and opinions of the others with whom they interact. It also has an effect on the actors' attitudes and

opinions and is, therefore, a foundation of actors' socialization, identity and decisions.
In networks like groups, communities and organizations, this influence process can produce agreements that define the culture of the group and that frame the collective activities of its members.
In this vein, social network analysis provides two broad approaches to indicating environments in which the influence between particular pairs of actors is more or less likely.

4.2.1 Two Perspectives on Social Influence

In literature (Friedkin [67], [68], [70]; Leenders [106], [107], [108]) there are different theories of social influence and in order to find ways to determine the effects of contextual structures, several solutions have been tried.
According to these theories, some attitudes and opinions of significant others influence the way in which a person comes to view a situation. The opinions of alters are seen as an appropriate standard against which ego evaluates his own opinion.
When forming his own opinion, ego uses other actors as his frame of reference and takes their opinions into account becoming more similar to others.
Within the social influence theory, according to Leenders [108], the notion of a frame of reference has crystallized around two processes:

1. Communication, i.e. actors use actors with whom they are directly tied to as their frame of reference.

2. Comparison, i.e. actors use actors they feel similar to as their frame of reference.

The most frequent and likely type of influence is social influence through direct contact between ego and alter, i.e. through **communication**.
In this process, through discussing matters with alter, ego comes to an understanding of an issue, because he accepts information from alter and adds new information to his own.
The other process of contagion is **social comparison**.
Ego, in order to search for his social identity, compares himself to those alters whom he considers similar to him in relevant respects, perceiving or assessing alter's behaviour. In this case, ego changes his behaviour, because ego wants to be like alter. This influence process resembles the mimetic isomorphism, discussed by DiMaggio and Powell [38].
These two types of social influences are empirically hard to distinguish [106].
Since individuals have multiple relations in their networks and are potentially influenced by all of them, it is necessary to extend the definition. Following Friedkin's [67] definition of social influence, social influence occurs when ego assimilates his attribute to become similar to all the friends in his network. As a result having more relations increases the level to which someone is influenced [33].
This approach, which is predominantly interpersonally oriented, ignores the vast amount of research which focuses on intergroup processes ([89]; [33]).
Social influence research should take into consideration both perspectives, since most natural settings are a mix of these two extremes ([14]; [148]; [145]; [36]; [33]).
According to Deaux and Martin [36], because social groups and the process of social identification are originated from the interpersonal network, the interpersonal influence will be more important.

4.2.2 Social influence network theory

The network of interpersonal influence, however, has a special theoretical status: if it is acknowledged as a noteworthy determinant of the actions under examination, it must be made part of a causal model [72].
Social influence network theory (for more details see Friedkin and Johnsen works) represented, in the late 1950s, a mathematical formalization of the social process of attitude change when it unfolds in a social network of interpersonal influences.
This theory advanced the hypothesis that networks of interpersonal influence were more important into the formation of interpersonal agreements and group consensus.
In 1956, Cartwright and Harary [21] with the theory of structural balance, create a link among social cognitions and social networks, and French, with the theory of social power create a link among social networks and group members' positions on issues [75].
These initial formulations describe the formation of group consensus and they do not provide an adequate account of settled patterns in case of disagreements. This limit was overcome in 1990 through Friedkin and Johnsen's generalization (Friedkin [68], [69]; Friedkin and Johnsen [72], [73], [74]).
The goal of a formal network theory of social influence is to try what is appropriate or correct under specific circumstances through a process of interpersonal influence, reducing uncertainty and conflict by the development of a shared attitude.
There is not a formal specification of mechanism that shows how interactions among group members operate to transform people' s uncertainty and

conflict into the interpersonal agreement that appears to be fundamental to the development of a norm.
For this reason, one assumes that a norm is a special case of an attitude for which the positive or negative evaluation of a feeling, thought or action becomes a shared normative evaluation, creating a link between the norm formation process and the theory of the formation of attitudes and the development of consensus.
The formal theory involves a two-stage weighted averaging of influential opinions. Actors start out with their own initial opinions on some matter. At each stage, then, actors form a "norm" opinion, which is a weighted average of the other opinions in the group. Actors then modify their own opinion in response to this norm, forming a new opinion, which is a weighted average of their initial opinion and the network norm.
This theory uses mathematical models and quantifications to measure the process of social influence.

4.3 The network effects model in social influence

Network theorists discovered that the network effects model could be modeled with the autocorrelation stemming from social proximity [158].
In such cases, $\mathbf{A}$ becomes a matrix of social distances and ρ is a substantive parameter estimating the extent to which an actor's outcome is affected by the behavior of those to whom he is socially proximate [42].
The network autocorrelation model allows the investigator to model an outcome variable for a single actor as a simultaneous consequence of both network and individual-level variables. This is a considerably clearer and more straightforward approach.

Two equations can describe the social influence network theory.
One of these concerns the origins of actors' initial opinions on an issue:

$$y = X\beta + \epsilon \tag{4.1}$$

$$E[\epsilon] = 0, E[\epsilon'\epsilon] = \sigma^2 I \tag{4.2}$$

The other equation concerns the transformation of these initial opinions through interactions among people.
We refer to it as the Network Effects Model, specifying it as follows:

$$y = \rho Ay + X\beta + \epsilon \tag{4.3}$$

As discussed by Leenders [108] and Marsden and Friedkin [116], the network effects term $\mathbf{A}y$ can be interpreted as a form of social influence, and thus provides a clear bridge between statistical analysis and social theories where comparison and reference processes are important.
This model is appealing because it integrates covariate or attribute effects of variables in X on the outcome y with network or interdependence effects of $\mathbf{A}y$.

4.4 Substantive interpretation of the Network Effects Structural Model

In order to give a substantive reading of the proposed approach in a specific application field, we consider the phenomenon of social influence that is a crucial issue for social network research, because it links the structure of

social relations to attitudes and behaviors of the actors in a network. In a statistical context, the specification of:

$$y = \rho Ay + X\beta + \epsilon \tag{4.4}$$

represents the construction of opinion of an actor, considering both his intrinsic opinion, in absence of social influence [106], and the opinion of his alters.

In the specific, the intrinsic opinion of the actor (ego) is represented by $X\beta$, when ρ is equal to zero, but it also verifies that, in determining his opinion, ego takes into account the opinions of his significant alters representing his frame of reference.

The significance of ego's alters is delineated by nearness, i.e. how the alters' opinions and beliefs are emulated by ego, represented by $\mathbf{A}y$.

If we denote with the subscript i the elements of ego, while the elements of ego's alters with *i'*, then, y_i is related to a weighted combination of the $y_{i'}$, where the weights are given by the $n \times n$ matrix $\mathbf{A}$, then y is related to $\mathbf{A}y$.

For instance, ego's political preference [108] can be represented in our approach by the endogenous Latent Variable ξ_j.

It could be a function of two elements:

i) ego's socio-economic status, education, and income, i.e. attribute data represented by manifest variables of the exogenous LV $\xi_{1,\ldots,Q-1}$;

ii) the political views expressed by ego's family, neighbours, and colleagues, represented by the exogenous LV ξ_Q, i.e. the latent variable associated to the network effects dependent variables.

Together, these effects then simultaneously determine ego's political stance. In this case the **Network Effects Structural Model** (eq. 3.3): can be

interpreted as follows:

- ξ_j is ego's political preference;
- $\beta_{q\in(1,\ldots,Q-1)}$ are the path-coefficients linking the exogenous latent variables associated to attribute data (e.g. ego's socio-economic status, education, and income) to ego's political preference;
- ρ_Q is the path coefficient linking the exogenous latent variable associated to network effects dependent variable, e.g. the political views expressed by ego's family, neighbours, and colleagues, to ego's political preference; ρ_Q measures the magnitude of the network effects;
- ζ_j is the part of ego's political preference that is not explained by the model.

It is possible to represent the process of social influence through a path diagram in a Network Effects Structural Model, as described in the Figure 4.1.

