Università degli Studi di Milano - Bicocca

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Dottorato di Ricerca in Statistica XXIII ciclo

Tesi di dottorato di Ricerca

New statistics for the parameter estimation of the stochastic actor-oriented model for network change

Discussant:
Viviana Amati

Tutor:
Prof. Piero Quatto

Co-Tutor:
Prof. Tom A.B. Snijders

Academic Year 2009-2010

Un sorriso non costa nulla e rende molto.
Arricchisce chi lo riceve, senza impoverire chi lo dona.
Non dura che un istante, ma il suo ricordo è talora eterno.
Nessuno è così ricco da poterne fare a meno.
Nessuno è così povero da non poterlo fare.
Crea felicità in casa; sostegno negli affari;
è segno sensibile dell'amicizia profonda.
Un sorriso dà riposo alla stanchezza;
nello scoraggiamento rinnova il coraggio.
nella tristezza una consolazione;
d'ogni pena è il natural rimedio.
Ma è un bene che non si può comprare, nè prestare, nè rubare,
poiché esso ha un valore solo nell'istante in cui si dona.
E se poi incontrerete talora chi non vi dona l'atteso sorriso,
siate generosi e date il vostro;
perché nessuno ha tanto bisogno di un sorriso
come chi non sa darlo agli altri.

(P. Faber)

Grazie di cuore a tutti coloro
che mi hanno donato e continuano a donarmi
il loro sorriso!

Acknowledgments

Various people directly or indirectly contributed to the realization of this thesis. Particular acknowledgment is given to Professor Piero Quatto who discussed some key questions with me and supported me in the writing of the thesis.
I wish to thank to Professor Tom A.B. Snijders, for his helpful comments and advice and for giving me the opportunity to spend three months of my PhD at the Nuffield College in Oxford. My life suddenly changed in September 2009 when I moved to this beautiful angle of the Earth: new people to work with, new friends and a new environment. This was one of the greater experience of my life!
Finally I express profound gratitude to Laura Terzera and Giulia Rivellini, who made this possible. Six years ago they introduced me to the "tangled" world of network analysis, and over the years they have helped me in finding my way. They also suggested that I "leave the nest" and spend some time abroad, an experience that I will remember for the rest of my life and that allowed me to take a glance at the rest of the world, far from "my house and my garden".

Contents

List of Tables

List of Figures

Introduction

Social Network Analysis (SNA) is a particular set of methods that has been developed since the 1930's. It has attracted considerable interest from the social and behavioral science in the last decades. A growing interest in applications and in the available statistical tools was registered in recent years.
Formally speaking a network is composed of a set of actors, having some members connected by a set of one or more relations. Thus, the concept of network strictly belongs to human nature, since networks are the result of social contacts which we experiment in everyday life. But, it can also be extended to a variety of other contexts, such as Biology, Medicine, Politics and so forth.
The peculiarity of SNA is represented by the data which can be analyzed. In addition to the classical statistical variables, which describe actor characteristics, there are relational data, which refer to the ties that exist among actors.
Different types of ties exist. Here we considered non-reflexive and directed ties, meaning that an actor cannot be related to himself and if i and j are two actors of a network, the presence of a tie from i to j, does not imply that there is a tie from j to i.
Intuitively, ties generate dependence among actors. If we consider two people as actors and friendship as a relation, then it is well-known that friends influence each other. This dependence requires specific statistical tools different from those that are usually used in common statistical analysis.
From a descriptive point of view, we can measure network properties using a variety of indexes. These measures are not sufficient because the interest also focus on network properties and on the mechanisms that leads to the observed network. For this reason a series of model were proposed. These models differ from the usual statistical models for the kind of inference that they require.
In fact, in statistics one usually deals with samples, and he/she wants to extrapolate information about a certain characteristic of a population. This requires a process of generalizing sample results to the population according to precise statistical rules which satisfy particular requirements. One of the widespread assumptions underlying this operation is the independence assumption between the units of the population.
In SNA the perspective is totally different. One studies a group of individuals (the set of actors) and he/she is interested in finding results concerning structures present in this specific group. As previously pointed out, the assumption of independence among actors is not reasonable because it cannot deal with the dependence structures existing

between them.
Thus, the usual sampling-based inference has no meaning in network context, instead the model based inference play a key role.
Different models for network data are proposed in the literature. Two bigger strands of models may be distinguished, according to the nature of network data. The former regards the models for cross-selectional data, i.e. one single network observation at a certain time point. In this situation the interest is to determine which are the main structures that characterize ties formation. The latter focus the attention on longitudinal models, i.e. on repeated observation of a same network at different time points. The purpose is to establish which are the leading forces that govern network dynamics.
Among the models for longitudinal network data, there is the Stochastic actor oriented (SAO) model (Snijders, 1996, 2001; Snijders *et al.*, 2010b), which assumes that the evolution process is distributed as a Continuous time Markov chain and the actor who has the opportunity to change can decide to modify or not to modify one of his outgoing ties according to a random utility model.
In SAO model, the most often used procedure for parameter estimation is the Method of Moments (MoM), which estimates the parameters using one observed statistic for each estimated parameter.
The idea of this work is to define new statistics which take into account the different ways of creating and deleting ties to which a certain effect can contribute. In this manner more information deriving from network observation at each time point can be included in the estimation process and intuition suggests that if the observation are close in time then the new statistics give rise to estimates with a smaller standard error than the usual one.
To empirically investigate what intuition suggests, the attention is focused only on two basilar reciprocity and transitivity effects, with the idea to extend the approach to other network effects. Furthermore, only two observational time points are considered.
The definition of new statistics leads to having more than one statistic for a single parameter. The result is that the principle of regular MoM gives rise to an over-identified system of equations, so that the ordinary MoM cannot be applied. A suitable method then is the Generalized Method of Moments (GMM), an estimation technique mainly used in econometrics, and potentially more efficient than the MoM.
Like the regular MoM, the GMM is based on the differences between the expected values of the statistics and their sample counterparts, but the GMM involves the minimization of a quadratic function of these differences rather than setting all differences to 0. This means that an extra problem arises: the determination of a matrix of weights reflecting the different importance and correlations of the statistics employed.
A stochastic optimization-simulation algorithm is used, following the approach suggested by Gelman (1995) and based on the Newton-Raphson algorithm, in order to approximate the solution. The algorithm is implemented using R, a free software for statistical computing and graphics.
In order to explain the idea and the proposed solution a brief introduction to social network analysis is provided in Chapter 1. This chapter does not pretend to give a complete

overview of all the available concepts developed since the 1930's, but it introduces only the key words that are useful for understanding the idea underlying this work.
Chapter 2 and Chapter 3 complete the description of statistical tools, focusing the attention on models for network data. The former focus on the difference between social network models and the usual statistical models and describes models for cross-sectional network data and simple models for network dynamics. The latter provides an overview of the SAO model, a model for longitudinal data which relaxes the inconvenient and unreal assumption of dyadic independence. An entire chapter is dedicated to the SAO model since the original part of the thesis focus on the estimation of its parameters.
At this point the innovative elements of the work are described in Chapter 4 and Chapter 5. In Chapter 4 the idea behind the definition of different statistics to estimate the parameters of the SAO model is introduced, also pointing out the problems that it causes. Then, the solution is theoretically presented, justifying the application of the GMM in the process of estimation and the implementation of a stochastic algorithm in order to obtain the estimate.
Chapter 5 shows how the theoretical considerations of Chapter 4 can be practically applied. The implemented algorithm that provides the estimates is described in detail and then the results are presented. Furthermore, computational considerations about the R code implemented and the use of parallel computing are discussed. A proposal of further improvements is also given in order to trace future directions and research developments.

Chapter 1

Social network analysis origins and tools

Social network concepts were developed in the early 1930's when different research groups became aware of the importance of relations to explains behaviors and patterns of ties within a set of people. Recently, the notion of social network has attracted curiosity and interest from many scientific communities leading to a wide variety of applications. The development of specific tools for social network analysis and the growing possibility of jointly analyzing relational and attribute data were the keys for the diffusion of social network techniques.

1.1 Networks are everywhere

In recent years, journal headlines have focused on the spread of "social networks" like Facebook, MySpace or Twitter, which are able to connect people who live in very different and distant areas of the world. This idea of "connected people" is not new and arose before the advent of Internet and its related features such as e-mails, chat-rooms, blogs and so on.
More specifically it takes its origins from the late 1960's when the psychologist Stanley Milgram talked about the well known *Six Degrees of separation* (Milgram, 1967; Travers & Milgram, 1969). He proved that everyone is linked to everyone else in the planet just through six intermediaries using a simple experiment. He asked some selected people to send a message to a target person, living in a distant place, using only a chain of friends and acquaintances. The result was that the chains which reached the target person had a mean length of 6. This surprising discovery is known as the *small world* phenomenon because of the well known expression "It's a really small world" that we claim when we encounter someone far from home, who turns out to share a mutual acquaintance with us.
Internet capabilities of joining people is just an example of social networks. Roughly speaking, a social network consists of social actors and one or more relationships defined over them. It is not difficult to imagine that networks play a key role in everyday life,

since individuals are part of a society through their contacts with others.
If we think about an ordinary day, we can recognize many networks in which we take part. We wake up and we meet our family members, then we go to work, maybe by train or underground or bus, and we meet our acquaintances or fellow travelers, and when we arrive at work we start interacting with our colleagues. Finally, we come back home and we go out with our friends.
This simplified reconstruction of our everyday life shows that we experience different relationships and different relationships mean different networks: family, acquaintances, cooperation and friendship networks are just very simple examples. But something is missing in our reconstruction. In fact, when we are traveling to work we face other types of networks which do not necessary involve other people: the map of the public services or the map that the navigator presents us are networks themselves.
An overview of the presence of networks in real life is in literature, which provides many studies.

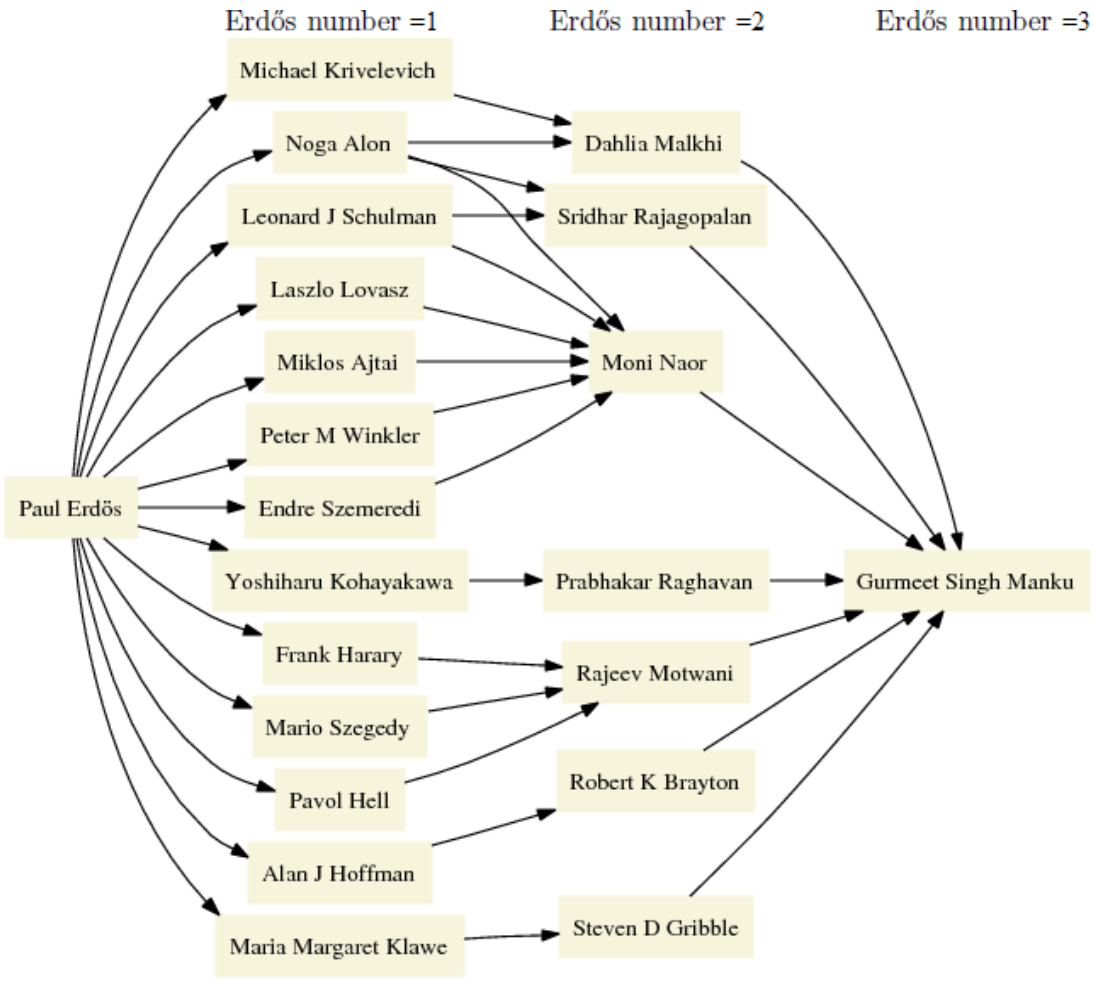
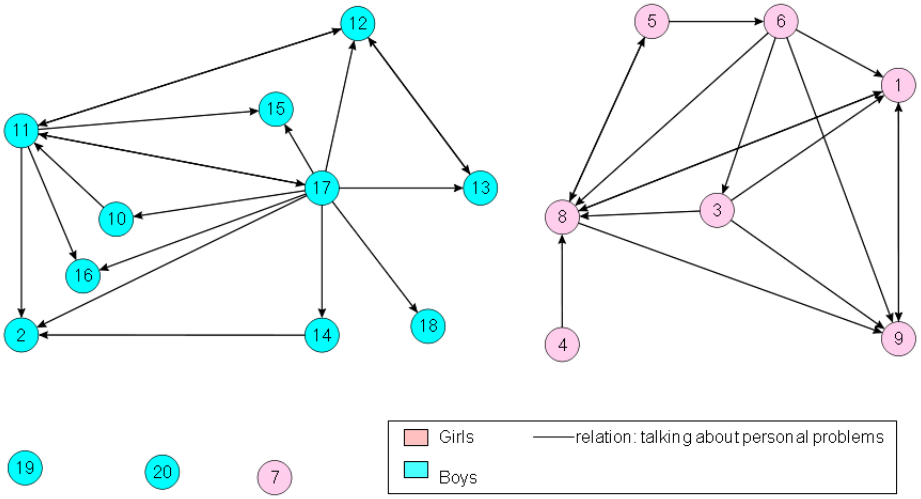
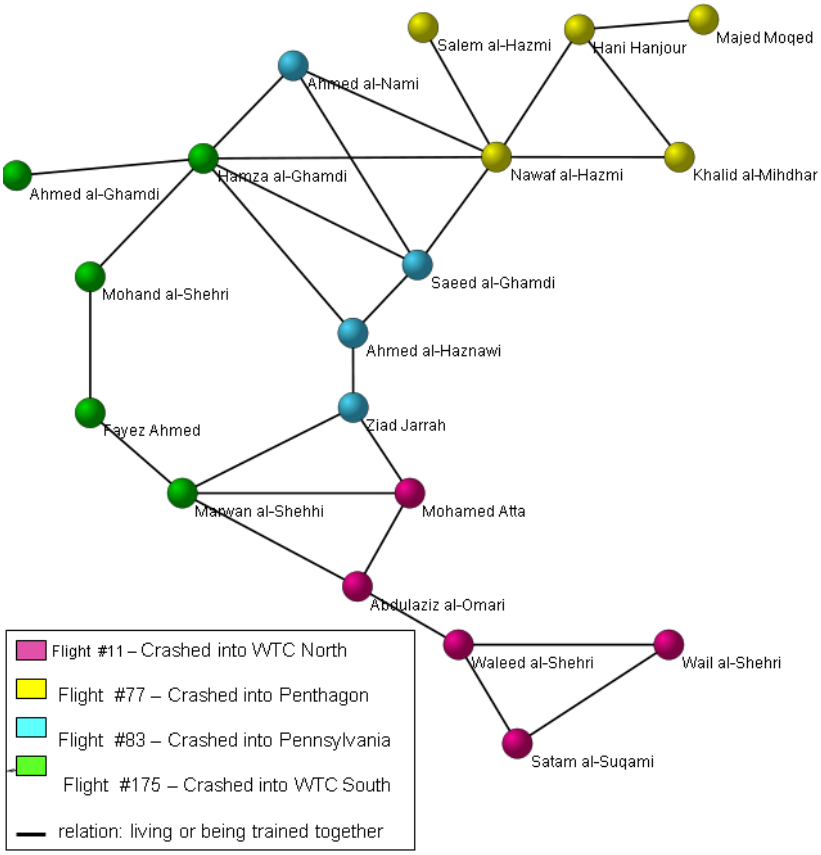