4.5 Further developments

In this work we have been interested in understanding how the social relationships influence people's choices, opinions and behaviors.
The substantive argument is that individuals modify their actions in response to other individuals' actions, therefore the networks of interpersonal influence that form these responses are a potentially noteworthy part of this work.
We have examined the nature and role of the social influence exerted by the network on its members ([2]; [133]; [37]), combining the elements of the

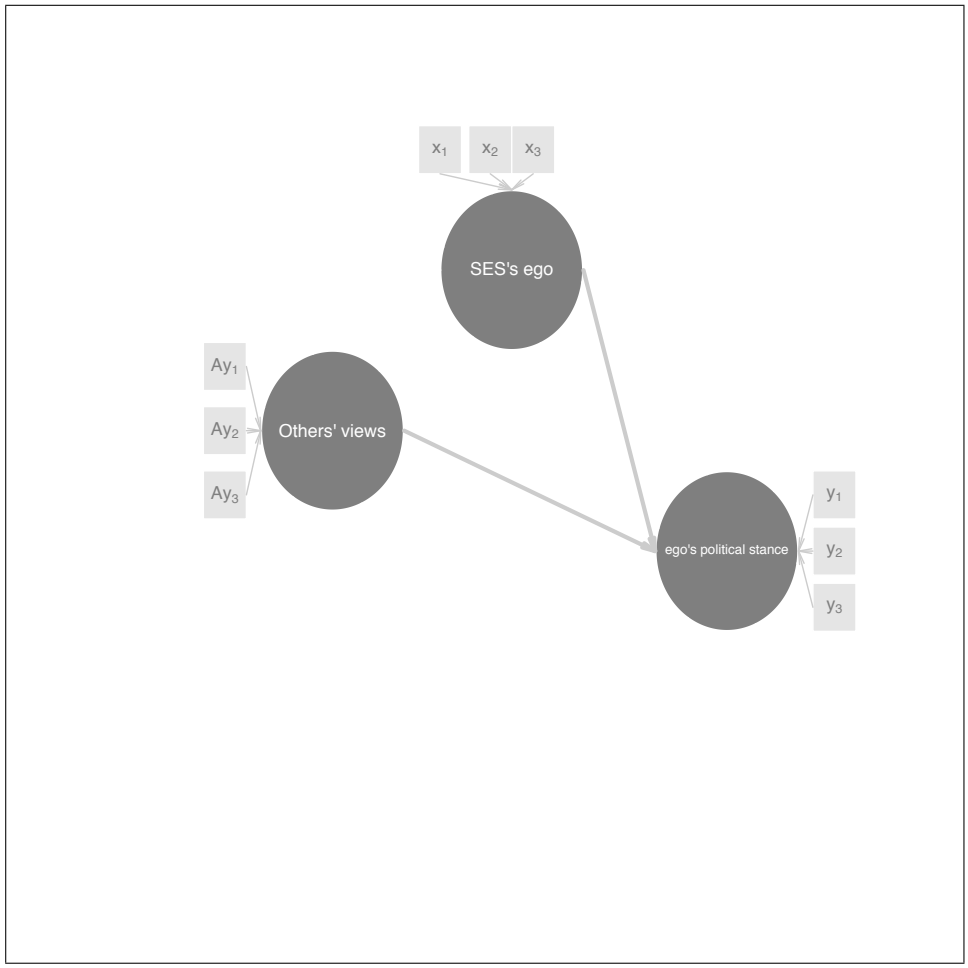

Figure 4.1: Social influence in the PLS-PM model

social network analysis and those of the PLS Path hypotheses.
The methodological argument is that controlling the social network effects is essential for understanding of individual actions, because an important kind of social structure emerges from the individuals' responses to other individuals' action.
In order to frame our discussion we use the study of Bagozzi and Dholakia [6] to start with.
Bagozzi and Dholakia [6] have modeled people's intentions as a function of individual-level and group-level variables that act separately to influence the people's attitudes.
There are two hypotheses whic are fundamental in socially targeted marketing on why people interact with others, they are:

- People want to get and share information in the network in order to know what others think or to validate on a decision already made or to buy a product (e.g., [84]; [118]);
- People want to understand and deepen salient aspects of one's self in order to obtain access to social resources and facilitate the attainment of one's future goals [118], i.e they may help one to form, clearly define and elaborate on one's own preference, taste and value.

4.5.1 Multiple networks

Analyzing the behavior of these individuals, such as their purchasing or technology adoption tendencies, requires statistical techniques that can handle both the scope and the complexity of the data.
One aspect of this complexity is the social network.

A popular approach to study this phenomenon is to use a model with explicit autocorrelation between individual outcomes, i.e. Network Autocorrelation Models, described above.
But this approach is defined with a single network structure term, ignoring the possibility that an individual may belong to two or more groups.
The advent of the Information Age has opened new possibilities in the field of social network analysis by making very large repositories of data available to researchers [77].
The depth of data now available, through phone calls, electronic communication via email, social networking services, blog-providers, etc provides researchers with a rich and publicly observable data to use in the analysis of social interactions.
An actor can be a member of multiple distinct but overlapping networks, such as a friend network, a work colleague network, a family network, and so forth, and each of these networks may have some connection to the outcome of interest, so a model that condenses all networks into one relation will be insufficient [165].

4.5.2 Homophily

It is necessary to consider the possibility that the network autocorrelation, due to some direct influence of an individual's neighbours on his behavior, can be opposed to the effect of homophily, in which social ties form among individuals with similar antecedent characteristics, who may then behave similarly as a result.
A pervasive feature of social and economic networks is that contacts tend to be more frequent among similar agents than among dissimilar ones.
The presence of homophily has important implications on how agents' char-

acteristics - genders, races, ethnicities, ages, class backgrounds, educational attainment, etc. - impinge on the information they receive, the attitudes they form, and the interactions they experience.
The result is that people's personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. Homophily therefore implies that distance in terms of social characteristics is translated into network distance [119].
It is therefore important to understand the generative process of homophilous social networks, and how the agents' preferences and their meeting opportunities concur in determining the observed mix of social ties [29]. Through the study of homophily it is possible to note how the network's surrounding contexts can drive the formation of its links.
One looks beyond the network to understand where the link comes from, in the specific, one looks at some social environments, e.g. schools and companies, to which the nodes belong. Therefore, in the same network it is necessary to consider both intrinsic effects and contextual effects on the formation of any single link.
If we want to represent this phenomenon in our model, through a path diagram, we can represent it as in the Figure 4.2.

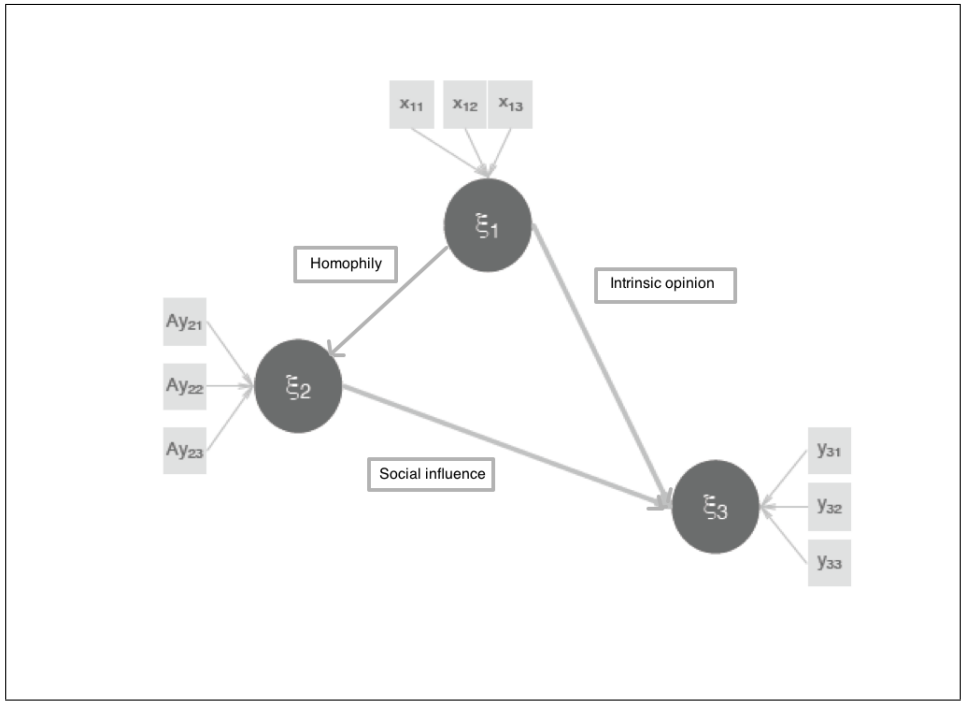

Figure 4.2: Homophily in the PLS-PM model

Conclusions

This thesis stems from the idea to draw a statistical soft-modeling framework to network data.
Network data arise in very different and multidisciplinary fields in order to study relational ties among units. The different fields highlighted in recent years the necessity to collect relational and attribute data, as well as metadata describing the actors in the network.
Since usual relational datasets are characterized by i) very different amount of units (from very few units to huge networks), ii) biased sampling (for instance, people with more social connections may have a higher chance of selection) and iii) a kind of heterogeneous information attached to both nodes and ties; these facets highlight the peculiarity for classical statistical tools and models to be applied.
In the specific, we are interested in processes where social relations provide a basis for the alteration of an attitude or behavior by one actor in response to another one.
This social process of attitude change, that appears in a social network, is known as social influence or contagion.
A mathematical formalization of the effects of social network on behaviors is given by the Network Effects Model.