Figure 1.1: *A small subgraph of Erdős collaboration network*

Let's start considering people as social actors. Common examples of relationships are friendships (Moreno, 1953; Hallinan, 1979), social support (Wellman, 1981) and e-mail exchange (Freeman, 1984). These are very classical topics that have been studied using network analysis. In recent years the studies of the existence of ethnic boundaries in populations and the integration of the second and third generation of migrants (Baerveldt *et al.*, 2004; Moody, 2001; Vermeij *et al.*, 2009), the pattern of sexual contacts (Bearman *et al.*, 2004) and the structure of communication between terrorists (Krebs, 2002; Rothenberg, 2001; Jordan, 2008) have drawn network analysts attention.
Many people work in organizations in which there are social relations such as "receive order from", "receive advice from", "collaborate with" and so on (Krackhardt, 1990; Krackhardt & Hanson, 1997; Lazega & Van Duijn, 1997; Contractor *et al.*, 2006). The last relationship defines the so called collaboration networks. A classical example is given by the co-authorship among academics (Newman, 2004b,a; Liu *et al.*, 2005; Rodriguez & Pepe, 2008). Actors are academics, and they are linked if they have co-authored at least one paper.
Erdős number is one of the most famous instance of co-authorship networks (Grossman *et al.*, 2003). Paul Erdős was an Hungarian mathematicians "who apparently spent a large portion of his later life living out of a suitcase and writing papers with those of his colleagues willing to give him room and board" (Newman, 2003). He published at least 1,401 papers during his life, working with a lot of people who defined Erdős number in his honor. Erdős number is a measure of proximity to the great mathematician. In more detail those who have published a paper with Erds have an Erdős number of one. Those who have published with a co-author of Erdős have an Erdős number of 2, and so on (Figure 1.1). This means that Erdős number is the shortest path between Erdős and people who can reach him through at least one path.
The analysis of this kind of network helps to discover the presence of social structures and to find an answer to some questions like: "Are friends of friends also friends?", "Are friends similar?", "Are ethnic differences a barrier against integration?", "Who plays a key role in the network?", "Which people should be isolated in order to avoid the spread of a pandemia?", "Who interacts with whom?", "Who gives advice to whom?" and so on.
In Figures 1.2 and 1.3 some networks are depicted. Circles represent individuals and lines represent relationships. Figure 1.2(a) illustrates a network of social support between secondary school classmates. The actors are the pupils, and the question that defines the relationship is "who do you talk to about personal problems?". Someone can be interested in the role played by gender and by reciprocity: "Do girls trust only girls?", "Do boys trust only boys?", "Is there mutual trust?". The picture shows that there are two subgraphs which are determined by sex, suggesting that there is homophily with regarding this attribute. Moreover, the prevalence of asymmetric arcs reveals that there is no reciprocity, so that there is often one confidant and one speaker in each couple of actors.
Figure 1.2(b) represents the relationship based on living or being trained together for the 19 hijackers who were responsible for the tragic events of September 11, 2001 (Krebs,



(a) A support network



(b) A terrorist network

Figure 1.2: *Some network examples*

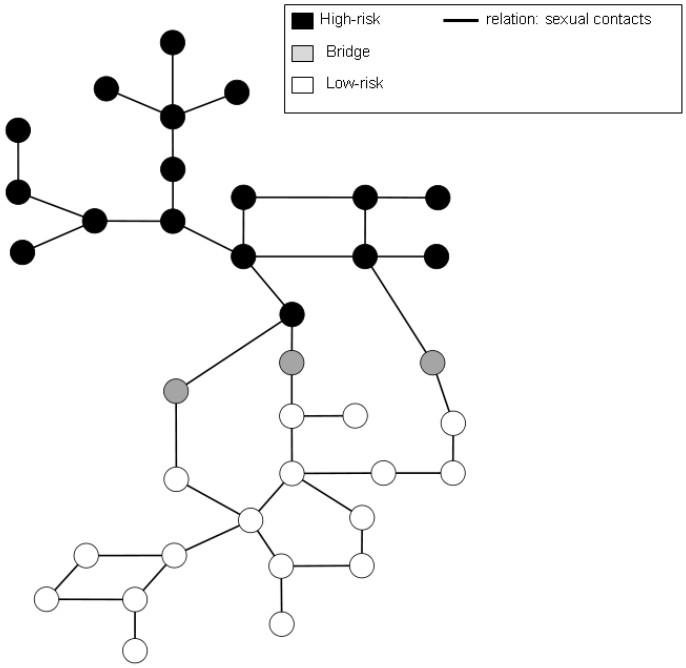


Figure 1.3: *A sexual network*

2002). In this context the key question could be "which actor should we remove to destabilize the network the most?". The answer is not simple since the network is surprisingly sparse and the terrorists, who were on the same plane, were distant from each other. For example, Hamza Alghamdi and Marwa Al-Shehhi, who were on the plane which crashed the WTC South, are 3-steps away from each other. This structure guarantees that if one of the cell members had been captured, the damage to the network would have been small because of the absence of direct contacts. It also explains Bin Laden's description of his strategy in one of his videotypes.

Finally, a hypothetical structure for a sexually transmitted infections (STDs) network is depicted in Figure 1.3. Diseases are transmitted from person to persons, so the nodes of the graph represent people, and the edges represent contacts. The graph reveals the presence a high-risk (black circles) and a low-risk population (white circles) linked by a few individuals who bridge the gap between them (grey circles). The real situation which is interpreted by this picture can be the following: the high-risk population is represented by intravenous drug users while the low-risk population by non intravenous drug users. An intravenous drug user who shares needles with his drug partners and who has sex with no intravenous drug users can play the role of cut-point and can be responsible for the spread of HIV. If someone wants to keep the risk low in the population represented by the white circles, one should remove the link between the grey and the white circles.

Social actors can also be groups of people. Padegett's studies about the marriage and

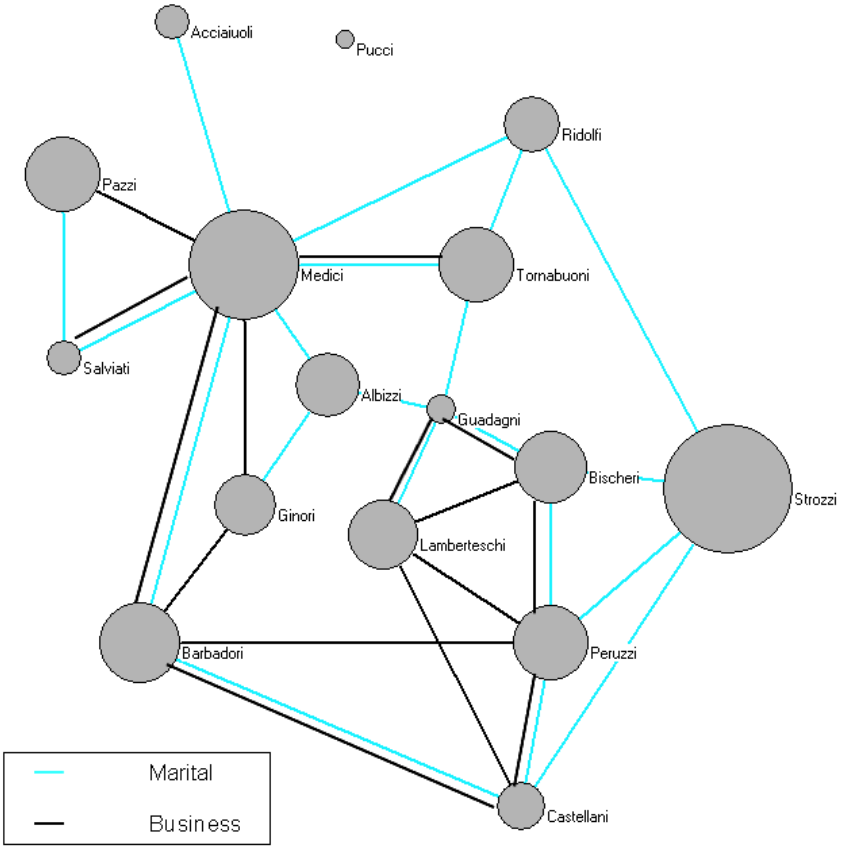


Figure 1.4: *Padgett's Florentine families, marital and business relations and wealth. The bigger the circle is, the greater the wealth is).*

business ties among 16 families in 15th century Florence (Padgett & Ansell, 1993) or the business relationships between companies (Mizruchi, 2003) are classical examples. Organizations and divisions within organizations play the role of social actors, if they are linked by the flows of money, other material resources or by the transfers of nonmaterial resources such as information. Countries and nations are also linked together by ties defined by trade, international cooperation, alliances and so on.
In this context questions of interest are "Who plays a key role in the network?", "How can alliances explain the power of a nation?", "Which directions follows the flows?", "Who is central, and who is peripheral in trade?".
For example, if we consider the marriage and business ties among some Florentine families and the wealth of each family (Figure 1.4), we can understand quite easily why the Medici succeeded in their rise to power and why not the Strozzi, which were the richest family at that time. Looking at the position of these two families in the network, it clearly appears that the Strozzi family does not show a central position, in fact it is mainly connected to families which are densely interconnected and share the same

status, so that each family had the same "right" to claim leadership. On the contrary the Medici family is the center of a simple sub-network system, in which the ties between families related to the Medici are connected to the others among themselves almost solely by the Medici. This privileged position allowed the Medici to control the others and to rise to power.
This example shows also that being more central has often some drawbacks. History tells us that the growing power of Medici family caused conspiracies against them. In 1478 Pazzi family organized a pot, but they did not succeed in their attempt. In a couple of hours the responsibles of the plot were arrested and hanged (Politianus & Adimari, 1769), confirming Medici family power.
At the end of this quick review of network applications it is necessary to give some examples of nonhuman actors, since they justify the general meaning of the term "social actors" and the popularity of network analysis in very different disciplines. Some lines before networks of public services were mentioned (Sen *et al.*, 2003; Brueckner, 2005; Guimera *et al.*, 2005). If we think about the railway or the subway paths, we can consider the stations as the social actors, and they are linked together if there is at least one train that stops in both of them.
In biology, networks are used to describe protein interactions (Bu *et al.*, 2003; Pellegrini *et al.*, 2004), the famous concept of "prey and predator" (Dunne *et al.*, 2002; Scotti *et al.*, 2007) and the neural networks (White *et al.*, 1986); in Physics to study the diffusion of innovation (Valente, 2005; Young, 2006), the network of citations between Web pages or academic papers (Garfield *et al.*, 1964; Hummon *et al.*, 1990; Small, 1999) and the virtual communities (Wellman *et al.*, 2003); in Economy to study the world economic system through trade or observable transactions in order to understand economic features of individual nations, such as the rate of economic development (Nemeth & Smith, 1985; Smith & White, 1992).
Networks can also be used in less serious contexts, for instance to describe the strategies designed by a team during a game or to reveal who "the heroes" are of a final game. In Figure 1.5 the 2010 Champions League final is analyzed according to a network perspective. Circles represent footballers (blue for the FC Internazionale Milano and read for the Bayern München) and arcs represent passes between them. It is clear that the teams adopted two different strategies: Inter prefers vertical lines, while Bayern horizontal ones. The former strategy works better since Inter can reach the Bayern area using few passes, and in light of the final result it was the best strategy.
The graph also shows the centrality of each footballer in terms of the proportion of received and sent passes with respect to the total number of passes for each team. The bigger the circle is, the higher the proportion of passes is. Other information are presented in the graph: the number of passes received and sent by each player through the width of the line (the thicker the line is, the higher the number of passes is) and the time played by each footballer through the color of the vertex (the darker the circle is , the higher the number of minutes played is).
These examples are barely a small amount of those present in the literature (Wasserman & Faust, 1994; Newman, 2003; Kolaczyk, 2009). They suggest that networks are every-

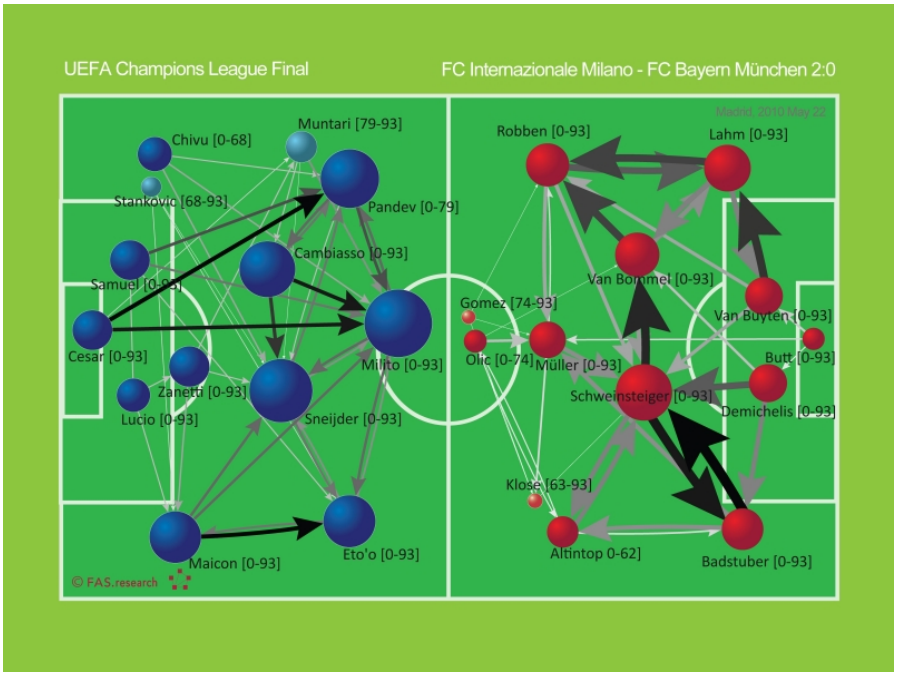


Figure 1.5: *The Champions League final 2010 (F.A.S. research)*

where and justify only in part the growing interest and curiosity in the last decades for social network analysis. To better understand why we study networks, it is necessary to go back to the early 1930's at the origin of social network analysis and briefly describe its evolution.

1.2 Why network analysis?

1.2.1 The origins of social network analysis

The terms "network analysis" are often preceded by the word "social", since the concept of network was first developed in a sociological and anthropological context. This also explains why a lot of network analysis studies involve human actors or human groups.

There have been three distinct research traditions that led to the birth of social network analysis: the British anthropologists of post World War II, the American sociometrists of the early 1930's and the American structural analysts of the 1960's. Even if the genesis of network started in different periods, most members of the three strands have not known the others'work in detail, so there was not much influence between them. This lack of influence is also justified by the variety of problems which the three traditions coped with. To solve them, at a certain moment, it was necessary to pay attention to the relations existing between human beings and to develop new methods for analysis.

In the early 1930's some German psychologists escaped from the Nazi regime to America

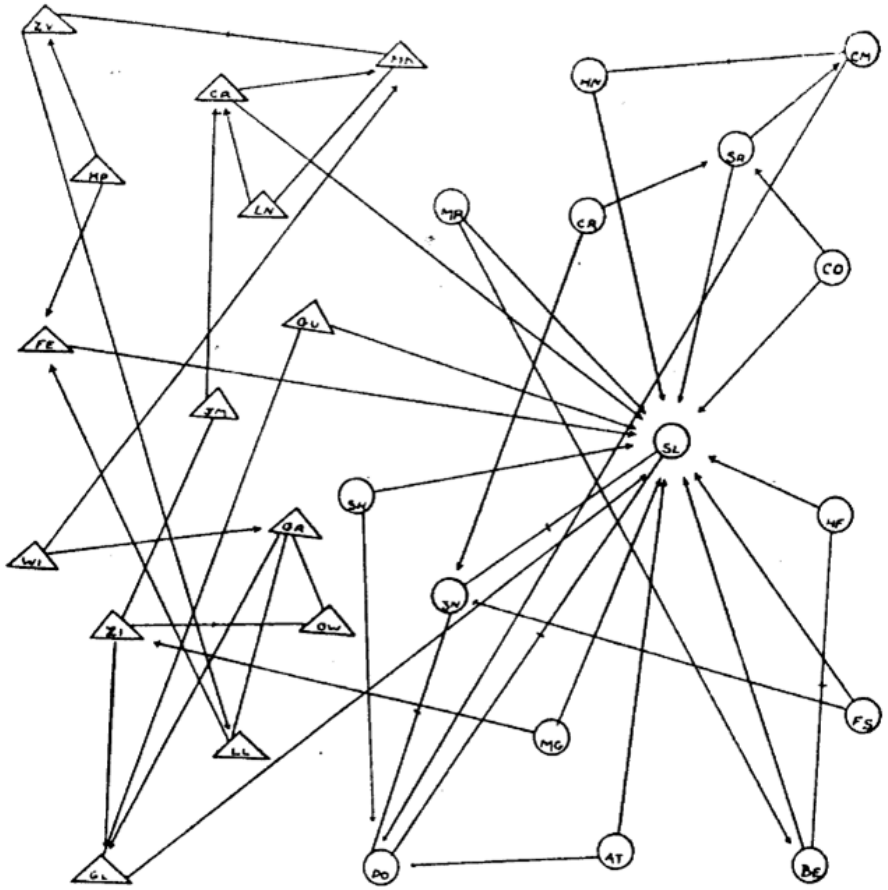


Figure 1.6: *An example of Moreno's sociogram: the relation is "studying together" and it was gathered on 29 classmates in a primary school*

and were influenced by the Gestalt theory[1]. They started to focus on the problems of sociometry and group dynamics. Their principal merit was to produce many technical advances with the methods of graph theory.
One of the major exponents of this group was Moreno (Moreno, 1953) who was interested in the role played by the relationships in a group both as limitations and opportunity for member actions. Specifically, he analyzed the correlation between the psychological well-being and a "social configuration", such as friendship, and he developed the so-called "sociogram" in order to represent them. The sociogram (Figure 1.6) is a two-dimensional pictures in which social actors are represented as points, and social relationships as lines linking the corresponding points. The sociogram is the ancestor of the modern graphical network representation. This graph allows the sociometrists to visualize the way

[1]The Gestalt theory was proposed by Khöler in 1925 and focused on the organized patterns through which thoughts and perception are structured

in which a piece of information can pass through an individual to another or how an individual can influence another one.
The usefulness of the sociogram was immediately understood by sociometrists and consequently it was applied to a broad range of situations. This led to its improvement. Twenty years after Moreno's idea, the structural analysts Cartwright (Cartwright & Zander, 1968) and Harary (Harary *et al.*, 1965) used the sociogram to study the relations living inside a group and to fill the gap between the micro and macro systems, introducing the concept of positive and negative relationships existing between triads, the bricks of complex social structure. Furthermore in the 1960's mathematicians such as Erdős and Reniy (Erdős & Rényi, 1960) began to study the distribution of random graphs.
Almost simultaneously at the University of Harvard Warner and Mayo (Mayo, 1977) studied the "informal relations" and discovered that large-scale systems contain cohesive sub-groups. Studying life in factories they found that there were two cohesive groups, the managerial elite and the worker group, whose existence is based on informal relationships. They also observed that a managerial elite who knew the influence of informal relationships in the worker groups on motivations could very successfully control worker behaviors. The finding of cohesive sub-groups gave rise to the concept of "cliques"[2] and subsequent work modified and extended this notion, allowing the network analyst to formalize the meaning of the words "social groups" used with slightly different significance by social scientists.
Thirty years later, the research group around Harrison White played a particular role in the development of social network analysis following two different strands. The formalists fixed their attention on the form of network patterns rather than on their content, neglecting the fact that similar structures can give rise to different behavioral consequences in different context. The structuralist works in the opposite direction looking for social structure according to two distinct approaches.
The first group considers the whole networks and the analysis aimed to describe the structure of the role relationships in a social system. Some sociological concept like social role, social positions and social status were now interpreted using the mathematical concept of structural equivalence. Networks help structural functionalist to find which people occupy the same positions inside the network or play the same key role, since they have identical ties to and from all the others in the network. Furthermore, for the first time in the history of the concept, the notion of social position was precisely defined. In this direction further steps were made by Heil and White (Heil & White, 1976), who came up with the idea of blockmodels. These are used to partition graphs (and matrices) so that social actors are grouped together into blocks that include structural equivalent nodes.
Considering again Figure 1.3. It is clear that the grey circles have the same social role and form a bridge between the two populations. They are also structurally equivalent since they share the same pattern of ties (one edge towards the high-risk population and

[2]A clique is a maximal complete subgraph of three or more nodes, all of which are directly connected to one another, with no other node in the network having direct ties to every member of the clique.

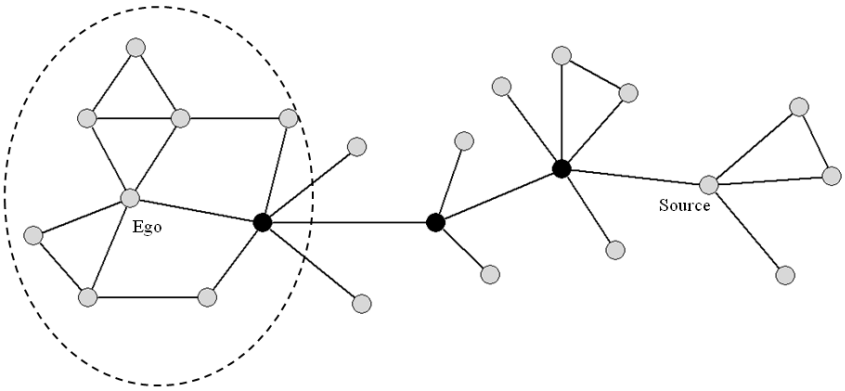


Figure 1.7: *Flows of job opportunities*

one edge towards the low-risk population). In this context knowing who the structural equivalent actors are helps to point out who is the responsible for the spread of HIV and how to apply prevention politics.
The second group views network in a different way, as Wellman suggested: "Rather than showing the universe as it is perceived by an outside observer, they provide Ptolemaic views of networks as they may be perceived by the individuals at their centers" (Wellman, 1988).
It is broadly agreed that individuals act to achieve goals, and that occupying a "strategic position" is relevant in order to reach their purpose. Granovetter (Granovetter, 1995) solved the problem of economists and sociologists in explaining how people acquire information about job opportunities through the social contacts that they have. Selecting a set of male workers in Boston who had changed their jobs in the last five years, it turned out that the primary channels through which individuals found a job opportunity were their informal personal contacts. This was not surprising, because economic and sociological theory had already established it, but Granovetter made a breakthrough, building the network that describes the flow of information of job opportunities and discovering something new.
Imagine that Ego wants to find a job (Figure 1.7). He starts asking friends and relatives, but this strategy is not very efficient because they share the same acquaintances, which means overlapping contacts and possession of the same knowledge about job opportunities (dotted line). For this reason it is through the relatively weak ties of less frequent contacts and of people working in different fields that new information comes (black dots). This is the now-famous argument of the " strength of weak ties", according to which "weak ties provide people with access to information and resources beyond those available in their own social circle; but strong ties have greater motivation to be of assistance and are typically more easily available" (Granovetter, 1983). Looking at Figure 1.7 it is clear that the actors between ego and the source of information play a key role in the flowing of news of job opportunities and their position is very important

in explaining how Ego get the job opportunity from the Source.
Both of these perspectives were also singled out by anthropologists at the University of Manchester. The major exponents were Barnes and Mitchell. Barnes (Barnes, 1954) was the first to introduce the "social network" concept in a rigorous and analytical way. Studying a community of Norwegian fishermen, Barnes realized the closed analogy among the nets used by the fishermen and the structure of the village community in which they lived. Mitchell deepened Barnes'idea of network (Mitchell, 1969), distinguishing between the "partial networks", i.e. the set of ties which links an individual to the others, and the "complete" network and underlining the multiple nature of the relations which allow the individual to exploit the network in order to achieve information, assets or resources.
Though American structuralist and British anthropologists had similar aims and adopted similar ways of looking at networks, they were animated by different interests. In fact, the former gave particular attention to the "centrality" of actors in the tie patterns of groups, while the latter were more interested in the density and the connectedness of large social networks. Furthermore, the first were more involved in the development of mathematical and statistical tools.
These three streams took the metaphor of network seriously and did not consider network analysis as a specialist technique within sociology, but they realized that social network analysis offered a new view of old problems, a new perspective in which the metaphors used by sociology could be formalized through graph theory, statistical and probability theory and algebraic models. These tools have become necessary since social network analysis solves analytical problems that are not standard, and this is a second reason for the birth of this scientific field.

1.2.2 Non standard data

The data analyzed by network analysis methods are quite different from the data typically encounter in "classical" social and behavioral science. For this reason the standard techniques used in classical social research were not able to analyze network data. Classic social research deals with *"attribute"* variables which refers to the characteristics and properties of the agents. Examples of attributes are attitudes, behaviors and opinions, but also properties of actors such as gender or educational level for individuals, number of employees for a factory, the wealth of a country, and so on. Since the attribute variables are analyzed under the hypothesis that they are gathered on a set of independent units, a huge amount of straightforward and complex statistical analyses can be applied to a range of research questions.
On the other hand network datasets can include two different kinds of variables: the "*composition*" variables and the "*structural*" (or *relational*) data. The former correspond to the attribute variables, while the latter measure contacts, ties and connections which link one actor to another, and for this reason they are collected on pairs of actors. Regarding the examples given in Paragraph 1.1, structural variables are: friendship ties between pupils, trade between nations, citations between web-pages, and so on.
The hypothesis of social agent's independence topples over when we consider relational

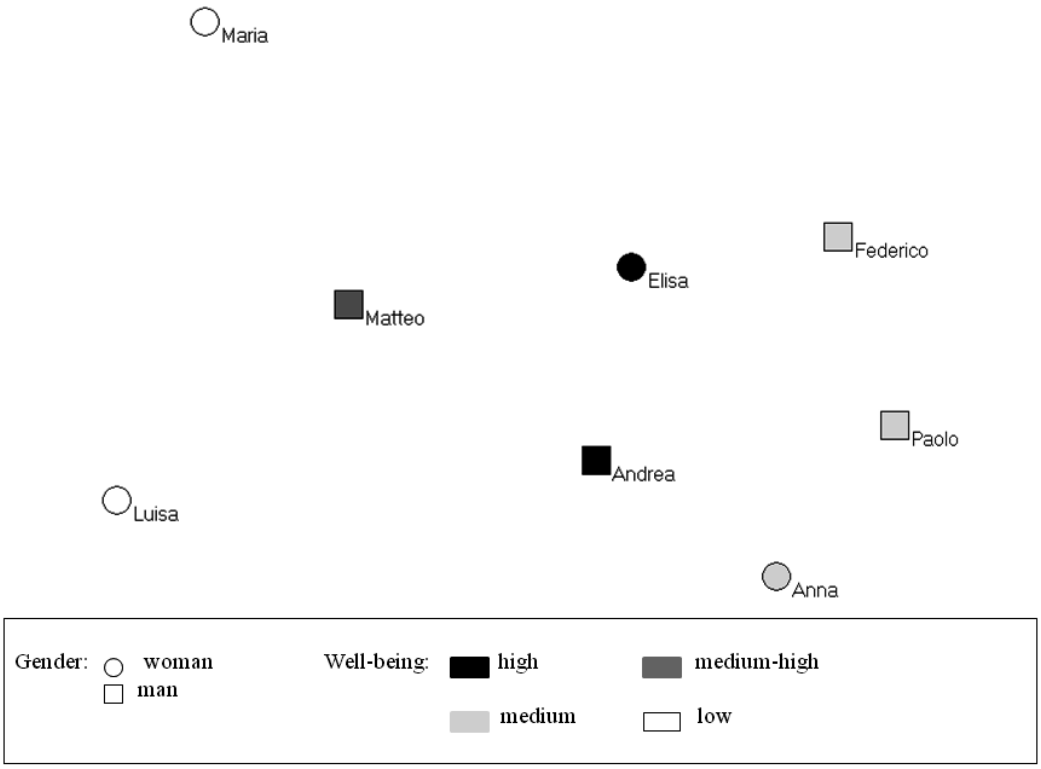
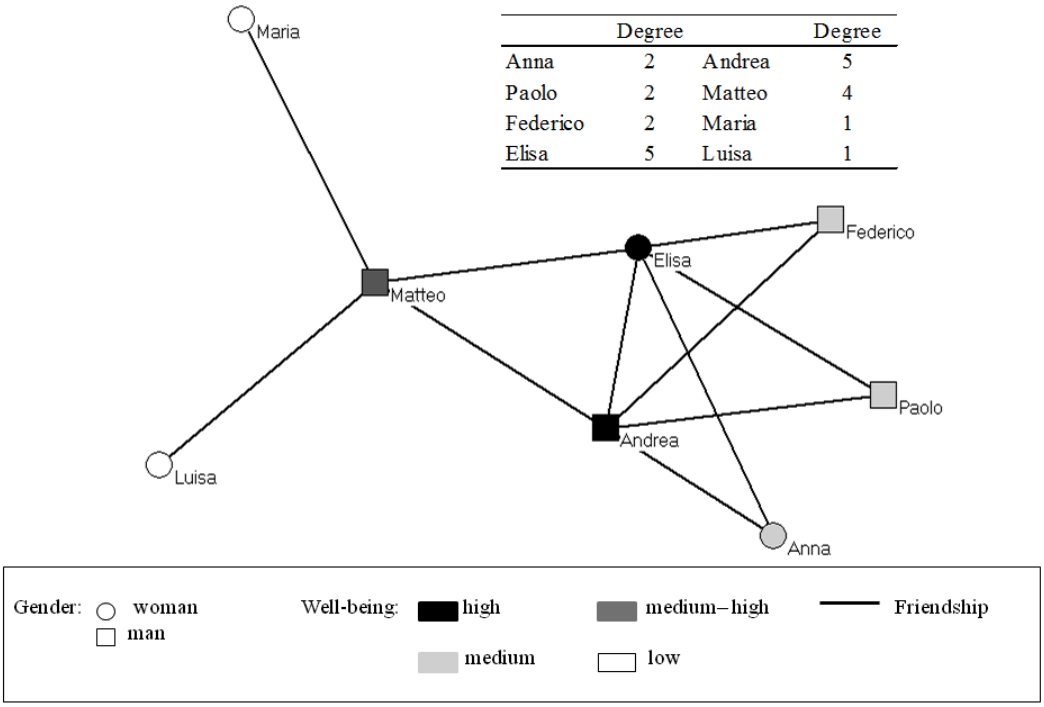