From an empirical point of view, these models are far from being directly observable.
The possibility of measuring them as latent factors depending from multi-dimensional constructs still remains.
All together, a component-based approach to network data through Partial Least Squares-path model algorithms is proposed.
A simulation study is presented in order to compare the Network Effects Model with the proposed approach by examining coefficients estimated from the two methods while controlling for specific attributes: network size, network density, autocorrelation coefficients and standard deviation of disturbances.
Numerical simulation seems to show significant results with more robust results of the proposed approach with respect to traditional Network Effect Model.
Anyway, further insight is due to assess some degeneracies that appeared in the Monte Carlo simulations in order to establish conclusive results.
The new approach proposed in this thesis can be considered a powerful tool because, through it, we can analyse the research hypotheses from several points of view, by means of different methodologies.
From the Social Network Analysis point of view, we can analyse Network Effect Models ([5]; [43]) through PLS approach.
From the Structural Equation Modeling point of view we can extend the Partial Least Squares - Path Modeling ([160]; [149]) to network data.
From the Sociological point of view, we can operationalize the social influence ([106]; [107]; [108]).
From the Statistical point of view, we can assess reliability of estimated coefficients, i.e. weights of PLS-path model.

This is the beginning of a new approach and a new vision within Partial Least Squares methodology.

Appendix A

Routines in R Language

A.1 Simulation data

Function useful in order to simulate data

```
MySim.data_2step <- function(k,n,p,q,lambdax,lambday,phi,tetax,Beta, Gamma,
nettuno,tetay,densita,var.e) {

e <- list()
x <- list()
y.sem <- list()
y <- list()
y_lagged.sem <- list()
w1 <- list()


w1_unica <- rgraph(n,,densita)

num.VLend <- nrow(as.matrix(Beta))
num.VMeso <- p+q
sigmaX <- lambdax%*%phi%*%t(lambdax) + tetax
sigmaY <- (lambday%*%
           (
    (solve(diag(num.VLend)-as.matrix(Beta)))
    %*%
    ((Gamma%*%phi%*%t(Gamma))+nettuno)
```

```
    %*%
    (t(solve(diag(num.VLend)-as.matrix(Beta))))
   )
   %*% t(lambday)
  )
         + tetay

sigmaYX <- lambday%*% (solve(diag(num.VLend)-as.matrix(Beta))) %*%
Gamma %*% t(phi) %*% t(lambdax)

# prima parte di sigma --> colonne di sigmaX e sigmaYX

pezzo1.sigma <- rbind(sigmaX,sigmaYX)
# seconda parte di sigma --> colonne di sigmaY e sigmaXY
pezzo2.sigma <- rbind(t(sigmaYX),sigmaY)

sigma <- cbind(pezzo1.sigma,pezzo2.sigma)

nomi <-c(paste("x_",c(1:(num.VMeso -1)),sep=""),"y_lagged","y")
colnames(sigma)<- nomi
rownames(sigma) <- nomi

true_beta <- (solve(sigma[1:num.VMeso,1:num.VMeso])%*%
sigma[1:num.VMeso,(num.VMeso+1)])[1:(num.VMeso -1),]
true_rho <- (solve(sigma[1:num.VMeso,1:num.VMeso])%*%
sigma[1:num.VMeso,(num.VMeso+1)])[num.VMeso,]
for (kk in 1:k) {

# Generate p+q multinormal variables with above defined covariance structure

xy <- mvrnorm(n, mu=rep(0,num.VMeso+q), sigma, empirical=F)
colnames(xy) <- nomi
x[[kk]] <- xy[,1:(num.VMeso-q)]
y_lagged.sem[[kk]] <- as.matrix(xy[,(num.VMeso-q+1):num.VMeso])
colnames(y_lagged.sem[[kk]]) <- c("y.lagged")
y.sem[[kk]] <- as.matrix(xy[,(num.VMeso+1):(num.VMeso+q)])
colnames(y.sem[[kk]]) <- c("y.sem")

#Assemble y from its components:

e[[kk]]<- mvrnorm(n,0,(var.e)^2,empirical=T)
```

```
y[[kk]]<-qr.solve(diag(n)-true_rho*w1_unica,x[[kk]]%*%true_beta+e[[kk]])
colnames(y[[kk]]) <- c("y")
w1[[kk]] <- w1_unica

}
list(x=x,y=y,w1=w1,y_lagged.sem=y_lagged.sem,y.sem=y.sem,beta=true_beta,
rho=true_rho,sigma=sigma,densita=densita,var.e=var.e,e=e)
}
```