Figure 1.8: *Attribute data about a group of eight people*

data because structural variables measure dependency between units. According to social network theories, actors do not act independently from one another, but they influence each other through the relations which link them. Thus, while the mainstream sociological studies consider structures and processes as the result of personal attributes, with the aim to group individuals according to the characteristics they show, social network perspectives move the attention from individual to network, since the world is composed of networks and not groups. By ignoring the social-structural context within which social actors are involved, an attribute-based analysis loses much of the explanatory potential that a relational analysis can offer. For this reason social network analysis makes it necessary to take into account both types of data and the creation of measures and analysis tools capable of incorporating them.
In more detail let us consider a simple example. Imagine a group of 8 people and their well-being. Figure 1.8 represents the 8 individuals according to their well-being and gender (represented by the color and the shape of each vertex respectively). These are attribute data. Considering only attributes we can compute the distribution of the two composition variables and the associations between them or we can group the agents, but our source of information is too poor to say anything else.
If we imagine also collecting information about friendship, we obtain the relational data represented in Figure 1.9. Exploiting the structural data we can also try to answer some



	Degree		Degree
Anna	2	Andrea	5
Paolo	2	Matteo	4
Federico	2	Maria	1
Elisa	5	Luisa	1

Figure 1.9: *Friendship network within a group of eight people*

questions like: "having more friends gives more well-being?" or "people who have a higher level of well-being develop more friendship relationships?", "a focal actors (ego) well-being is dependent on the level of well-being of their friends" or "people select each other because they have the same level of well-being?".
It is clear that in order to reply to such questions it is necessary to consider the structure of dependency between actors and to jointly analyze the composition and relational variables. For example, the straightforward notion of "*degree*"[3] together with the well-being distribution is a good way to observe that there is a significant association between degree and well-being: people who have more friends also have a high level of well-being. It is necessary to point out that this observation does not describe a causal relationship between having more friends and showing a high level of well-being, so at this point it is not yet possible to answer the first two questions. A longitudinal approach is needed in order to decide the causality between the well-being and the friendship ties formation. This means that we should have observed the network evolution through time, a topic recently considered by social network analysts.
Just to give an idea, and considering this time the last two questions, we can imagine that we have observed the network from two different time points t_0 and t_1 and we fix our attention on the relation between Elisa and Andrea.
Assuming that only the attributes of these two agents or the ties between them are

[3]The degree is the number of lines incident with a node in the graph.

allowed to change between the two time points, different situations can arise. Two of them are depicted in Figure 1.10. Case a) suggests that at time t_0 Elisa and Andrea are friends but the boy feels a lower level of wellness with respect to the girl. Then at time t_1 both the actors present the same level of well-being. This is a case of influence which suggests that Andreas well-being is dependent on the level of well-being of his friends.
Instead, case b) suggests that people select each other because they have the same level of well-being. In fact, at time t_0 Andrea and Elisa are not friends but have the same level of wellness. Then a friendship relationship is born between the two observational points. This fact is known in literature as the "homophily selection".
To investigate these patterns in a network some statistical tools have been developed, such as specific models for longitudinal data which will be treated in more detail in Chapter 3 and which are the object of this project.

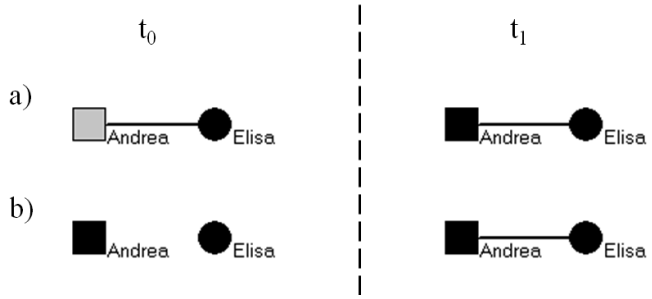


Figure 1.10: *Selection and influence*

At this point it is clear that the dependence between actors and their actions and the particular nature of relational data have required both descriptive and inferential statistical tools to describe the structure and to test theoretical propositions about networks respectively. In fact, the exploitation of structural information to test theories prevents the use of multiple regression, t-tests, canonical correlations and so forth.
For this reason Wasserman and Faust (Wasserman & Faust, 1994) suggested that "social network analysis may be viewed as a broadening or generalization of standard data analytic techniques and applied statistics which usually focus on observational units and their characteristics."
At the end of this paragraph some words should be spent about the collection of relevant and reliable data through suitable network study design. If the aim of the analysis is to compare one network to another, it is necessary to dispose of the complete network data on all social ties linking the units of the populations. In this case the problem consists in defining the boundaries of the population and in gathering all the ties. The issue of missing data (Robins *et al.*, 2004; Huisman, 2009) is still an open question and the standard techniques seem to solve it only in parts since once again relational and compositional variables should be considered jointly.
Other matters arise when studying egocentric networks; in fact, in this case it is necessary to define a sample design. Conventional random sampling procedures can be used to collect egocentric network data and generalize results about the networks surrounding

units to a large population, but the literature on network sampling has also developed (Frank, 1977, 1988; Wasserman & Faust, 1994).
These few lines suggest that sampling and missing data theory should also be adjusted to social network analysis (Marsden, 1990; Carrington *et al.*, 2005; Knoke & Yang, 2008; Wasserman & Faust, 1994) like all the statistical tools and their development has contributed to the spread and the growing interest in social network analysis and to the cooperation between sociologists, graphical theorist, mathematicians and statisticians.

1.3 Notations and basilar concepts

1.3.1 The variety of network typology

A social network is a structure composed of a set of actors, some of whose members are connected by a set of one or more relations. As suggested by Paragraph 1.1, graphs constitute a nice way to represent network data. A graph is a pair $(\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is a finite set of nodes, usually represented by points, and $\mathcal{L}$ is a set of lines, which is a subset of the set $\mathcal{N} \times \mathcal{N}$ of ordered pairs of distinct vertexes. Nodes correspond to social actors, while lines describe relations between them.
Different types of networks, and consequently of graphs, are defined according to types of relations. On one side there are relations that are directional, that is, they are directed from an actor to another, such as friendship or trade from one country to another. Formally considering an ordered pair of actor $< i, j >$, node i is the initial node (or the sender) of the line and node j is the terminal node (or the receiver). Directed relations are represented using arrows according to the direction of the relation, and the presence of a tie from i to j does not imply that there is a tie between j and i. A network described by directed relations is called *directed network* and the corresponding graph is called *digraph*. Figure 1.2(a) and Figure 1.5 are examples of digraphs.
On the other side there are relations that are undirected; that is, it is not possible to distinguish between a line from i to j and a line from j to i, such as co-authorship, collaboration or living near each other. A network based on a non directional relation is called *undirected network*. There is only one measurement to be made for each pair, rather than two measurements as is in the case of a directional relation. Figure 1.2(b) and Figure 1.3 are examples of undirected graphs.
These concepts can be generalized to multirelational networks, described by more then one relation, like Padgett's Florentine example or the transportation system of a city. In the first, marriage and business are relations; in the second, the different lines defines the multirelational aspects.
Beyond the direction or the multiple presence of relations, there are other properties of ties that identify different types of networks. If we think about the collaboration networks between academics, where link is identified by "having written a paper with", we can imagine that two authors who have published 10 articles together have a stronger relation than two authors who have only published 1 article together. This suggests that we can distinguish ties according to their strength, interpreted by the number of

co-authored articles. When it is possible to measure ties'intensity, lines of the graph are weighted and they are represented by different width, the thicker the line is the stronger the relation is. Figure 1.5 represents a weighted network, where each line is characterized by the number of passages between two football players. This kind of graph (or network) is called "*valued*".
Connections between social actors can also be positive or negative. Alliances and hostilities among nations are two different kinds of relations. The former can be considered in a positive way since alliances mean cooperation and peace, while hostilities have a negative meaning because it causes tensions and wars. In a multirelational network we can represent them by assigning a "+" sign to edges or arcs if the relation is positive and a "-" sign if it is negative. A Graph (and consequently a network) of this type is called a "*signed*" graph (network).
Networks can also be classified according to the set of actors, which defines the *mode* of a network. If a network includes only a set of actors, it is called a "*one-mode*" network. The examples showed in the last pages are all one-mode networks.
A network may also include two sets of actors. A famous dataset collected by Davis and colleagues in the 1930s, contains the observed attendance at 14 social events by 18 Southern women (Davis, 1979). Women and attendance constitutes two different sets of actors on which the relation "participate to an event" is defined. This is an example of a *two-mode* network. Other examples can be: people and journals as sets of actors and reading as a relation or parliamentarians and decrees laws as sets of actors and a positive vote as a relation. Relations are usually defined from a set of actors to another, but sometimes ties are also allowed for homogeneous pairs, that is the sender and the receiver actor belong to the same group.
The combination of modes and relational properties leads to different kinds of networks. In the next chapters the focus is on one mode network (neither signed nor valued) because they have been studied for a long time and are widespread. Furthermore, they frequently represent the starting point for new developments, since they are the more straightforward type of networks. For this reason notations and subsequent concepts will be introduced referring to this kind of networks.
It is not difficult to imagine that the graphical representation of networks is not useful when someone tries to depict large networks, that is networks with a lot of nodes. Even if there are a lot of algorithms implemented in specialized drawing programs such as Pajek, Ucinet or Visone (Borgatti *et al.*, 2002; Huisman & van Duijn, 2005; Nooy *et al.*, 2005; Brandes & Wagner, 2003), graph representation suffers from the limited space of sheets and screens even for few hundreds nodes, where it is quite easy to observe many crossings of lines.
To solve this problem an algebraic representation of network relations can be used. Matrices are able to express all the quantitative information embedded in a graph, and at the same time they enable a much larger set of analysis than possible with the corresponding visual representation.
Basically, different types of matrix representation exist according to the role played by rows and columns, but the adjacency matrix (or sociomatrix), commonly denoted by X,

	1	2	3	4	5	6
1	-	1	0	0	1	1
2	1	-	0	0	1	0
3	0	0	-	1	0	0
4	0	0	1	-	1	0
5	1	1	0	1	-	0
6	1	0	0	0	0	-

	1	2	3	4	5	6
1	-	1	0	0	1	1
2	1	-	0	0	1	1
3	0	0	-	1	0	0
4	0	0	0	-	0	0
5	0	0	0	1	-	0
6	0	0	0	0	0	-

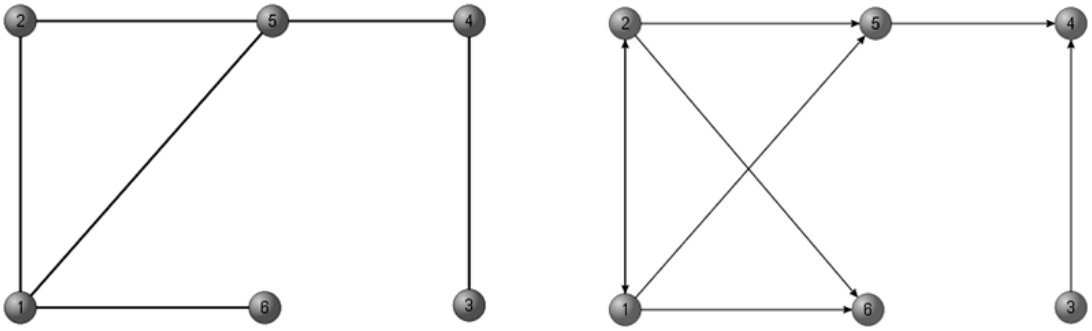


Figure 1.11: *Adjacency matrices for an undirected and a directed graph*

is more useful. If there are g social actors in a network, the adjacency matrix is a $g \times g$ matrix, in which there is a row and a column for each agent. The generic entry of X takes value 1 if there is a link between actors i and actors j and 0, otherwise:

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ is related to } j \\ \\ 0 & \text{otherwise} \end{cases}$$

and it is denoted by a small letter. Usually the entries on the diagonal are null since relations are not often reflexive.
Figure 1.11 shows that when the network is undirected, the sociomatrix is symmetric, while if ties are directed this symmetry is not guaranteed because in the first case we are not able to distinguish between ties from i to j, and so $x_{ij} = x_{ji}$ while in the second case it can also happen that $x_{ij} \neq x_{ji}$.
Regarding agent attributes the usual cases for variable matrix is used. Each row represents an actor and each column an attribute, so that each entry of this matrix contains the value assumed by the j^{th} variable on the i^{th} actor. The matrix will be denoted by Z, while the cell by z_{ij}. Actor attributes can be continues, such as the wealth in the Padgett's Florentine network in Figure 1.4 or discrete such as gender in the support network in Figure 1.2(a).

1.3.2 Measures for network description

Graphical and matrix representation contains all the available information about the considered network. As usual it is necessary to summarize the network's characteristics in order to describe its features. Different indexes were proposed.
If we consider the entire network, the first properties are the number of actors and lines and density. Density is a measure that describes the presence of ties, and it has the advantage of allowing comparison among different networks with respect to the number of ties.
Let A and B be two networks gathered observing the same relation on two different sets of actors and 40 and 60 be the number of ties recorded in A and B respectively. The risk is to conclude that B is denser than A since there are more lines, but this may be only "apparent". In fact, if we suppose that network A included 30 nodes and network B 50 nodes, the perspective changes. If one considers the possible number of ties in the two networks, which is given by $g \times (g-1)$, and he/she computes the ratio, denoted by Δ, between the lines observed and the maximum possible number of lines, he/she obtains:

$$\Delta_A = \frac{40}{30 \times 29} = 0.046 \qquad \Delta_B = \frac{60}{50 \times 49} = 0.024$$

In this way he/she discovers that network A is more dense than network B. Therefore if we want to compare two or more networks with respect to the number of lines, it is necessary to account for the maximum possible number of lines. This is the idea behind the density index which represents the number of lines in a simple network as a proportion of the maximum possible numbers of lines. Denoting by ℓ the observed number of lines, density is defined by the ratio:

$$\Delta = \frac{\ell}{g(g-1)}$$

and can be interpreted as the proportion of possible lines present in a network. We can observe that density is inversely related to network size: the larger the social network is, the lower the density is, since the number of possible lines increases rapidly with the number of vertices, whereas the number of ties which each person can maintain is limited.
If we look at actors, different properties can be described. The most relevant one regards the number of lines incident to a node, defined in the literature as "*degree*". If the network is oriented one can distinguish between "indegree" and "outdegree", which are defined as the number of arcs an actor receives and the number of arcs the actor sends respectively. The sum of indegree and outdegree is equal to the degree.
Indegree and outdegree assume values between 0, if no ties are sent or received, and $g-1$, if all the possible ties are sent or received. The combination of indegree and outdegree allows to classify nodes. We speak about "isolate" node if indegree and outdegree are both equal to 0, "initial" node if indegree is null and "terminal" node if outdegree is 0. We will also say that two nodes that are connected by a line are "adjacent".
Indegree and outdegree measure centrality of an actor in terms of the involvement degree

of an actor in the network. The higher the degree of an actor is, the more he/she is involved in relations, and therefore he/she is central. There are other measures of actors'centrality because there are different definitions of centrality, according to the collected relation.
The first one is based on the reachability of a vertex within a network and is based on the concept of distance. The *distance* between two vertexes is the length of the the shortest path (geodesic) that connects the vertexes. In a directed network the shortest path must take into account the direction of the lines in the path. This kind of centrality suggests that an actor is central if it is easily reachable, and it is measured through the so-called "closeness-centrality" index which is computed dividing the number of adjacent vertexes to a node by the sum of all distances between the vertex and all the others[4].
A different concept is expressed by the "betweeness-centrality" which relies on the fact that a person is more central if he or she is more important as an intermediary in the communication network. In this context "the centrality of a person depends on the extent to which he or she is needed as a link in the chains of contacts that facilitate the spread of information within the network" (Nooy *et al.*, 2005). The "betweeness-centrality" index corresponds to the proportion of all geodesics between pairs of other vertexes that include this vertex[5]. The indexes considered in the previous lines (except density) focus on actor properties. There are other measures that focus on the pattern of lines.
Sometimes a network is cut up into pieces, which are called *components*. Components are subsets of actors among whom there are relatively strong or weak ties and describe the presence of cohesive subgroups, all of which are based on the way in which vertexes are interconnected. An analysis of components allows to test if structurally delineated subgroups differ with respect to social characteristics.
Just to clarify Figure 1.2(a) can be considered. It clearly appears that the network can be split into five components, three of them are isolated nodes, and the other two are two subsets of ten and seven nodes respectively. The bigger components are clearly identified by gender, which plays a determinant role in explaining the pattern of support relations.
Components and cohesive subgroups are the results of particular configurations presented in a network. Some of these will be the object of the next paragraph, but since there is a wide variety of configurations only the fundamental ones in the next few pages.

1.3.3 Configurations in networks

The structure of a network is the result of combination of configurations which describes the interaction between actors.
The simplest manner of interconnection between actors is represented by dyads. A dyad "consists of a pair of actors and the possible tie(s) between them" (Wasserman & Faust, 1994). Let i and j be two vertexes. In an oriented network different kinds of dyads can

[4]If the graph is disconnected, (i.e. some nodes are not reachable by the others) this kind of measure cannot be computed since the distance between two non-adjacent nodes is equal to infinity

[5]Details about centrality indexes can be found in Wasserman & Faust (1994), Nooy *et al.* (2005), Hanneman & Riddle (2005).

be defined (Figure 1.12).

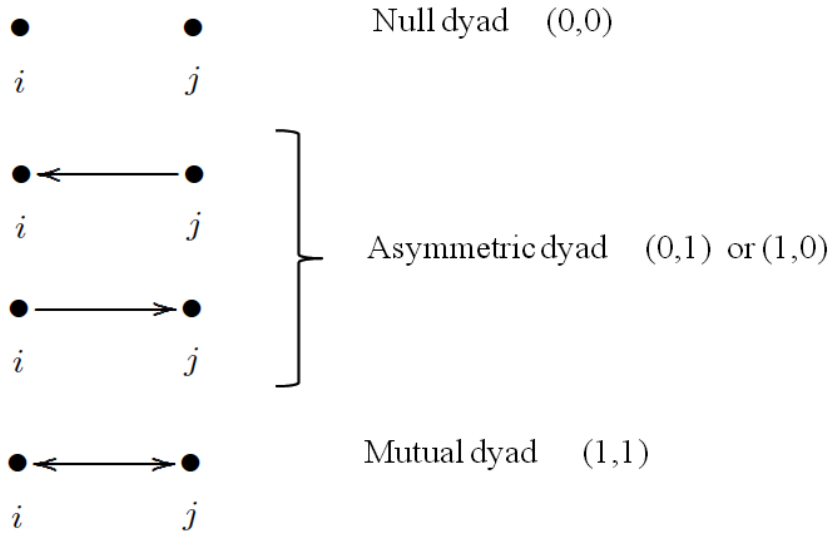


Figure 1.12: *Possible configurations for a dyad*

A mutual relationship between i and j exists when both the ties from i to j and from j to i exist. A reciprocal dyad is apparent in an adjacent matrix X when both cells (i, j) and (j, i) are equal to 1, so that $D_{ij} = (x_{ij}, x_{ji}) = (1, 1)$. This means that we can observe a mutual dyad when both actors in a pair of nodes choose the other in the relation.
On the contrary if there are no ties between i and j, the dyad is "*null*" and the (i, j) and (j, i) off-diagonal cells of X are equal to 0, i.e. $D_{ij} = (x_{ij}, x_{ji}) = (0, 0)$.
The last case occurs when there is one tie from i to j or from j to i. The dyad is "*asymmetric*" and in terms of the cells of the adjacency matrix it is represented by $D_{ij} = (x_{ij}, x_{ji}) = (0, 1)$ or $D_{ij} = (x_{ij}, x_{ji}) = (1, 0)$. The asymmetric dyad is often interpreted as an intermediate state of relations that are starving for a more stable equilibrium of a mutual or null dyad. A different interpretation suggests that asymmetries indicate unequal resources within a dyad.
To describe dyad configurations within a network different approaches can be applied. The first one is the so-called "dyad census" which corresponds to the frequency distribution of dyad configurations.
The second approach is based on a set of indexes which describes the proportion of mutual dyads within the network. Mutual relations are often the objects of interest since a mutual dyad represents the smallest cohesive subgroup, and it is a state of equilibrium between two actors. Reciprocity has also been recognized as one of the most important forces in structuring social relations (Goulder, 1960) and has been widely studied in the context of exchange-networks. A detailed discussion about indexes for studying reciprocity is given by Wasserman & Faust (1994).
More complex configurations arise if three actors and the possible relations among them

are considered. These types of configurations are called "triads". Figure 1.13 represents

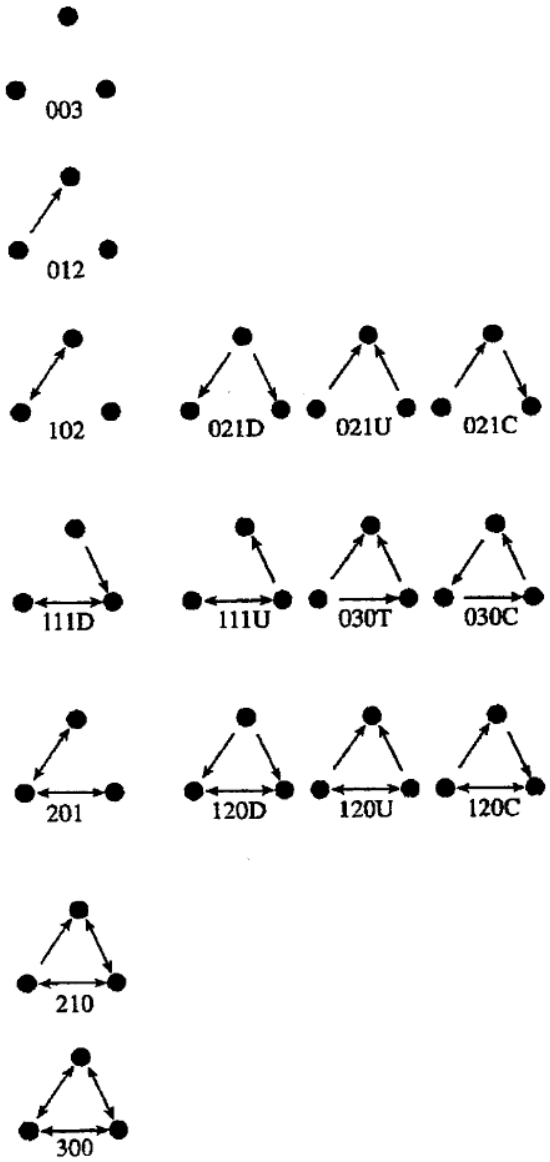


Figure 1.13: *Possible configurations for a triad*

the possible sixteen triad configurations. Triads are coded by a string of three numbers and eventually one letter. Regarding numbers, they represent in order the number of mutual, asymmetric and null dyad involved in a triad, while the character, if present, is "D" for down, "U" for up, "T" for transitive and "C" for cyclic according to the directions of the arrows.
One of the relevant configurations for triads is the triplet coded by 030T, which is called *transitive triad*. Let i, j and h be three actors within a network. Transitivity is defined

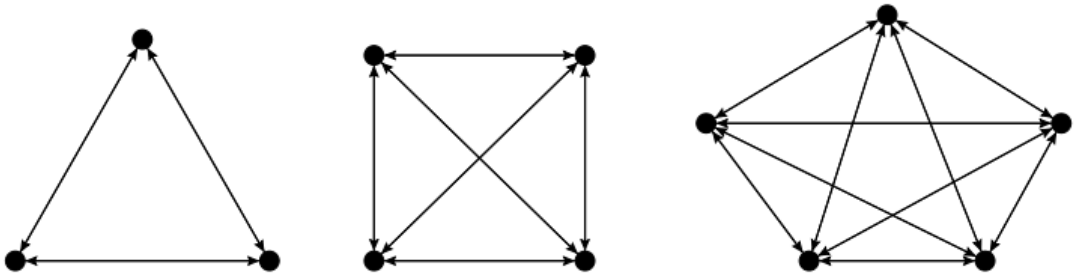


Figure 1.14: *Examples of cliques*

as the tendency for a tie from i to h to occur at greater than chance frequencies if there are ties from i to j and from j to h and in terms of a well-known relation such as friendship can be described by the sentence "the friend of my friend is also my friend". Holland and Leinhardt (Holland & Leinhardt, 1972) provide strong statistical evidence that transitivity is a very important structural tendency in social networks because it is one of the breaks of network structure together with reciprocity.
Some remarks should be spent about reciprocity and transitivity. The first observation concerns some advanced statistical tools which can establish if the presence of mutual dyads and transitive triads is significant with respect to the other dyad and triad configurations. For this aim statistical tests were performed (Katz & Powell, 1955). The second remark is related to the computation of triad census which is computationally burdensome. This problem has been very relevant in the past and has led to the search of efficient algorithms (Moody, 1998). Now the problem seems to be solved, in fact there are some network programs which are able to compute the triad census in an efficient way (Batagelj & Mrvar, 2001; Schank & Wagner, 2005).
Finally, some words will be spent about more complex structures that involve more than three actors and show particular dependence patterns.
Previously cliques were mentioned. A clique is a maximal connected subgraph of three or more actors, i.e. a subset of nodes in which each actor is adjacent to all the other members of the clique. The restriction that the clique contains at least three nodes is necessary so that a mutual dyad is not considered to be a clique.
One can interpret a clique as a collection of actors which all "choose" each other. Some examples of cliques are depicted in Figure 1.14, and they show that all the directed ties between each couple of actors are included. Cliques are structures that describe cohesive groups, and they were particularly important in the study of affective relations such as friendships. In this context cohesive subgroups are characterized by a subset of nodes of which all friendship relations are mutual. Empirical studies proved that cliques are structures that characterize large networks, i.e. networks with hundred of nodes, while they are rarely observed in small networks.
Cliques are also very strict concepts because the absence of a single tie prevents a subgraph from being a clique. For this reason the concept was extended to "n-clique", a

subgroup of actors so that the geodesic distance among subgroup members is less or equal to n.
More complex structures are described in the literature but they are not the focus of this work, since in the next chapters reciprocity and transitivity will play the main role. A detailed review of complex structure can be found in Wasserman & Faust (1994) and Scott (1988).

Chapter 2

Statistical models for social network data

Most social network methods are descriptive, attempting to represent underlying structures which govern ties formation or to characterize network properties through algebraic computation or a large variety of indexes. Statistical models for social network analysis move beyond description to explain the presence of ties as a function of individual and graph level explanatory factors.

2.1 Why modelling social networks and what is different?

There are many well-known techniques that measure properties of a network and a set of actors in order to describe and reveal network features which might be important for particular research questions. Given the profusion of descriptive techniques one can wonder why network statistical models are necessary. The answer is not different from the reasons that justify the moving from descriptive to inferential statistics.
First of all, network models are able to jointly consider the regularities underlying the process and the random parts related to unknown factors. Since social behaviors and network tie formation are complex phenomena, characterized by both variability and regularities, statistical models are a key for combining these elements in a parsimonious way and to better understand the uncertainty associated with the observed network.
The thrift of a statistical model is also useful in order to explain complex network data structures, such as the longitudinal observation of a single network or multiple observations of a same relation on more than one group. If we have a network observed at different time points, descriptive statistics can be computed for each network, but a longitudinal model is able to enclose all the information concerning the evolution of the network in a few parameters. Likewise, if we observed a same relation on different groups, we can apply a multilevel model in order to combine the different information derived from each set of actors.
Once the parameters of the model are estimated, inference about the parameters can be performed in order to establish if specific structures, arisen from a previous descriptive

analysis, are significant (more details about the kind of inference, necessary for network statistical model, are given later). Significance of certain structures helps to develop hypothesis about the social process that might produce the significant structural properties revealed by descriptive techniques.
Furthermore, models allow to unbundle structures and to determine the net contribution of each effect. Just to clarify, "clustering in networks might emerge from endogenous (self-organizing) structural effects (e.g. structural balance), or through node level effects (e.g. homophily). To decide between the two alternatives requires a model that incorporates both effects and then assesses the relative contribution of it" (Robins *et al.*, 2007).
Finally, network models allow to point out how localized social processes and structures interact to give rise to the complete network patterns. This means that they are able to fill the gap between micro and macro levels, and they allow researchers to answer longstanding questions in social network analysis.
Non standard models should be applied in the network context to fulfill all these aims. In the world of complex networks, it is dependence that matters and assumptions involving traditional statistical independence, although helpful for fitting models, are likely to be empirically inadequate. In statistics one usually deals with samples, and he/she wants to extrapolate information about a certain characteristic of a population. This requires a process of generalizing sample results to the population according to precise statistical rules which satisfy particular requirements. One of the widespread assumptions underlying this operation is the independence assumption between the units of the population. In social network analysis the perspective is totally different. In fact, one studies a group of individuals (the set of actors) and he/she is interested in finding results concerning structures present in this specific group. As previously pointed out, the assumption of independence among actors is not reasonable because it cannot deal with the dependence structures described in the previous lines.
The result is that, on one hand, there are samples and independence hypothesis; while on the other hand, there is a complete set of actors whose observational units depend from each other through the relationships existing among them. Since statistical models are not able to treat dependence structures[1], they cannot be applied to study network data.
These two approaches give rise to two different kinds of inference: a "sampling (or designed)-based" and a "model-based" inference. Design-based inference is applied in statistical analysis when random selection of observational units or random allocation of units to different experimental treatment are the only assumption, i.e. the sampling-design is the only source of randomness explicitly accounted for in estimation and inference.
In model-based inference assumptions external to the sampling design are required, so that the starting point is a probabilistic model which is assumed to be a correct representation of the data. Consequently, the finite population quantities of interest are random with respect to the model.
The usual statistical approaches are based on sampling-based inference, instead network

[1]Transitivity is the first dependence structure that statistical models cannot deal with

analysis considers model based inference.
Following the notation in Chapter 1, we suppose to have observed a set of g actors on which a dichotomous relation is defined. The result is represented as a digraph or as an adjacency matrix X whose generic cell takes value 1, if there is a tie between actor i and actor j, and 0 otherwise. The dependent variable of the model is the tie from i to j indicated by X_{ij}:

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ is related to } j \\ \\ 0 & \text{otherwise} \end{cases}$$

The presence of a tie between i and j can be explained on the basis of two elements measured at different levels. A first set of explanatory variables collects *covariates* and can be a function of individual actors or of directed or undirected pairs. The covariates, which are a function of individual actors, are denoted by the terms "actor-based covariates", and they correspond to the classical statistical variables, such as gender, age, welfare, ethinicity and so forth. The covariates which are function of directed or undirected pairs are known as "dyad-based covariates", and they can be collected for each combination of actors or they can be derived from actor attributes as well, such as similarity indicators. If we considered a network defined by the relation "working together", being in the same office is an example of dyadic covariates which is collected for each combination of pairs of actors, while homophily with respect to gender or seniority can be derived from actor attributes.
The second set of explanatory variables is related to *patterns of ties* in the network, and it constitutes the major difficulty in modeling social networks. These variables describe the dependence structure between ties, and examples are given by reciprocity or transitivity. As already mentioned, reciprocity is dependence of couples of actors, while transitivity between triples of actors. Other dependence structures are given by: the number of incoming or outgoing relations of the same actor, which is a dependence form within each row and each column of the adjacency matrix, respectively; popularity, which describes the tendency that choices lead to more choices; balance, which expresses a preference for others who make the same choices as the actor himself.
Since the dependent variable of a network model is dichotomous, one can imagine that the straightforward model for describing the presence or the absence of a tie between two actors is a logistic regression. A logistic regression model expresses the logit transformation of the probability of observing a tie between actors i and j as a linear function of the covariates:

$$logit\left[P\left(X_{ij}=1\right)\right] = \beta_0 + \beta_1 Z_1 + \cdots + \beta_k Z_k$$

where $Z_1, Z_2, \cdots, Z_k$ are explanatory variables. In more detail, the probability of observing a tie between actors i and j can be computed in the following way:

$$\pi(x_{ij}) = P\left(X_{ij}=1\right) = \frac{e^{\beta_0+\beta_1 Z_1+\cdots+\beta_k Z_k}}{1+e^{\beta_0+\beta_1 Z_1+\cdots+\beta_k Z_k}}$$

The logit transformation is the logarithm of the odds ratio related to $\pi(x_{ij})$, i.e.

$$logit\left[P\left(X_{ij}=1\right)\right]=ln\left[\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right]$$

The problem related to this model is that it neglects the dependence structure of network data and consequently, from a statistical point of view, the parameters estimates are inefficients, and their standard errors are unreliable. For this reason, a logistic regression is not adequate, and different models should be defined. In the following paragraphs a short review of network models will be given.