A.2 The PLS - PM Algorithm

The PLS-PM algorithm used in this work

```
myPLSPM_all_Wold<-function(X, p_blocchi, path, scaling=NA, outer.mode=NA, PLScomp=NA, inner.scheme=NA){



if (is.na(scaling)==T) {
scaling<-vector("list", length(p_blocchi))
for (i in 1:length(scaling)) {
scaling[[i]]<-c(rep("NUM",p_blocchi[i]))
}
}

#REFLECTIVE WAY

if (is.na(outer.mode)==T) {
outer.mode<-vector("list", length(p_blocchi))
for (i in 1:length(outer.mode)) {
outer.mode[[i]]<-c("RIF")
}
}

if (is.na(PLScomp)==T) {
PLScomp<-array(1, length(p_blocchi))
}
#CENTROID SCHEME
if (is.na(inner.scheme)==T) {
inner.scheme<-c("CEN")
}


X <- as.matrix(X)
path <- as.matrix(path)
link <- t(path)+path
N <- nrow(X)
P<- ncol(X)
blocchi<-list()
mean_X <-list()
    var_X <- list()
    correzione<-(sqrt(N/(N-1)))
    QQ <- list()
p_blocchi<-c(1,p_blocchi)
for (q in 1:(length(p_blocchi)-1)) {
blocchi[[q]]<-as.matrix(X[,(sum(p_blocchi[1:q])):(sum(p_blocchi[1:q])+p_blocchi[q+1]-1)])
QQ[[q]] <- blocchi[[q]]
```

```
}
p_blocchi<-p_blocchi[2:length(p_blocchi)]
nbloc<-length(p_blocchi)
w <- vector("list", nbloc)
z <- vector("list", nbloc)



    y <- vector("list", nbloc)
    e <- matrix(,nbloc,nbloc)
    converg<-numeric()
ncicli<-0
z_temp<-matrix(0,N,1)

#####################################################################
    #                       data pre-handling                       #
#####################################################################
for (q in 1:(nbloc)) {
for (p in 1:(p_blocchi[q])) {
if (scaling[[q]][p]=="NUM") {
        QQ[[q]][,p]<-scale(QQ[[q]][,p])*correzione
        }
        if (scaling[[q]][p]=="RAW") {
        QQ[[q]][,p]<-QQ[[q]][,p]
        }
        }
}

#####################################################################
    #                       initialization                          #
#####################################################################

for (q in 1:nbloc) {
w[[q]]<-svd(scale(blocchi[[q]]))$v[,1]
w[[q]]<-w[[q]]/sqrt(as.numeric(t(w[[q]])%*%w[[q]]))
y[[q]]<-QQ[[q]] %*% w[[q]]
#y[[q]]<-QQ[[1]][,1]

    }

   ####################################################################
    #                        iterative cycle                          #
#####################################################################

repeat {
ncicli<-ncicli+1
```

```
        y_old <- y[[nbloc]]

for (q in 1:nbloc) {

         ###################################################################
         # --- updating the weights ["e"] ---- ############################
         ###################################################################

        # --- updating the weights ["e"]: CENTROID scheme ---- #
        if (inner.scheme=="CEN") {
          z[[q]] <- z_temp
          for (k in 1:nbloc) {
          e[q,k]<-cor(y[[q]],y[[k]])

          if (e[q,k]>0) {e[q,k]<-1}
          else {e[q,k]<- -1}

          z[[q]]<-(z[[q]])+(link[q,k]*e[q,k]*y[[k]])
          }
         }
         # --- updating the weights ["e"]: FACTORIAL scheme ---- #
         if (inner.scheme=="FAC") {
          z[[q]] <- z_temp
          for (k in 1:nbloc) {
          e[q,k]<-cor(y[[q]],y[[k]])

          z[[q]]<-(z[[q]])+(link[q,k]*e[q,k]*y[[k]])
          }
          }

          # --- standardize inner estimates if PLScore mode ---- #

          if (outer.mode[q]!="PLScow") {
           z[[q]]<-scale(z[[q]])*correzione
         }


###########################################################
       # ---- updating the weights ["w"] ---- ###################
        ############################################################


        # --- updating the weights ["w"]: REFLECTIVE WAY ---- #

         if (outer.mode[q]=="RIF") {
     w[[q]]<-(1/N)*(t(QQ[[q]]) %*% z[[q]])
         }
         # --- updating the weights ["w"]:FORMATIVE WAY ---- #
```

```
     if (outer.mode[q]=="FOR") {
     w[[q]]<-solve(t(QQ[[q]]) %*% QQ[[q]]) %*% t(QQ[[q]]) %*% z[[q]]
         }

         # --- updating the weights ["w"]:PLScore WAY ---- #

         if (outer.mode[q]=="PLScore") {
         w[[q]]<-myPLSRdoubleQ(z[[q]],,QQ[[q]],,PLScomp[q])$B
         }

         # --- updating the weights ["w"]:PLScow WAY ---- #
         if (outer.mode[q]=="PLScow") {
         w[[q]]<-myPLSRdoubleQ(z[[q]],,QQ[[q]],,PLScomp[q])$B

         w[[q]]<-w[[q]]/sqrt(as.numeric(t(w[[q]])%*%w[[q]]))
         }

        # --- outer estimations ["y"] ---- #


         y[[q]] <- QQ[[q]] %*% w[[q]]

         if (outer.mode[q]!="PLScow") {
         y[[q]] <- scale(y[[q]])*correzione
    }


       }

      #print(e)

#num_converg <- sum((y_old-z[[nbloc]])^2)
#den_converg <- sum(y_old^2)
#num_converg <- sum((w_old-w[[nbloc]])^2)
#den_converg <- sum(w_old^2)
        #converg <- num_converg/den_converg
        converg <- sum((abs(y_old)-abs(y[[nbloc]]))^2)
        print("converg")
        print(converg)
        print("ncicli")
        print(ncicli)
        if (converg<0.0000001 | ncicli>101)  {break}
}


     ###############################################################
     #       computation of the LVs using the outer weigts w       #
```

```
 ###################################################################

       VL <- list()
       sqm_VL <- array(, nbloc)
       w_tilde <- list()
       abs_w_tilde<-list()
VLS <- list()
       somma_w_tilde<-array(,nbloc)
w_tilde_normal <- list()
       for (q in 1:nbloc)
       {
        VL[[q]] <- QQ[[q]] %*% w[[q]]

        sqm_VL[q] <- sd(VL[[q]])/sqrt(N/(N-1))
        #cat("stDev di VL[q]:  ")
        #print(sd(VL[[q]]))
      w_tilde[[q]] <- w[[q]]/as.numeric(sqm_VL[q])
      #cat("w_tilde: ")
        #print(w_tilde[[q]])
        VLS[[q]] <- QQ[[q]]  %*% w_tilde[[q]]
        abs_w_tilde[[q]] <- abs(w_tilde[[q]])
  somma_w_tilde[[q]] <- sum(abs_w_tilde[[q]])
  w_tilde_normal[[q]] <- w_tilde[[q]]/somma_w_tilde[[q]]
       }
               # ----- the LVs are standardized  ----- #

        ###############################################################################
        # computation of the correlation between each LV and the corresponding MVs #
        ###############################################################################

       CORR_VL <- list()
   COMM_vm <- list()
       COMM <- list()
       for (q in 1:nbloc) {
         #CORR_VL[[q]] <- (t(QQ[[q]])) %*% VLS[[q]]/N
         CORR_VL[[q]] <-cor(VLS[[q]],QQ[[q]])

       # ----- computation of the Communality and Redundancy indexes  ------ #

  COMM_vm[[q]] <- CORR_VL[[q]]^2
  COMM[[q]] <- sum(COMM_vm[[q]])/p_blocchi[[q]]
       }

       ###############################################################################
   #                           AVERAGE COMMUNALITY                                 #
       #                                                                             #
       #       (the average communality is obatined taking into account all the      #
       #                     comunality indexes, i.e. one per block)                 #
```

```
   ###############################################################################

 COMM_M <-0
 for (i in 1:nbloc) {
  if (p_blocchi[i]>1) {
  COMM_M<-COMM_M+(p_blocchi[[i]]*COMM[[i]])
  }
 }
 COMM_M<-COMM_M/sum(  p_blocchi[which(p_blocchi>1)] )

   ###################################################################
       # computation of the parameters of the inner model by regressing #
       #each endo on own predictors #
       ###################################################################
 n_eso<-0
repeat {
n_eso<-n_eso+1
if (length(grep(1, path[n_eso,])) > 0) {break}
#if (path[n_eso,1]==1) {break}
}
n_eso<-n_eso-1
n_endo<-nbloc-n_eso
#print(n_endo)
pred<-vector("list",n_endo)
inn_regr<-vector("list",n_endo)
R2<-array(,n_endo)
RED_blocco<-array(,n_endo)

RED_vm<-vector("list", n_endo)
for (i in 1:n_endo) {
pred[[i]]<-matrix(,N,sum(path[n_eso+i,]))
count<-0

for (j in 1:ncol(pred[[i]])) {

repeat {
count<-count+1
if (sum(path[n_eso+i,1:count])==j) {break}
}

pred[[i]][,j]<-VLS[[count]]
}
inn_regr[[i]]<-lm(VLS[[n_eso+i]]~pred[[i]])
R2[i]<-(var(VLS[[n_eso+i]])-(var(residuals(inn_regr[[i]])))/var(VLS[[n_eso+i]])
RED_blocco[i]<-R2[i]*COMM[[n_eso+i]]
RED_vm[[i]]<-R2[i]*COMM_vm[[n_eso+i]]
}
R2_M<-mean(R2)
```

```
GOF<-sqrt(R2_M*COMM_M)
list(QQ=QQ, w=w,pred=pred,ncicli=ncicli,VLS=VLS,VL=VL,
CORR_VL=CORR_VL, w_tilde=w_tilde,w_tilde_normal=w_tilde_normal,
COMM=COMM, COMM_M=COMM_M,COMM_vm=COMM_vm,
blocchi=blocchi,N=N, inn_regr=inn_regr, GOF=GOF,R2=R2,
R2_M=R2_M, RED_blocco=RED_blocco ,RED_vm=RED_vm)

}
```