2.2 Models for cross-selectional data

2.2.1 A first model for undirected graphs

In 1951 Solomonoff and Rapoport made a first attempt to model a network and to define a distribution for random nets (Solomonoff & Rapoport, 1951). They focused their attention on the entire collection of undirected graphs, defined on a set of g actors, and they assumed that each edge within a network has a probability p to be present. In other words they defined a Bernoulli model for random nets, which was subsequently described in a rigorous way by Erdős and Rényj (Erdős & Rényi, 1960).
According to these two Hungarian mathematicians, the probability distributions express the probability of observing a particular graph. The probabilities are defined on the set of all graphs of g nodes, which has cardinality equal to $2^{\frac{g(g-1)}{2}}$, since each of the $\frac{g(g-1)}{2}$ edges may or not may be present. Let $X_{ij}=\{0,1\}$ be a random variable which indicates the presence of an edge between actors i and j. It is assumed that it is distributed as a Bernoulli random variable of parameter p:

$$X_{ij}=\begin{cases}1 & \text{with probability } p\\ \\ 0 & \text{with probability } \textit{1-p}\end{cases}$$

Since edges are supposed to be independent, the probability of each undirected random graph G is:

$$P(G)=p^{\ell}(1-p)^{\frac{g(g-1)}{2}-\ell}$$

where ℓ is the number of edges within the graph. The set of all possible graphs defined on g actors and their corresponding probabilities represents the so called *random graph distribution.*
If one is interested only in the distribution of graphs defined on a set of g actors and exactly with ℓ edges, he/she should modify the previous formula. Assuming that each graph with these characteristics is equiprobable, it is not difficult to draw the following

distribution[2]:

$$P(= G \mid \mathcal{L} = \ell) = \frac{\ell!(g-\ell)!}{g!}$$

which may be regarded as a random graph distribution conditional on the number ℓ of edges.
The random graph distribution also had a main role because it allows the definition of the first ever proposed algorithm for the formation of random graphs (Erdős & Rényi, 1959).
After Erdős and Rényj's proposal not much interest was directed to statistical models. In fact, the attention towards statistical modeling for relational data has increased since the beginning of the 1980's when Holland and Leinhardt proposed an exponential family of probability distributions for directed graphs (Holland & Leinhardt, 1981). Since then a wide range of models has been proposed.

2.2.2 The p_1 model

The p_1 model was developed by Holland and Leinhardt in 1981 as a reaction to the paucity of available statistical tools for analyzing social networks (Holland & Leinhardt, 1981). Model p_1 is based on the assumption of dyads independence, which allows to easily determine the probability distribution of the entire network X. Having defined the probability of occurrence of a mutual dyad, an asymmetric dyad and a null dyad as:

$$P[D_{ij} = (0,0)] = n_{ij} = n_{ji} \qquad P[D_{ij} = (1,0)] = a_{ij}$$

$$P[D_{ij} = (1,1))] = m_{ij} = m_{ji} \quad P[D_{ij} = (0,1)] = a_{ji}$$

where $m_{ij} + n_{ij} + a_{ij} + a_{ji} = 1 \quad \forall\, i \neq j$, the probability distribution of the entire network X is computed as the product of dyad probabilities:

$$P(X = x) = \prod_{i<j} m_{ij}^{X_{ij}X_{ji}} \prod_{i \neq j} a_{ij}^{X_{ij}(1-X_{ji})} \prod_{j<i} n_{ij}^{(1-X_{ij})(1-X_{ji})} \tag{2.1}$$

Some re-arrangements allow to express the probability distribution in equation (2.1) in an exponential form:

$$P(X = x) = K \cdot exp\left[\sum_{i<j} \rho_{ij}^{X_{ij}X_{ji}} + \sum_{i,j} \mu_{ij} X_{ij}\right] \tag{2.2}$$

where for all $i \neq j$:

- $\rho_{ij} = ln\frac{m_{ij}n_{ij}}{a_{ij}a_{ji}}$ is an index of reciprocity

[2]The number of graphs on g actors with ℓ edges is given by the combination of g elements in class ℓ: $\frac{g!}{\ell!(g-\ell)!}$. Since the graph are equiprobable the probability of observing each graph is the inverse of the ℓ-combination previously described.

- $\mu_{ij} = ln\frac{a_{ij}}{n_{ij}}$ is a log-odds measure of the probability distribution of an asymmetric dyad between i and j
- $K = \prod_{i<j} [1/(1 + exp(\mu_{ij} + exp\mu_{ji} + exp(\rho_{ij} + \mu_{ij} + \mu_{ji})))]$ is a normalizing constant, which assures that the probabilities sum to 1.

Assuming that the reciprocity parameter is constant over all dyads , i.e. $\rho_{ij} = \rho$ for all $i \neq j$, and the parameter μ_{ij} depends additively on the propensity of arcs to emanate from node i and the propensity for arcs to have node j as a target, i.e. $\mu_{ij} = \mu + \alpha_i + \beta_j$ for all $i \neq j$, the resulting model can be written as:

$$P(X = x) = K \cdot exp\left[\rho \sum_{i,j} X_{ij}X_{ji} + \mu X_{++} + \sum_i \alpha_i X_{i+} + \sum_i \beta_i X_{+i}\right] \tag{2.3}$$

where a subscript "+" indicates a sum over the i-th row ($i+$) or the i-th column ($+i$) of the adjacency matrix, while the double subscript "++" represents the sum of all the cells of the adjacency matrix.
The parameters α_i and β_i must satisfy the two constrains that their sum with respect to i should be 0, i.e. $\sum_{i=1}^{g} \alpha_i = 0$ and $\sum_{i=1}^{g} \beta_i = 0$.
It clearly appears that the distributions in equations (2.2) and (2.3) belong to the exponential family. This suggests that the statistics X_{++}, X_{+i}, X_{i+} and $\sum_{i,j} X_{ij}X_{ji}$ are sufficient statistics for the vector of parameters $(\mu, \rho, \alpha_1, \cdots, \alpha_n, \beta_1, \cdots, \beta_n)$.
The parameters ρ and μ can be interpreted as uniform *reciprocity* and *density* parameters, and the node - dependent parameters α_i and β_i reflects respectively *expansiveness* and *attractiveness*, of each node i. In particular: if α_i is large and positive, node i will tend to have a relatively high out-degree, or it tends to send a huge amount of ties; if β_i is large and positive, node i will tend to have a relatively high in-degree, or it tends to receive a lot of ties; if μ is large and positive the network is dense, and a consistent number of ties is present; if ρ is higher, there is a tendency towards reciprocation so that X_{ij} and X_{ji} tend to be more alike.
The model in (2.3) can be re-expressed in terms of the probability distribution related to each dyad[3]

$$P\left[(X_{ij} = x_1, X_{ji} = x_2)\right] = \frac{exp\left[x_1(\mu + \alpha_i + \beta_j) + x_2(\mu + \alpha_j + \beta_i) + x_1x_2\rho\right]}{k_{ij}} \tag{2.4}$$

[3]This distribution is obtained computing the probability of each mutual dyad using the model in (2.3). For instance, if we consider a mutual dyad were $X_{ij} = 1$ and $X_{ji} = 1$ the sufficient statistics in the model take values: $\sum_{i,j} X_{ij}X_{ji} = 1$, $X_{++} = 2$, $X_{i+} = 1$, $X_{+i} = 1$, $X_{i,+} = 1$ and $X_{j+} = 1$, so that the probability of a mutual dyad corresponds to:

$$P\left[(X_{ij} = 1, X_{ji} = 1)\right] = \frac{(\rho + 2\mu + \alpha_i + \alpha_j + \beta_i + \beta_j)}{k_{ij}}$$

Reasoning in a similar way, the probability associated to each dyad can be determined, and then the distribution in (2.4) follows.

where $k_{ij} = 1 + exp(\mu + \alpha_i + \beta_j) + exp(\mu + \alpha_j + \beta_i) + exp(2\mu + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho)$ and $x_1, x_2 = 0, 1$.
Equation (2.4) allows to show that the p_1 model can be regarded as a bivariate logistic regression model for the dyad. In fact, using the logit as a link function, one can write:

$$logit\left[X_{ij} = 1 | X_{ji} = x_2\right] = ln\left[\frac{P(X_{ij} = 1 | X_{ji} = x_2)}{P(X_{ij} = 0 | X_{ji} = x_2)}\right] = \mu + \alpha_i + \beta_j + x_2\rho$$

and this makes more clear the role played by each parameter. It follows that the estimation of the p_1 model does not require a particular estimation method, in fact the parameters can be estimated using the same methods for estimating log-linear model (Wasserman & Faust, 1994).
In this context the drawback relies on the total number of parameters, which must be estimated. Having $2g$ parameters[4] makes the interpretation of the model difficult, but this is not the real problem of the p_1 model. The big issues rely on the basilar assumptions that dyads are independent, a hypothesis which is not reliable in a network context, and on the fact that covariates are not included.
The first issue was overcome developing a series of latent variable model, in which the hypothesis of dyad independence is relaxed assuming that dyads are independent conditionally on latent (i.e. unobserved) variables representing some potential underlying structure. One of this model is the p_2 model. Several other important latent variable models have been proposed for some particular networks. The idea at the base is that actors have latent locations in a space (euclidean or ultra metric), and given these latent positions ties are conditionally independent (Hoff *et al.*, 2002; Schweinberger & Snijders, 2003; Shortreed *et al.*, 2006).
Regarding the second issue Fiendberg and Wasserman (Fienberg & Wasserman, 1981) included covariates within the p_1 model, defining cluster on the basis of categorical attributes and assuming that expansiveness and attractiveness parameters are constant within the same group.

2.2.3 The p_2 model

The p_2 model (Duijn *et al.*, 2004; Lazega & Van Duijn, 1997) represents a natural extension of the p_1 model and "can be regarded as a random effects version, or also a multilevel version, of the p_1 model" (Snijders, 2009b). In this model the sender and receiver effects α_i and β_i constitutes the latent random variables, and they are explained by both "actor-based" and "dyad-based" covariates.
The form of the p_2 model does not apparently differ from equation (2.4), since this formula is still valid. The difference relies on the fact that now parameters are expressed as a linear combination of covariates and random effects, suggesting that density, reciprocity, attractiveness and expansiveness depend on actor attributes.

[4]There are two parameters for the reciprocity and density effects and two parameters for each actor, for a total number of $2 + 2g$ parameters. Since the parameters α_i and β_j are subjected to constrains, the number of independent parameters is equal to $2g$.

In more details, let $Z_1, \cdots, Z_k$ be a set of explanatory actor-based variables. Attractiveness and expansiveness parameters are regressed with respect to these variables (different explanatory variables may be used to model productivity and expansiveness):

$$\alpha_i = \gamma_0 + \gamma_1 Z_{1ij} + \gamma_2 Z_{2ij} + \cdots + \gamma_k Z_{kij} + U_i$$

$$\beta_i = \delta_0 + \delta_1 Z_{1ij} + \delta_2 Z_{2ij} + \cdots + \delta_k Z_{kij} + V_i$$

where U_i and V_i are random variables which represent the unexplained parts of the sender and receiver effect of actor i.
It is assumed that U_i and V_i are normally distributed with expectation 0 and variance σ_U^2 and σ_V^2, respectively. Attractiveness and expansiveness parameters of a same node i are correlated, i.e. $Cov(U_i, V_i) = \sigma_{UV}$, while independence is assumed for parameters of different actors, so that $Cov(V_i, V_j) = Cov(U_i, U_j) = Cov(U_i, V_j) = 0$.
Since only the regression parameters should be estimated, a p_2 model is usually more parsimonious than the p_1 model. The reduction of parameters enables relaxation of the assumption that density and reciprocity are constant over dyads, so that also these parameters are linearly related to dyadic attributes which will be denoted by $W_1, \cdots, W_k$:

$$\mu_{ij} = \mu + \upsilon_1 W_1 + \cdots + \upsilon_k W_k$$

$$\rho_{ij} = \rho + \lambda_1 W_1 + \cdots + \lambda_k W_k$$

Both parameter equations contain a constant part, represented by μ and ρ and a part that varies across dyads, i.e. the regression terms. It is also assumed that the reciprocity parameter is constant within dyads, that is $\rho_{ij} = \rho_{ji}$.
The parameters of the model are: μ and ρ, the regression coefficients $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\upsilon}$, $\boldsymbol{\lambda}$[5], the variances σ_U^2 and σ_V^2 and the covariance σ_{UV}. The interpretation of the parameters is similar to that of the p_1 model, except for the variances and covariance. In particular, the higher the variance is, the higher the importance of unknown characteristics is to explain the popularity and attractiveness of an actor, i.e. covariates has a poor explanatory capability. Regarding the covariance, it explains the correlation between sending and receiving ties, i.e. the structure of "give and take" in the tie formation process.
In order to estimate the parameters of the p_2 model, it is convenient to use MCMC methods because the maximum likelihood estimation of the random effects requires the solution of intractable integrals. Details can be found in Duijn *et al.* (2004) and Zijlstra & van Duijn (2003).
There is a multilevel version of the p_2 model which deals with multiple networks (Zijlstra *et al.*, 2006) that are the result of observing a same relation on different sets of actors. The multilevel p_2 model can be regarded as a three-level random effects model, where the first level is formed by tie observation, the second level by the actors and the third level by the network itself.

[5]The following vector notation is used: vector are denoted by a bold character, but when they are the argument of a function the bold character is neglected

To cope with this multilevel structure, random effects for the fixed p_2 regression parameters at the actor level and at the network level are included in the model. The random effects allow to account for differences between the networks considered. Even for the p_2 multilevel model it is assumed that random effects are normally distributed with zero means and covariance matrix Σ, and that the random effects at the actor level are independent from the random effects at the network level.
The p_2 model solves the problem of including covariates into the model, but it represents structural network effects only to a very limited extent. In fact, like the model p_1, it includes parameters only for density and reciprocity, so that transitivity effects or more complex dependence structures cannot be modeled. The crucial insight to solve this problem was given by Frank and Strauss who defined the well-known notion of "conditional independence" (Frank & Strauss, 1986).

2.2.4 The p^* model or ERGM

The hypothesis of dyad Independence of p_1 and p_2 models represents a severe limitation in network modeling as pointed out in the previous paragraph. Consequently, developments which allow to relax this condition have become more and more important. In 1986 Frank and Strauss proposed a Markov dependence idea (Frank & Strauss, 1986): two network tie variables are supposed to be conditionally independent, given the values of all other network tie variables, unless they have a node in common. The concept of conditional dependence states that a tie between actors i and j is conditionally independent of ties involving any pair of actors k and l, but it could be conditionally dependent on any other ties that involves i or j.
Conditional dependence is represented by the well-known concept of *dependence graph*, that indicates which ties are conditionally dependent. "The dependence structure of a random graph is simply a graph whose nodes are all possible relational ties in the original and whose ties specify which ties in the relation are conditionally dependent, given the remaining relational ties" (Wasserman & Pattison, 1996).
According to the conditional dependence assumption Frank and Strauss defined a particular Markov graph distribution which was generalized into the well-known p^* model also called Exponential Random Graph Model (ERGM) (Wasserman & Pattison, 1996; Robins *et al.*, 2007; Wasserman & Robins, 2005). Application of the Hammersley-Clifford Theorem (Besag, 1974) leads to the following probability distribution of this kind of model:

$$P(X = x) = \frac{exp\left[\sum_k \theta_k s_k(x)\right]}{\kappa(\theta)} = \frac{exp\left[\theta' s(x)\right]}{\kappa(\theta)} \tag{2.5}$$

where $s(x)$ is the vector of statistics corresponding to the k configurations (or dependence structure), $\boldsymbol{\theta}$ is a k-dimensional vector of parameters and $\kappa(\theta)$ is a normalizing constant which ensures that the probabilities sum to 1. With appropriate homogeneity constrains, the model parameters represent the contributions of the k network configuration to the probability of an observed network.

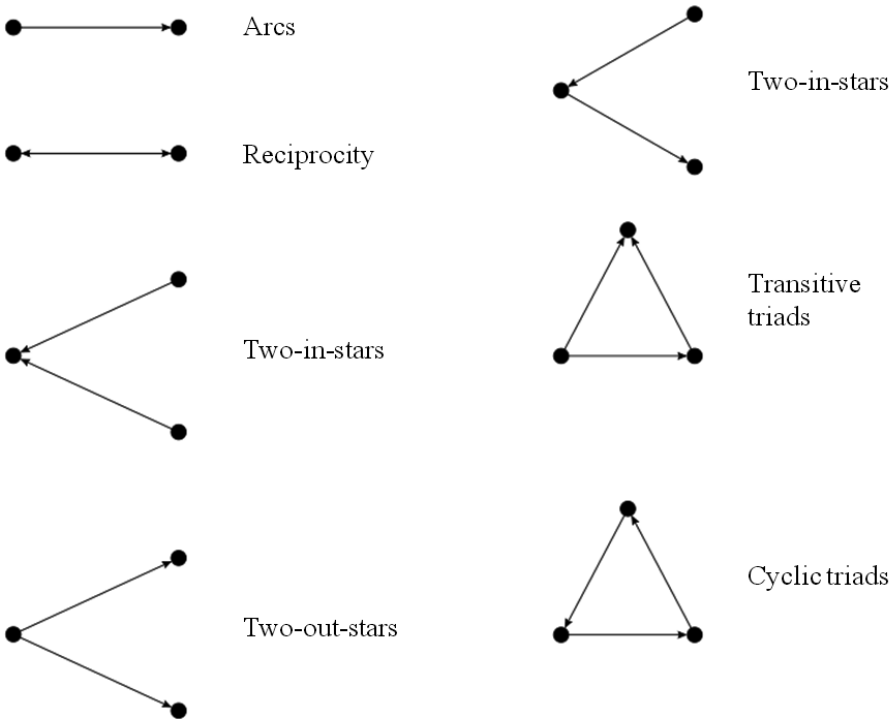


Figure 2.1: *Some examples of configurations*

Again the probability distribution belongs to the exponential family, so that $s_k(x)$ are sufficient statistics. The statistics are network configurations that involved Markov graph dependence, and the summation in (2.5) is over these dependence structures. Some examples of these configurations are depicted in Figure 2.1.
The p^* model presented some difficulties in the estimation process, because of the presence of the normalizing constant $\kappa(\theta)$, which makes standard likelihood techniques not immediately applicable. Since $\kappa(\theta)$ is the sum over all the possible networks of $\sum_k \theta_k s_k(x)$, it is impossible to compute $\kappa(\theta)$ by direct enumeration. Just to clarify, if we consider for purposes an undirected graph and 6 actor, there are $6 \times 5/2 = 15$ possible dyads, each of them can take the value 0 or 1. So there are $2^{15} = 32768$ graphs which should be enumerated to compute $\kappa(\theta)$. This is very hard work, and since networks usually include dozens of nodes this is not feasible. For this reason, Frank and Strauss suggested that if the number of nodes is more than six, it is not possible to evaluate the normalizing constant by direct enumeration.
To solve the estimation problem Strauss and Ikeda proposed to use a pseudo-likelihood approach and showed that pseudo-likelihood estimation can be conducted using standard logistic regression procedures (Strauss & Ikeda, 1990). Wasserman and Faust described how to set up a data set for this estimation procedure and listed possible configurations which can be included in the model (Wasserman & Pattison, 1996).
Focusing the attention on a tie between the actors i and j and conditioning with respect

to the rest of the graph, denoted by X_{ij}^c, the pseudo-likelihood of the p^* model is given by:

$$ln\left[\frac{P(X_{ij}=1|X_{ij}^c)}{P(X_{ij}=0|X_{ij}^c)}\right] = ln\left[\frac{\frac{exp\left[\theta' s(x_{ij}^+)\right]}{\kappa(\theta)}}{\frac{exp\left[\theta' s(x_{ij}^-)\right]}{\kappa(\theta)}}\right] = \boldsymbol{\theta}'\left[s(x_{ij}^+) - s(x_{ij}^-)\right] \tag{2.6}$$

where x_{ij}^+ and x_{ij}^- indicates the network x when the tie between i and j is present and absent respectively. Equation (2.6) suggests that computing the ratio between the two probabilities, the normalizing constant vanishes; and that maximizing the pseudo-likelihood is equivalent to maximizing the likelihood function for the fit of a logistic regression model, so that any statistical package for logistic regression analysis can be used to estimate the vector of parameters $\boldsymbol{\theta}$. The problem related to the pseudo-likelihood approach is given by the fact that the properties of the resulting estimator for exponential graph models are unknown and maximum pseudolikelihood estimators are not admissible for a squared error loss function, since it is not a function of the complete sufficient statistic $s_k(x)$. Furthermore, the model in (2.5) is not a standard logistic regression because the dependencies within the data and the usual methods to test the model fit are not reliable.
For this reason social network analysts look for an alternative approach, based on an MCMC methods to approximate the maximum likelihood estimator (Corander *et al.*, 1998; Snijders, 2002). The method presents a degeneracy problem when the model is not well specified. In this case the distribution of the exponential random graphs is concentrated on the full graph (all ties are present) or on the empty graph (no ties are present), which may lead to convergence problems of estimation algorithms and a poor fit to empirical data.
To solve the problem new specification for ERG models were proposed (Snijders *et al.*, 2006), which allow to relax the Markov dependence assumption.
The idea is to consider a partial conditional dependence which can be incorporated into ERGM, thanks to the Hammersely-Clifford. This partial dependence is described by the so called *social circuit dependence*, according to which two possible network ties are conditionally dependent if their observation leads to a 4-cycle. The corresponding configuration is depicted in Figure 2.2. According to Markov dependence, we will expect that two distinct possible edges (i, j) and (k, l) are conditionally independent. In particular circumstances, it can happen that if a node i is related to j and node k is related to l, then the presence of a tie between i and k can make the presence of a tie between j and l more likely (and vice versa), so that they are conditionally dependent. For instance, in a business, cooperation between two bosses may lead their employees to work together or in families the presence of a friendship between two children may increase the chances that their mothers become friend. Extensions to the p^* model towards multirelational networks and valued networks were developed by Wasserman and Faust (Pattison & Wasserman, 1999; Robins *et al.*, 1999).

To sum up briefly, we saw that the first model proposed was based on the assumption in which all edges are conditionally independent. This hypothesis was relaxed suppos-

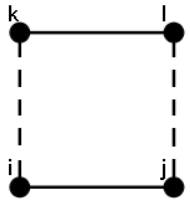


Figure 2.2: *Social circuit dependence*

ing that dyads are independent and was included in the p_1 and p_2 models. Finally, a more complex dependence structure, the Markov dependence, was considered and then generalized to social circuit dependence and included in the p^* model.
The models described in the previous paragraphs can be estimated using specific softwares, such as SIENA (Huisman & Van Duijn, 2003), PNet (Wang *et al.*, 2006; Harrigan, 2007) and statnet (Handcock *et al.*, 2008), which also include hypotheses testing for the parameters and for goodness of fit. The principle behind goodness of fit relies on the fact that a good estimate should be able to reproduce network structures. For this reason goodness of fit procedures simulate the network according to the estimated values for the parameters, and they compare observed network configurations to the simulated one through an adequate test. If configurations are well-reproduced, then the model is a good description of the considered network.

2.3 Models for longitudinal data

In the last decade the interest of social networks analysts has been extended from cross-selectional data, deriving from a single observation of one or more networks at a certain time point, to longitudinal data, collected through the observation of a network over time.
We suppose to focus on a set of g actors and on a relationship defined on them. These two elements defined a network which we observed at M (with M at least equal to 2) discrete time points, indicated by $t_1, \cdots, t_M$. Thus, the M network observations are denoted by $X(t_1), \cdots, X(t_M)$. The aim is to describe network evolution and the leading force that characterize and determine it. In order to do that it is assumed that network dynamics are continuous time processes, even though we observed the network at discrete time points. Consequently, there is an unobserved network evolution going on between the observational time points.
Doreian and Stockman (Doreian & Stokman, 1997) classified studies focusing on network longitudinal data into some categories: studies that predict attributes from structural information; descriptive studies of network change, and studies where network change is seen as a transition in network structures between time points.
A growing emphasis in network literature on the need to model network change has also been registered. Different longitudinal models were proposed. A first strand of models for network evolution implicitly includes temporal processes and their aim is to

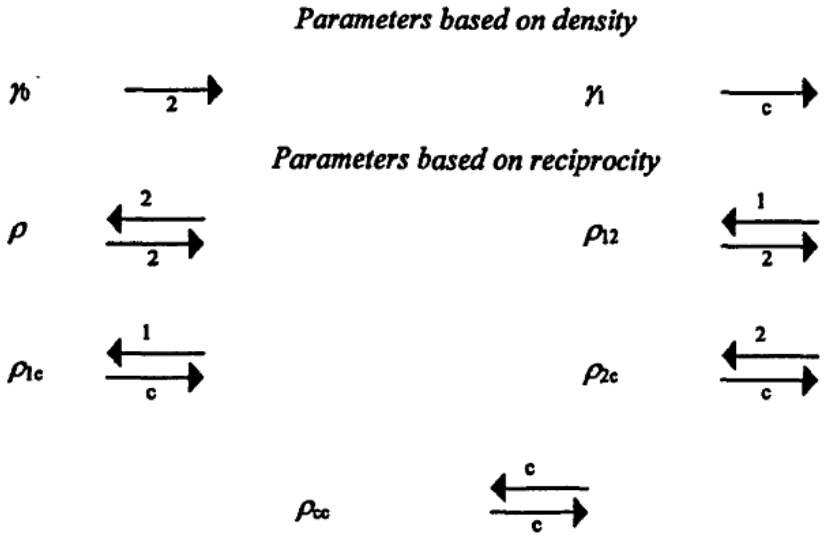


Figure 2.3: *Parameters and associated configurations based on density and reciprocity for the p^* model for network evolution. Ties are numbered depending on whether they occur at time 1 or 2; "c" refers to a constant tie, where both time 1 and time 2 ties are present.(Robins & Pattison, 2001)*

model social influence and social selection. Some of these models are based on the p^* probability distribution (Robins *et al.*, 2001b,a), some others on autocorrelation models (Doreian *et al.*, 1984; Marsden & Friedkin, 1993; Leenders, 2002). In Stockman and Doreian's classification these kinds of models belong to the first category.
A second strand of longitudinal network models explicitly includes time. Among them we can distinguish between two main classes of longitudinal network models: the log-linear models (Wasserman & Iacobucci, 1988) and those based on continuous-time Markov chains (Wasserman, 1980a).
Regarding log-linear model, in 2000 Robins and Pattinson extended the p^* model to longitudinal network data (Robins & Pattison, 2001). The starting point consists in observing that p^* implicitly includes the notion of temporal process. In fact, configurations, such as reciprocity and transitivity or more complex dependence structures, observed at a certain time point, are the result of network evolution through time. "Indeed a p^* model can be seen as the equilibrium distribution of what we might term a network "birth and death" process, in which, at any point time, a network link is added or removed with probability determined by p^* model" (Robins & Pattison, 2001).
The formulation of the model is equal to that reported in equation (2.5). To model time, the usual configurations of the p^* model are classified according to the moment at which each tie occurred (Figure 2.3), and each class constitutes a new configuration. Consequently, new parameters and new statistics are associated to new configurations, and the resulting model explicitly includes time through the statistics.
The drawback of this procedure relies on the fact that the model is less parsimonious,

and it shows the same estimation issues of the p^* model for cross-selectional data.
The second class of longitudinal models identifies the underlying process as continuous-time Markov chains, a process well-known in statistical analysis. There are many books that explain in details continuous-time Markov chains (Ross, 1996; Norris, 1998; Durrett & Durrett, 1999), but to understand the following models only some notions are necessary.
Let $\{X(t), t \geq 0\}$ be a continuous-time stochastic process taking values in a countable set I. Markov chains are based on the the assumptions that the future depends on the present and on the past only through the present. This implies that for any possible outcomes $i, j \in I$ and for any pair of time points $s < t$:

$$P\left(X(t) = j | X(s) = i, X(u) = x(u), 0 \leq u < s\right) = P\left(X(t) = j | X(s) = i\right) \tag{2.7}$$

If this probability does not depend on the time points s and t, but only on the time elapsed between them, i.e. $(t - s)$, then the process is said to have stationary (or homogeneous) transition distributions and the *rate of jump* can be defined. Rates of jump (rate of change) are the derivatives at 0 of the transition probabilities:

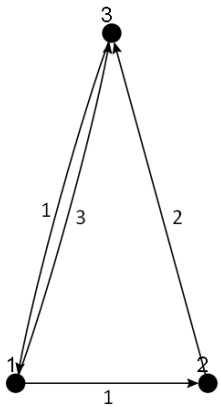
$$\begin{aligned} q(i,j) &= \lim_{dt \to 0} \frac{P(X(t+dt)=j | X(t)=i)}{dt} \quad \text{for } i \neq j \\ q(i,i) &= \lim_{dt \to 0} \frac{P(X(t+dt)=j | X(t)=j)}{dt} \end{aligned} \tag{2.8}$$

and can be interpreted as the rate at which the process changes from i to j in the short time interval $(t, t+dt)$. The rates of change for each couple of outcomes can be collected in a matrix $Q = (q(i,j) : i, j \in I)$ that is called "intensity matrix". Q must satisfy the following conditions:
i) $0 \leq -q(i,i) < \infty$
ii) $q(i,j) \geq 0$ *for all* $i \neq j$
iii) $\sum_{j \in I} q(i,j) = 0$ *for all* i
Knowing the rate of jump from i to j allows also to determine the rate of leaving state i just summing up the rate of jumps from i to j for all $j \neq i$.
A convenient way to present the data for a continuous-time Markov chain is with a diagram and each diagram corresponds to a unique Q-matrix (Figure 2.4) Thus, the rate of going from i to j is the value attached to the (i,j) arrow on the diagram.
Denoting by $P(t)$ the matrix which includes the conditional probabilities given in (2.7), the following relation can be proved:

$$P(t) = e^{tQ}$$

If a Markov chain has a stationary transition distribution and each one of its states communicates with the other states, i.e. from each state i it is possible to reach any other state j, then the limiting distribution of the process $\{X(t), t \geq 0\}$ is unique, and it is defined as:

$$\lim_{t \to \infty} P(X(t) = j | X(t) = i) = \pi$$



$$Q = \begin{bmatrix} -4 & 1 & 3 \\ 0 & -2 & 2 \\ 1 & 0 & -1 \end{bmatrix}$$

Figure 2.4: *A simple example of continuous time Markov chain*

In a social network context the arbitrary finite outcome space I corresponds to the set of all directed graphs $\mathcal{X}$, so that the adjacency matrix $x(t)$ represents the state of the process at time t. The rate of change is then the rate of transition between two different network configurations.
An additional assumption was proposed by Holland and Leinhardt known as the "conditional change independence" (Holland & Leinhardt, 1977). They suggested that the arcs of a digraphs are statistically independent, and this implies that the probability of any two arcs changing simultaneously is 0. This means that at each time point just one tie is allowed to change.
Different kinds of models are defined according to the two hypotheses previously described (networks evolve according to continuous-time Markov chains and the conditional change dependence). These models differ for the dependence hypotheses between arcs underlying the model, as with cross-selectional data. The simplest model to describe network evolution is based on the unrealistic assumption of independent arcs. The model is useful as a baseline since it allows explicit calculations of the transition probability matrix $P(t)$ from the infinitesimal transition rate matrix Q. Snijders et al. provided a short description of this model and the method of estimation of its parameters (Snijders & Van Duijn, 1997; Snijders, 2005). More interesting models are briefly described in the next two paragraphs.