A.3 Application of SNA, OLS and PLS-PM

Function useful to apply SNA, OLS and PLS-PM on simulated data

```
MyFit.simulation <- function(k,p,q=1,data,SNA=TRUE,OLS=TRUE,PLSnew=TRUE) {

  results<-matrix(,k,(5*p+13))
  weights.X<-matrix(,k,p)
  weights.Y<-matrix(,k,q)
  weights.lagged<-matrix(,k,q)
  colnames(results) <- c(paste("True_beta",c(1:p),sep=" "),"True_rho",
  paste("beta_LNAM",c(1:p),sep=" "), "rho_LNAM","R2_LNAM", "AIC_LNAM",
  paste("beta_OLS",c(1:p),sep=" "), "rho_OLS","R2_OLS",
  "beta_PLSnew", "rho_PLSnew","R2_PLSnew","GoFnew","True_Sigma",
  paste("cor_y_x",c(1:p),sep=" "),"cor_y_ylag","Density",
  paste("beta_PLSnew_X",c(1:p),sep=""))
  colnames(weights.X) <- paste("PLS_w_x",c(1:p),sep=" ")
  colnames(weights.Y) <- paste("PLS_w_y",c(1:q),sep=" ")
  colnames(weights.lagged) <- paste("PLS_w_lag",c(1:q),sep=" ")


  if (PLSnew == TRUE) {
    inner_matrix<-matrix(c(0,0,0,0,0,0,1,1,0),3,3,byrow=T)
    colnames(inner_matrix)<-c("x","y_lag","y")
    rownames(inner_matrix)<-c("x","y_lag","y")
  }

  for (kk in 1:k) {

  results [kk, 1:p] <- data$beta
results [kk, p+1] <- data$rho
results [kk, 3*p+11] <- data$var.e
results [kk, 4*p+13] <- data$densita
y_lagged<-data$w1[[kk]]%*%data$y[[kk]]
results[kk,(3*p+12):(4*p+11)] <- cor(data$y[[kk]], data$x[[kk]])
results[kk,4*p+12] <- cor(data$y[[kk]],y_lagged)

     if (OLS == TRUE) {
   LM <- lm(data$y[[kk]] ~ 0+ y_lagged+ data$x[[kk]])
   results[kk,(2*p+5):(3*p+4)] <- LM$coefficients[2:(p+1)]
```

```
   results[kk,(3*p+5)] <- LM$coefficients[1]
   results[kk,(3*p+6)] <- summary(LM)$r.squared
   }


   if (SNA == TRUE) {
   LNAM <-lnam(data$y[[kk]],data$x[[kk]],data$w1[[kk]])
   rss <- sum(LNAM$residuals^2)
   mss <- sum((LNAM$fitted - mean(LNAM$fitted))^2)
    results[kk,(p+2):(2*p+1)] <-t(LNAM$beta)
    results[kk,(2*p+2)] <-LNAM$rho1
   results[kk,(2*p+3)] <-mss/(mss + rss)
   results[kk,(2*p+4)] <-round(-2 * LNAM$lnlik.model + 2 * LNAM$df.model,2)
   }

   #fit PLS new A
    if (PLSnew == TRUE) {
     t<-cbind(data$x[[kk]],y_lagged,data$y[[kk]])
     colnames(t)<- c(paste("x",c(1:p),sep="_"),"y_lag","y")
     t <- scale(t,T,F)
     PMnew <- myPLSPM_all_Wold(t,c(p,1,1),inner_matrix,scaling=
list(rep(c("RAW"),p),rep(c("RAW"),q),rep(c("RAW"),q)),outer.mode=
list(c("PLScow"),c("PLScow"),c("PLScow")),PLScomp= c(1,1,1),
inner.scheme= c("CEN"))
path.model <- lm(PMnew$QQ[[3]] %*% PMnew$w[[3]]~PMnew$QQ[[1]]
%*%PMnew$w[[1]] + PMnew$QQ[[2]] %*%PMnew$w[[2]])
   results[kk,(3*p+7)] <- path.model$coefficients[2]
   results[kk,(3*p+8)] <- path.model$coefficients[3]
   results[kk,(3*p+9)] <- summary(path.model)$r.squared
   results[kk,(3*p+10)] <- PMnew$GOF
   weights.X[kk,]<- unlist(PMnew$w)[1:p]
   weights.lagged[kk,]<- unlist(PMnew$w)[(p+1):(p+q)]
   weights.Y[kk,]<- unlist(PMnew$w)[(p+q+1):(p+q+q)]
   results[kk,(4*p+14):(5*p+13)] <- ((unlist(PMnew$w)[1:p])*
path.model$coefficients[2])
    }
  }
list(results=results,weights.X=weights.X,weights.lagged=weights.lagged,
weights.Y=weights.Y)
}
```

A.4 Results of three methods

Function useful to obtain the results of SNA, OLS and PLS-PM on simulated data

```
MySim.fit.data<-function (k,n,p,q=1,var.e,densita,sem){

dataset <- list()
risultati <-list()
i<-0
time.sim <- system.time(
for (t in 1:length(densita)){
 for (s in 1:length(var.e)){
  i<-i+1
    print("Results for simulation scheme number:")
print(i)
    dataset[[i]] <-MySim.data_2step(k,n,p,q,lambdax=sem[[1]],
    lambday=sem[[2]],phi=sem[[3]],tetax=sem[[4]],Beta=sem[[5]],
    Gamma=sem[[6]],nettuno=sem[[7]],tetay=sem[[8]],densita[t],var.e[s])
   risultati[[i]] <- MyFit.simulation(k,p,q,dataset[[i]],SNA=TRUE,OLS=TRUE,PLSnew=TRUE)
  }
 }
)
print (time.sim[1])
list(dataset=dataset, risultati=risultati)
}
```

Bibliography

[1] Abell, P., (1987). *The Syntax of Social Life*. Oxford, UK: Clarendon.

[2] Alon, A., Brunel, F. B., Schneier Siegal, W. L., (2004). Ritual behavior and community life cycle: Exploring the social psychological roles of net rituals in the development of online consumption communities. In *Online consumer psychology: Understanding how to interact with consumers in the virtual world*, C. Haugvedt, K. Machleit, Yalch (Eds), Hillsdale, NJ: Erlbaum.

[3] Amato, S., Esposito Vinzi, V., Tenenhaus, M., (2004). *A global goodness-of-fit index for PLS structural equation modeling*. Oral Communication to PLS Club, HEC School of Management, France, March 24.

[4] Anselin, L., (1982). A note on the small sample properties of estimators in a first order autoregressive model. In *Environment and Planning A*, pp. 1023 - 1030.

[5] Anselin, L., (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.

[6] Bagozzi, R. P., Dholakia, U. M., (2002). Intentional social action in virtual communities. *Journal of Interactive Marketing*, 16(2), pp. 2 - 21.

[7] Bollen, K. A., (1989). *Structural equations with latent variables*. Wiley, New York.

[8] Bollen, K. A., Bauldry, S., (2010). Model Identication and Computer Algebra. *Sociological Methods & Research*, 39, pp.127-156.

[9] Bollobas B., (2001). Random Graphs, Second Edition, *Cambridge studies in advanced mathematics*, 73 Cambridge University Press, Cambridge.

[10] Bonacich P., (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology* 92(5), pp. 1170 - 1182.

[11] Borgatti S. P., (2009). Social Network Analysis, Two-Mode Concepts. In *Encyclopedia of Complexity and Systems Science*, R. A. Meyers (Ed), Springer New York.

[12] Boudreau, M., D. Gefen, D. Straub (2001). Validation in IS Research: A State-of-the-Art Assessment. In *MIS Quarterly*, 25:1, pp. 1 - 23.

[13] Bramoullé, Y., Fortin B., (2009). The Econometrics of Social Networks, *Cahiers de recherche* 0913, CIRPEE.

[14] Breiger, R.L., (1974). Duality of persons and groups. *Social Forces* 53, pp. 181 - 190.

[15] Burt, R.S., Lin, N. (1977). Network time series from archival records. In Heise, D.R (ed.), Sociological Methodology. 1977, pages 224-254. San Francisco: Jossey-Bass.

[16] Burt, R.S., Doreian, P., (1982). Testing a structural model of perception: conformity and deviance with respect to journal norms in elite sociological methodology. *Quality and Quantity* 16, 109 -150.

[17] Butts C.T., (2008). Social network analysis: A methodological introduction. In *Asian Journal of Social Psychology* 11, pp. 13–41.

[18] Byrne, D., (1971). The attraction paradigm. New York: Academic Press.

[19] Cairns R.B., Perin J.E., Cairns B.D., (1985). Social structure and social cognition in early adolescence: Affiliative patterns. In *Journal of Early Adolescence*, 5(3), pp. 339-355.

[20] Cartwright, D. (1965). Influence, leadership, control. In *Handbook of organizations*, J. March (Ed.), Chicago: Rand McNally, pp. 1 - 47.

[21] Cartwright, D., Harary, F., (1956). Structural Balance: A Generalization of Heider's Theory, *Psychological Review*, 63, pp. 277-293.

[22] Cassel, C., Hackl, P., Westlund, A. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. In *Journal of Applied Statistics*, 26, pp. 435 - 446.

[23] Cassel, C., Hackl, P., Westlund, A., (2000). On measurement of intangible assets: a study of robustness of partial least squares. In *Total Quality Management*, 11, pp. 897 - 907.

[24] Chin, W. W., (1998). The partial least squares approach for structural equation modeling. In *Modern methods for business research*, G. A. Marcoulides (Ed.) London: Lawrence Erlbaum Associates, pp. 295 - 236.

[25] Chin, W. W., Newsted, P. R., (1999). Structural equation modelling analysis with small samples using partial least squares. In *Statistical strategies for small sample research* R. H. Hoyle (Ed.), Thousand Oaks, CA: Sage, pp. 307 - 341.

[26] Cliff, A. D. and Ord, J. K., (1973). *Spatial Autocorrelation*. Pion, London.

[27] Cliff, A.D., Ord, J.K., (1981). *Spatial Processes: Models and Applications*, Pion, London.

[28] Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.

[29] Currarini S., Vega Redondo F., (2010). *Search and Homophily in Social Networks*, Working Papers 24, Department of Economics, University of Venice "Ca' Foscari".

[30] Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., Baum, S., (2012). Net- works and geography: Modelling community network structures as the outcome of both spa- tial and network processes. *Social Networks*, 34 (1), pp. 6 - 17.

[31] Daudin, J.J., Picard, F., Robin S., (2008). A mixture model for random graphs. In *Statistics and computing*, 18(2), pp. 173 -183.

[32] Davis J.A. (1970). Clustering and hierarchy in interpersonal relations: Testing two theoretical models on 742 sociograms. *American Sociological Review*, 35, pp. 843-852.

[33] de Klepper, M., Sleebos, E., van de Bunt, G., Agneessens, F. , (2010). Similarity in friendship networks: selection or influence? The effect of constraining contexts and non-visible individual attributes. *Social Networks*, 32, pp. 82–90.

[34] de Miguel Luken, V., Tranmer, M. (2010). Personal support networks of immigrants to Spain: A multilevel analysis. In *Social Networks*, 32, pp. 253–262.

[35] de Sola Pool, I., Kochen, M. (1978). Contacts and influence. *Social Networks*. 1, pp. 5 - 51.

[36] Deaux, K., Martin, D., (2003). Interpersonal networks and social categories: specifying levels of context in identity processes. *Social Psychology Quarterly*, 66, pp. 101 - 117.

[37] Dholakia, U. M., Bagozzi, R. P., (2004). Motivational antecedents, constituents and consequents of virtual community identity. In *Virtual and collaborative teams: Process, technologies, and practice*, S. Godar, S. Pixie-Ferris (Eds.), London: IDEA Group, pp. 252– 267.

[38] DiMaggio P. J., Powell W., (1983). The iron cage revisited institutional isomorphism and collective rationality in organizational fields, *American Sociological Review*, 48, pp. 147-60.

[39] Dijkstra, T., (1983). Some comments on maximum likelihood and partial least squares methods. In *Journal of Econometrics, 22*, pp. 67- 90.

[40] Doreian, P., (1980). Linear-models with spatially distributed data-spatial disturbances or spatial effects. In *Sociological Methods Research*, 9(1), pp. 29 - 60.

[41] Doreian, P., (1981). Estimating linear models with spatially distributed data. In S. Leinhardt (editor), *Sociological Methodology*, Jossey-Bass, San Francisco, pp. 359 - 388.

[42] Doreian P. (1989a) Network autocorrelation models: problems and prospects. In D.A. Griffith, (editor), *Spatial Statistics: Past, Present, Future*, Michigan Document Services, Ann Arbor, pp.369 - 389.

[43] Doreian, P. (1989b) Models of network effects on social actors. In *Research methods in social analysis*, L. C. Freeman, D. R. White and K. Romney (Eds), George Mason University Press, Fairfax, pp. 295 - 317.

[44] Doreian P., (1996). When the Data Points are not Indipendent. In: *Developments in Data Analysis, Metodološki zvezki*, A. Ferligoj and A. Kramberger (Eds),12, Ljubljana: FDV, pp. 27 - 46.

[45] Doreian P., (2001). Causality in Social Network Analysis. In: *Sociological Methods & Research*, Sage Publications, Vol. 30 No. 1, pp. 81-114.

[46] Doreian, P., Stokman, F. N., (1997). *Evolution of Social Networks*. Gordon and Breach Publishers, Amsterdam.

[47] Doreian, P., Conti, N. (2012). Social context, spatial structure and social network structure. In *Social Networks*, 34, pp.32 - 46.

[48] Doreian, P., Porzio, G. C., Vitale , M. P., (2013). Including Network Effect in Structural Equation Models: Student Communities and Performance In *Book of abstracts of International Workshop ARS'13 Networks in Space and Time: Models, data collection and applications*, pp. 33 - 34.

[49] Doreian P., Stokman F. N., (2003). Introduction to Special issue on Evolution of Social Networks. Part III. In *Journal of Mathematical Sociology*, 27, pp. 85-87.

[50] Dow, M. M., Burton, M.L., White, D.R., (1982). Network autocorrelation: A simulation study of a foundational problem in regression and survey research. *Social Networks, 4*, pp. 169 - 200.

[51] Durkheim, E., (1938). *The rules of sociological method*, University of Chicago sociological series, University of Chicago.

[52] Efron, B., Tibshirani, R.J., (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

[53] Erdős, P., Rényi, A., (1959). On Random Graphs. I. *Publicationes Mathematicae*, 6: pp. 290 - 297.

[54] Esposito Vinzi, V., Trinchera, L., Amato, S. (2010). PLS Path Modeling: From Foundations to Recent Developments. In Esposito Vinzi, Chin, Henseler, Wang (editors), *Handbook of Partial Least Squares*, Springer: Heidelberg, pp. 47 - 82.

[55] Esposito Vinzi, V., Russolillo, G., (2013) Partial least squares algorithms and methods. In *WIREs Computational Statistics*, 5 (1), pp. 1 - 19.

[56] Esposito Vinzi, V., Trinchera, L. (2013). Composite-based Predictive and exploratory path modelling and multi-block data analysis. In *59th World Statistics Congress of the International Statistical Institute (ISI)*

[57] Festinger L, Schachter S, Back KW. (1950). *Social Pressures in Informal Groups: A Study of Human Factors in Housing*. New York: Harper.

[58] Fienberg, S. E., Wasserman S., (1981). An Exponential Family of Probability Distributions for Directed Graphs: Comment. *Journal of the American Statistical Association*, 76, pp. 54 - 57.

[59] Folmer, H., Oud, J.H.L., (2008). A Structural Equation Approach to Models with Spatial Dependence. In *Geographical Analysis*,40 (2), pp. 152 - 166.

[60] Fornell, C., Bookstein, F. L., (1982). Two structural equation models: LISREL and PLS appliead to consumer exit-voice theory. In *Journal of Marketing Research 19*, pp. 440 - 452.

[61] Forsyth, E., Katz, L., (1946). A matrix approach to the analysis of sociometric data: Preliminary report. *Sociometry*, 9, 340-347.

[62] Frank O., Strauss D., (1986). Markov graphs. *Journal of the American Statistical Association*, 81, pp. 832 - 842.

[63] Freeman Linton C., (1979). Centrality in social networks: Conceptual clarification,*Social Networks*, 1(3), pp. 215 - 239.

[64] Freeman Linton C., (1989). Network Representations. In *Research Methods in Social Network Analysis*, L. C. Freeman, D. R. White, A. K. Romney, Fairfax, Va: George Mason University Press, pp. 11-40.

[65] Freeman, L.C., Freeman, S.C., Michaelson, A.G. (1988). On human social intelligence. lournal of Social and Biological Structures. 11, 41 5-425.

[66] Freeman L. C., Webster C.M. (1994). Interpersonal proximity on social and cognitive space. *Social cognition*, 12(3), pp. 223-247.

[67] Friedkin, N.E., (1995). *A Structural Theory of Social Influence*. Cambridge University Press, Cambridge.

[68] Friedkin, N.E., (1998). *A Structural Theory of So- cial Influence*. Cambridge, UK: Cambridge Univ. Press.

[69] Friedkin, N. E., (1999). Choice Shift and Group Polarization. *American Sociological Review*, 64, pp. 856-75.

[70] Friedkin, N.E. 2003. Social Influence Network Theory: Toward a Science of Strategic Modification of Interpersonal Influence Systems. In *Dynamic Social Network Modeling and Analysis*, R. Breiger, K. Carley, P. Pattison. (Eds). National Academy of Sciences/National Research Council Committee on Human Factors. Washington, D.C., pp. 89-100.

[71] Friedkin, N.E., Cook, K.S., (1990). Peer group influence. In *Sociological Methods & Research*. 19, pp. 122 - 143.

[72] Friedkin, N.E., Johnsen, E.C., (1990). Social Influence and Opinions. *Journal of Mathematical Sociology*, 15, pp. 193-205.

[73] Friedkin, N.E., Johnsen, E.C., (1997). Social Positions in Influence Networks. *Social Networks* 19, pp. 209-22.

[74] Friedkin, N.E., Johnsen, E.C., (1999). Social Influence Networks and Opinion Change. *Advances in Group Processes*, 16, pp. 1-29.

[75] Friedkin, N.E., Johnsen, E.C., (2011). *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. New York: Cambridge University Press.

[76] Gilbert, E. N. (1959). Random Graphs. *The Annals of Mathematical Statistics* 30 (4), pp. 1141 -1144

[77] Goldberg M., Kelley S., Magdon-Ismail M., Mertsalov K., Wallace A., (2010). *Finding overlapping communities in social network*, SocialCom.

[78] Granovetter, M., (1973). The strength of weak ties. *American Journal of Sociology*, 81, pp. 1287-1303.

[79] Griffith, D. A. (1976) Spatial autocorrelation problems: some preliminary sketches of a structural taxonomy . In *The East Lakes Geographer, 11*, pp. 59 - 68.

[80] Haining, R. (1978). The moving average model for spatial interaction. In *Transactions of the Institute of British Geographers 3*, pp. 202 - 225.

[81] Hanafi, M. (2007). PLS Path modelling: computation of latent variables with the estimation mode B, *Computational Statistics* 22, pp. 275 - 292.

[82] Handcock, M.S., Raftery, A. E., Tantrum, J. M., (2007). Model-Based Clustering for Social Networks. In *Journal of the Royal Statistical Society*, 170 (2), pp. 301 - 354.

[83] Harary, E, Norman, R.Z., Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs*. New York: John Wiley and Sons.

[84] Hars, A., Ou, S. (2002). Working for free? Motivations for participating in open-source projects. *International Journal of Electronic Commerce*, 6(3), pp. 23 - 37.

[85] Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology*, 21, pp. 107-112.

[86] Henseler, J., (2010). On the convergence of the partial least squares path modeling algorithm. *Computational Statistics*, 25 (1), pp. 107 - 120.

[87] Hoff, P.D. (2005). Bilinear Mixed Effects Models for Dyadic Data. In *Journal of the American Statistical Association*, 100(469), pp. 286 - 295.

[88] Hoff, P. D., Raftery, A.E., Handcock, M.S., (2002). Latent Space Approaches to Social Network Analysis. In *Journal of the American Statistical Association*, 97(460), pp. 1090 -1098.

[89] Hogg, M. A., Terry, D. J., (2000). Social identity and self-categorization processes in organizational contexts. *Academy of Management Review*, 25, pp. 121-140.

[90] Holland, P.W., and Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*. 2, pp. 107-124.

[91] Holland P.W., Leinhardt S., (1972). Some evidence on the transitive of positive interpersonal sentiment. *American Journal of Sociology*, 72, pp. 1205 -1209.

[92] Holland, P.W., Laskey, K.B., Leinhardt S., (1983). Stochastic blockmodels: First steps. *Social Networks*, 5 (2), pp. 109 - 137.

[93] Huisman, M., Snjiders, T.A.B. (2003) Statistical analysis of longitudinal network data with changing composition. In *Sociological Methods & Research*, 32, pp.253 - 287.

[94] Huston, T. L., Levinger, G., (1978). Interpersonal attraction and relationships. In *Annual Review of Psychology*, M. R. Rosenzweig and L. W. Porter (Eds), 29.

[95] Hwang, H., and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, 69, pp. 81 - 99.

[96] Jöreskog, K. (1970) A general method for analysis of covariance structure. In *Biometrika, 57*, pp. 239 - 251.

[97] Kaplan, D. (2000). *Structural equation modeling. Foundations and extensions.* Thousand Oakes, CA: Sage Publications.

[98] Koehly L. M., Pattison P., (2005). Random Graph Models for Social Networks: Multiple Relations or Multiple Raters. In *Models and Methods in Social Network Analysis*, Carrington, P.J., Scott, J., Wasserman, S. (Eds.), Cambridge University Press, pp.162 - 191.

[99] Kogovšek T., (2006). Reliability and validity of measuring social support networks by Web and telephone. *Metodološki zvezki*, pp. 239 - 252

[100] Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9, pp. 109 - 134.

[101] Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In *Networks and Organizations: Structure, Form, and Action*, N. Nohria, R. Eccles (Eds.) Boston MA: Harvard Business School Press, pp. 216 - 239.

[102] Krämer, N. (2007). *Analysis of high-dimensional data with partial least squares and boosting.* Doctoral dissertation, Technischen Universitt Berlin.

[103] Laumann, E. (1979). Network analysis in large social systems: Some theoretical and methodological problems. In *Perspectives on social network research*, Holland, P., Leinhardt, S., (Eds), Academic Press, pp. 379–402.

[104] Laumann, E.O., Marsden, P.V., Prensky, D. (1989). The boundary specification problem in network analysis. In *Research Methods in Social Network Analysis*, Freeman, L.e., White, D.R., Romney, A.K. (Eds.), Fairfax, VA: George Mason University Press, pp. 61 - 87.

[105] Lazarsfeld, P.F., Henry, N. W., (1968). *Latent Structure Analysis*, Boston: Houghton Mifflin.

[106] Leenders, R.Th.A.J., (1995). *Structure and influence: statistical models for the dynamics of actor attributes, network structure and their interdependence*. Thela Thesis Publishers, Amsterdam.

[107] Leenders, R. Th. A. J., (1997). Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. In *Evolution of Social Networks*, P. Doreian, and F. N. Stokman (Eds.), Amsterdam: Gordon and Breach, pp. 165 -184.

[108] Leenders, R.Th.A.J., (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24 (1), pp. 21 - 47.

[109] Liu A., Folmer, H., Oud, J.H.L. (2011) W-based versus latent variables spatial autoregressivemodels: evidence from Monte Carlo simulations. In *The Annals of Regional Science, 47 (3)*, pp. 619 - 639.

[110] Lohmöller, J., (1989). *Latent variable path modeling with partial least squares*. Heildelberg:Physica-Verlag.

[111] Luce, R.D., Perry, AD. (1949). A method of matrix analysis of group structure. *Psychometrika*. 14, 95-116.

[112] Marin A., Hampton K.N. (2007). Simplifying the personal network name generetor. Alternatives to traditional multiple and single name generators. *Field Methods*, 19(2), pp. 163 - 193.

[113] Marsden, P.V., (1987). Core discussion networks of American. *American Sociological Review*, 52(1), pp. 122 - 131.

[114] Marsden, P.V., (1990). Network Data and Measurement. In *Annual Review of Sociology*. 16, pp. 435 - 463.

[115] Marsden, P.V. (2011). Survey Methods for Network Data. In *The Sage Handbook of social Network Analysis*, John Scott and Carrington, Peter J (Eds), London: Sage Publications, pp. 370 - 388.

[116] Marsden, P., Friedkin, N., (1994). Network studies of social influence. In *Advances in social network analysis: Research in the social and behavioral sciences*, S. Wasserman, J. Galaskiewicz (Eds.), Thousand Oaks, CA: SAGE Publications, Inc., pp. 3 - 26.

[117] McDonald, R. P., (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, 31(2), pp. 239–270.

[118] McKenna, K. Y. A., Bargh, J. A. (1999). Causes and consequences of social interaction on the internet: A conceptual framework. *Media Psychology*, 1, pp. 249 - 269.

[119] McPherson, M., Smith-Lovin, L., Cook J. M., (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27: 415 - 444.

[120] Mead, G., H., (1925). The Genesis of the Self and Social Control, *International Journal of Ethics*, 35, pp. 251 - 277.

[121] Mead,R. (1967). A mathematical model for the estimation of inter-plant competition. In *Biometrics*, 23, pp. 189 - 205.

[122] Mizruchi M.S., Neuman E. J., (2008). The effect of density on the level of bias in the network autocorrelation model. *Social Networks*, 30, pp. 190 - 200.

[123] Moreno J.L., (1953). *Who shall survive? Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Student ed. Roanoke, VA: Royal.

[124] Nadel, S.P., (1957). *The Theory of Sociol Structure*. New York: Free Press.

[125] Newcomb, T.M. (1953). An approach to the study of communicative acts. *Psychological Review*. 60, pp. 393 - 404.

[126] Nowicki, K. and Snjiders, T.A.B. (2001) Estimation and Prediction for Stochastic Blockstructures. In *Journal of the American Statistical Association*, 96, pp. 1077 - 1087.

[127] O'Malley, A.J., Marsden, P.V., (2008). The Analysis of Social Networks. In *Health Services and Outcomes Research Methodology, 8*, pp. 222 - 269.

[128] Ord, J.K., (1975). Estimation Methods for Models of Spatial Interaction, *Journal of American Statistical Association*, 70.

[129] Oud J.H., Folmer, L., (2008). A Structural Equation Approach to Models with Spatial Dependence, *Geographical Analysis*, 40, pp. 97 - 211.

[130] Pattison P.E., Wasserman S. (1999). Logit models and logistic regressions for social networks. II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52, pp. 169 -194.

[131] Pattison, P. E., Robins, G. L. (2002). Neighbourhood-based models for social networks. *Sociological Methodology*, 32, pp. 300 - 337.

[132] Paxton P., Curran P.J., Bollen K.A., Kirby J., Chen F. (2001). Monte Carlo Experiments: Design and Implementation. *STRUCTURAL EQUATION MODELING*, 8(2), pp. 287–312.

[133] Postmes, T., Spears, R., Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3), pp. 341 - 371.

[134] Robins G.L., Pattison P.E., Wasserman S. (1999). Logit models and logistic regressions for social networks. III. Valued relations. *Psychometrika*, 64, pp. 371 - 394.

[135] Robins G., Pattison P., Elliott P. (2001). Network Models For Social Influence Processes, *Psychometrika*, 66 (2), pp. 161-190.

[136] Robins,G.L., Snijders T.A.B., Wang P., Handcock M., Pattison P.. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*,29 (2), pp. 192 - 215.

[137] Shalizi, C. R., Thomas A. C., (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, *Sociological Methods and Research*, 40, pp. 211–239.

[138] Simon, H.A., 1957, *Models of Man*, New York, Wiley & Sons.

[139] Snjiders, T.A.B. (2005). Models for Longitudinal Network Data. In *Models and methods in social network analysis*, P. Carrington, J. Scott and S. Wasserman (Eds), New York: Cambridge University Press, pp. 215 - 247.

[140] Snjiders, T.A.B. and Nowicki, K.(1994). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. In *Journal of Classification*, 14, pp. 75 -100.

[141] Snjiders, T.A.B., Spreen, M. and Zwaagstra, R., (1995). The use of multilevel modelling for analysing personal networks (Networks of cocaine users in an urban area). In *Journal of Quantitative Anthropology, 5*, pp. 85 - 105.

[142] Snjiders, T.A.B., Steglich, C.E.G. and Schweinberger, M. (2007) Modeling the co-evolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences*, K. van Montfort, H. Oud and A. Satorra (Eds), Lawrence Erlbaum, pp. 41 - 71.

[143] Snjiders, T.A.B., van de Bunt, G. G. and Steglich, C.E.G. (2010). Introduction to stochastic actor-based models for network dynamics. In *Social Networks, 32*, pp. 44 - 60.

[144] Steglich, C.E.G., Snjiders, T.A.B. and Pearson, M. (2010) Dynamic Networks and Behavior: Separating Selection from Influence. In *Sociological Methodology, 40*, pp. 329 - 392.

[145] Stets, J.E., Burke, P.J., (2000). Identity theory and social identity theory. *Social Psychol- ogy Quarterly*, 63 (3), pp. 224 - 237.

[146] Stokman, F. N., Doreian P., (1996). Concluding remarks. In *Journal of Mathematical Sociology (Special double issue on Evolution of Social Networks)*, 21, pp. 197-199

[147] Stokman, F. N., Doreian P., (2001). Introduction to Special issue on Evolution of Social Networks. Part II. In *Journal of Mathematical Sociology* , 25, pp. 1-4

[148] Tajfel, H., (1978). Interindividual behaviour and intergroup behavior. In *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*, Tajfel, H. (Ed.), Academic Press, London, pp. 27–60.

[149] Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. and Lauro, C (2005) PLS path modeling. In *Computational Statistics and Data Analysis, 48*, pp.159–205.

[150] Tenenhaus, A. and Tenenhaus, M. (2011):Regularized generalized canonical correlation analysis. *Psychometrika, 76 (2)*, pp.257–284.

[151] Travers, J., Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*. 32, 425-443.

[152] van Duijn M. A. J. , Snijders T. A. B., Zijlstra B. J. H. (2004). p2: a random effects model with covariates for directed graphs *Statistica Neerlandica* 58 (2), pp. 234 - 254.

[153] Vehovar V., Lozar Manfreda K., Koren G., Hlebec V. (2008). Measuring egocentered social networks on the Web. *Social networks*, 30(3), pp. 213-222.

[154] Wasserman, S., C. Anderson (1987) Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks* 9, pp. 1-36.

[155] Wasserman S., Faust K., (1994) *Social Network Analysis: Methods and Applications*. New York: Academic Press.

[156] Wasserman, S. Pattison, P. (1996). Logit models and logistic regressions for social networks. I: An introduction to Markov graphs and p*. *Psychometrika*, 60, pp. 401 - 425.

[157] Wertz, C., Linn, R., Joreskog, K. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34(1), pp. 25–33.

[158] White, D, Burton, M, and Dow, M (1981) Sexual division of labor in Africa: A network autocorrelation analysis. *American Anthropologist*, 83, pp. 824-849.

[159] Whittle, P. (1954) On stationary processes in the plane. In *Biometrika, 41*, pp. 434 - 449.

[160] Wold H., (1975). Path models with latent variables: the nonlinear iterative partial least squares (NIPALS) approach. In *Quantitative Sociology: Intentional Perspective on Mathematical and Statistical Modeling*, Blalock H.M., Aganbegian A., Borodkin F.M., Boudon R. and Capecchi V., (Eds), New York: Academic Press, pp. 307 – 357.

[161] Wold, H. (1980) Model construction and evaluation when theoretical knowledge is scarce: On the theory and application of Partial Least Squares. In *Model Evaluation in Econometrics*, J. Kmenta and J. Ramsey (Eds), New York: Academic Press, pp. 47 - 74.

[162] Wold, H. (1982) Soft modeling: the basic design and some extensions, In *Systems under Indirect Observation, 2*, K. G. Joreskog and H. Wold (Eds), North-Holland, Amsterdam, pp. 1–54.

[163] Wold, H. (1985) Partial Least Squares. In *Encyclopedia of Statistical Sciences*, Kots S., Johnson N.L. (Eds). Wiley: New York, 6, pp. 581-591.

[164] Zeggelink, E. (1995): Evolving friendship networks: An individual-oriented approach implementing similarity. *Social Networks*, 17, pp. 83-110.

[165] Zhang, B., Thomas, A. C., Doreian, P., Krackhardt, D., Krishnan R., (2013) Contrasting Multiple Social Network Autocorrelations for Binary Outcomes, With Applications To Technology Adoption. *ACM Trans. Management Inf. Syst.*, 3(4):18.

[166] Zijlstra B. J. H. , van Duijn, M. A. J., Snijders, T. A. B. (2006) The Multilevel p2 Model A Random Effects Model for the Analysis of Multiple Social Networks *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2, pp. 42-47.