2.3.1 The reciprocity model

In 1980 Wasserman proposed two simple models for network evolution (Wasserman, 1980a,b). The first one is the reciprocity model, which assumes that each dyad is independent of all the other dyads, and the process $\{D_{ij}(t) = (X_{ij}, X_{ji})\}$ is a continuous-time Markov chain. The outcome space of the Markov chain is constituted by the four dyad configurations $\mathcal{X} = \{11, 01, 10, 00\}$, i.e. one mutual state, two asymmetric states and one null state in the order. The entire $X(t)$ digraph process is the result of $\frac{g(g-1)}{2}$ independent dyadic processes with identical infinitesimal transition matrix Q (Figure 2.5).

The corresponding graphical representation is depicted in Figure 2.5

$$Q = \begin{array}{c} \\ (0,0) \\ (1,0) \\ (0,1) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (1,0) & (0,1) & (1,1) \end{array} \\ \begin{bmatrix} -2\lambda_0 & \lambda_0 & \lambda_0 & 0 \\ \lambda_1 & -(\lambda_1+\lambda_0+\mu_0) & 0 & \lambda_0+\mu_0 \\ \lambda_1 & 0 & -(\lambda_1+\lambda_0+\mu_0) & \lambda_0+\mu_0 \\ 0 & \lambda_1+\mu_1 & \lambda_1+\mu_1 & -2(\lambda_1+\mu_1) \end{bmatrix} \end{array}$$

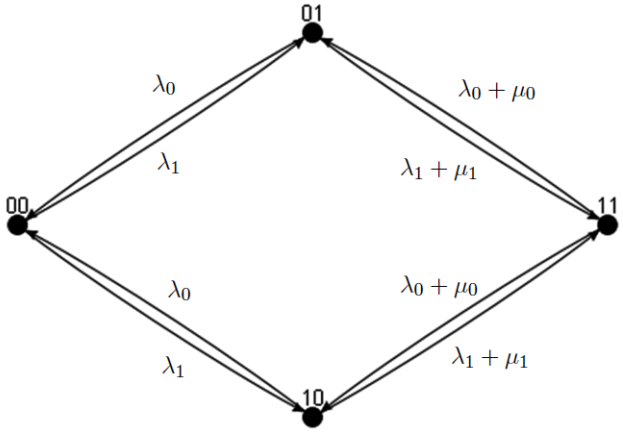


Figure 2.5: *Transition rates between dyads and the corresponding matrix of infinitesimal transition rates*

Generally speaking, the transition rates can be summarized in the following equations:

$$\begin{aligned} \lambda_{0ij}(x,t) &= \lambda_0 + \mu_0 x_{ji} \\ \lambda_{1ij}(x,t) &= \lambda_1 + \mu_1 x_{ji} \end{aligned} \tag{2.9}$$

They are assumed to be time homogeneous, so that the resulting stochastic process is stationary in time. Equations (2.9) shows that the transition rates depend on x_{ji} by a linear combination with coefficients λ_0 and μ_0 or λ_1 and μ_1.
Looking at Figure 2.5, it is possible to give an interpretation of the four parameters. It is clear that μ_0 and μ_1 measure the importance of a reciprocated tie and the importance of a dissolution of a reciprocated tie respectively. Instead λ_0 and λ_1 are measures of the overall rate of changes which described the pass from a null to an asymmetric dyad and from an asymmetric to a null dyad, respectively.
Since the presence of a tie between j and i should increase the tendency for a choice from i to j, it should happen that $\lambda_0 + \mu_0 \geq \lambda_0$. Reasoning in the same way, one should expect that the presence of a tie from j to i should decrease the tendency of a tie from j to i to disappear; thus, $\lambda_1 + \mu_1 \leq \lambda_1$.
A drawback of the reciprocity model is that the probability transition matrix for the process is quite difficult to compute. Wasserman shows that they are function of the eigenvalues and of the eigenvectors of the infinitesimal generator matrix. Fortunately, this difficulty does not echo to the limiting stationary transition distribution of the

process, which assumes the following forms according to each state:

$$\pi_{11} = \frac{\lambda_0(\lambda_0+\mu_0)}{(\lambda_0+\lambda_1)(\lambda_1+\mu_1)+\lambda_0(\lambda_0+\mu_0+\lambda_1+\mu_1)}$$

$$\pi_{10} = pi_{01} = \frac{\lambda_0(\lambda_1+\mu_1)}{(\lambda_0+\lambda_1)(\lambda_1+\mu_1)+\lambda_0(\lambda_0+\mu_0+\lambda_1+\mu_1)}$$

$$\pi_{00} = \frac{\lambda_1(\lambda_1+\mu_1)}{(\lambda_0+\lambda_1)(\lambda_1+\mu_1)+\lambda_0(\lambda_0+\mu_0+\lambda_1+\mu_1)}$$

Since the limiting stationary transition distribution depends on the four parameters λ_0, μ_0, λ_1 and μ_1, it is quite easy to estimate the three probabilities. In fact, once one has estimated the four coefficients, the corresponding values should be replaced in the above expressions. Because of the simplicity of the reciprocity model, this can be done using the data on the transitions on pairs of actors.

2.3.2 The popularity model and the expansiveness model

In 1980 Wasserman (Wasserman, 1980a,b) also proposed the popularity model in which individual i's choice or non choice of individual j depends only on the in-degree of j, denoted by x_{+j}. In fact, the transition rates are defined by the following equations:

$$\lambda_{0ij} = \lambda_0 + \pi_0 x_{+j}$$
$$\lambda_{1ij} = \lambda_1 0 + \pi_1 x_{+j}$$

It is assumed that they are time homogeneous, so that the resulting stochastic process is stationary. The popularity model describes the network evolution in terms of the social status of each member. In fact, it describes how an individual, who has high in-degree within a group, can influence the choices within the network.
This suggests that the entire digraph process is represented by g independent processes consisting of the colums of the adjacency matrix X(t).
Regarding the parameters of the model, λ_0 and λ_1 play the same role that they have in the reciprocity model. π_0 and π_1 measure the importance of the "popularity" of an individual j to be chosen or not chosen, respectively, by an individual i. Since it seems reasonable that more choices lead to more choices and a large in-degree should decrease the possibility of choices disappearing, π_0 can assume only positive values, while π_1 only negative values.
Instead of assuming that the columns of the adjacency matrix are on independent stochastic process, one can assume this hypothesis for the row of the adjacency matrix. The resulting model is the expansiveness model, whose rate of transitions are functions of the out-degree of actors i:

$$\lambda_{0ij} = \lambda_0 + \epsilon_0 x_{+j}$$
$$\lambda_{1ij} = \lambda_1 0 + \epsilon_1 x_{+j}$$

were ϵ_0 and ϵ_1 measure the importance of the aptitude to make many choices. As with the reciprocity model, computations of the probability distribution matrix are quite

difficult, and in this case, even the stationary limiting transition distribution is difficult to determine.

The models previously described make assumptions that do not allow to consider the entire network structure to describe network changes. In fact, independence assumptions break the network in parts, which are considered independent. For this reason Snijders (Snijders, 1996) proposed the well-known Stochastic actor-oriented model, based on the idea that individuals act in order to maximize a utility function, given constraints determined by the network. This model is fully described in the next chapter.

Chapter 3

A model for longitudinal data: the Stochastic actor-oriented model

Social networks are dynamic by nature. Ties may change over time: they can be established to form more complex structures or they can be broken. In recent years there has been growing interest towards longitudinal network analysis with three special issues in the *Journal of Mathematical Sociology* in 1996, 2001 and 2003 edited by Doreian and Stockman, and one special issue in *Social Networks* in 2010 edited by Snijders and Doreian.
Among models that have described network evolution over time, the Stochastic actor-oriented model allows to relax the hypothesis of dyad independence, including more complex dependence structures. In particular, the model described in this chapter deals with the compromise between the usual statistical requirement of parsimonious modeling and the flexibility to represent the complicated dependencies in network evolution process.

3.1 Assumptions of the Stochastic actor-oriented model

The Stochastic actor-oriented (SAO) model is a flexible class of models for network panel data, which is based on the assumption that networks evolve following continuous-time Markov chains. In order to introduce and describes this model some notations should be identified and arranged.
Let us focus the attention on a set of actors $\mathcal{G} = \{1, \cdots, g\}$ over which a relation $\mathcal{R}$ is defined. We can formally define a relation as a subset $\mathcal{R}$ of the Cartesian product $\mathcal{X} \times \mathcal{X}$. If $(i, j) \in \mathcal{R}$, then there is a tie between the actor i and the actor j, otherwise the two actors are disconnected. As already mentioned in the previous pages, we assume that the relations are nonreflexive, i.e. $(i, i) \notin \mathcal{R}$, and directed, i.e. $(i, j) \in \mathcal{R}$ does not imply that $(i, j) \in \mathcal{R}$.

The two sets $\mathcal{G}$ and $\mathcal{R}$ define a network X which we observed at $M \geq 2$ time points $t_1, \cdots, t_M$. The networks will be denoted by $x(t_1), \cdots, x(t_n)$. Furthermore, some actors'characteristics can be collected, such as gender, seniority rank, and so forth. These characteristics are called actors'attributes or covariates and will be denoted by $Z_1, \cdots, Z_H$ where H is the number of collected attributes.
The aim of the analysis of longitudinal network data consists in taking an insight into the network evolution, so that it is possible to determine the leading forces which govern the process. In more detail, the network evolution is the dependent variable, which we want to describe as a function of structural effects, explanatory random variables and explanatory dyadic variables.
Some assumptions about the evolving process should be formulated.
A fundamental assumption of the model regards the distribution of the process that describes network evolution. First of all, we assume that time, denoted by t, is a continuous variable. This does not mean that it is assumed to observe the network evolution in continuous time, but the observations are made at discrete time points $t_1, \cdots, t_M$. Indeed we assume that there is a latent process underlying the network observed and going on between the observation moments. This hypothesis allows us to consider the dependencies between network ties as the result of processes where one tie is formed as a reaction to the existence of other ties.
Just to clarify, we can consider a simple effect such as reciprocity and focus our attention on two observation moments. We can imagine that at the first observation two actors are mutually disconnected, and they form a mutual dyad at the second observation. If we consider a discrete time model, this reciprocal dyad is formed out of nothing, while if we use a continuous time approach, such a reciprocal dyad can be formed tie by tie, as a consequence of the creation of a tie from the first actor to the second and its reciprocation. Thus, we can explain the mutual dyad arisen between the two observation moments as the result of the reciprocity process.
The second assumption regards the distribution of the latent process which determines network evolution. The changing network is the outcome of a continuous-time Markov chain, which in the following pages will be denoted by $\{X_t, t \geq 0\}$. This is a very strong assumption that relies on the well-known *Markovian property*. This property states that: "The future depends on the present and on the past, only through the present". In our context this mean that "for any point in time, the current state of the network determines probabilistically its further evolution, and there are no additional effects of the earlier past" (Snijders *et al.*, 2010b).
Having assumed that the underlying process is a continuous-time Markov chain, network ties represent a state with a tendency to endure over time, rather than a brief event (Brandes *et al.*, 2009). For instance, if we consider ties determined by friendship, trust or cooperation, they can change but they endure over time. On the contrary, telephones calls or e-mail exchanges among a group of actors at any given time point are brief events, which start and finish in a short time.
The assumption of a continuous-time Markov chain is untenable for a network of events, while it will usually not be totally realistic for states, since further evolution of a network

can also depend on the past. If one disposes of three or more waves, he can test the Markovian assumption of time dependence and propose more complex models, which makes assumptions less restrictive. The drawback is that the resulting models are too complicated to manage. For this reason we can look at the Markov process as an approximation of the process that governs network evolution.
The third assumption determines the name of the model considered here. We assume that the actors control their outgoing ties, i.e. they decide to change one of their outgoing ties according to their position in the network, their attributes and the characteristics of the other actors. This means that actors are not changing their outgoing ties at will, but as will be shown later, they want to maximize a utility function under the structural constraints of the network. This is well-expressed in the structural individualism theory proposed by Udehn (Udehn, 2002).
To maximize the utility function actors should have full knowledge of the whole network, in terms of the existing ties and the attributes that characterize the other actors in the network. Thus, we assume that actors have complete information about the network and all the other actors. This assumption explains why the use of the SAO model is usually confined to a restricted number (few hundreds) of actors; in fact if the network is too large, it is unrealistic that actors has a global knowledge of the whole network.
From the previous lines it clearly appears that actors can decide which outgoing tie should be changed, giving them the power to act. This justifies why the longitudinal model considered here is termed "actor-oriented" (actor-based). A different perspective is used in the "tie-oriented" model which focuses on a pair of actors and on the existing ties between them. An overview of this approach is given by Snijders (2006) and Snijders (2009a).
Finally it is assumed that no more than one tie can change at any given moment t, i.e. only one actor has the opportunity to change one of his outgoing ties at t. "This implies that tie changes are not coordinated, and depend on each other only sequentially, via the changing configuration of the whole network" (Snijders *et al.*, 2010b). Just to clarify, we can consider again the reciprocity effect, and as already mentioned in the previous lines, we suppose there are no ties between two actors at the first observation time, and a mutual dyad at the second observation time. At this point this assumption suggests that the two actors involved in the dyad cannot negotiate or coordinate the creation of the ties between them, but again first one tie should be created and then the other, as a reaction to the first. In other words, tie changes are not coordinated and depend on each other only sequentially, according to the changing configuration of the whole network.
According to these four assumptions, the evolution process can be decomposed into its smallest possible components, which are called *micro-steps*. At each micro-step one probabilistically selected actor has the opportunity to change. He/She can decide to change or not to change one of his outgoing ties, so that the utility function is maximized. Formally we can describe each micro-step as a pair of elements: the first determines the waiting time between one opportunity to change and the next one, while the second determines the precise change which is made. The set of the all micro-steps represents the so-called *complete data*, which are the result of the latent process underlying the

evolution of the network. Thus, we cannot observe them, but only the final result.
Furthermore, each component of a micro-step can be modeled by a specific process, so that the actor-oriented process can be decomposed into two stochastic sub-processes, which will be described in the next paragraphs. The first one is the change opportunity process which models the frequencies of the changes, in terms of the waiting time between a change to another, while the second one is the change determination process which models the precise tie made by the actor who has the opportunity to change.

3.2 The formulation of the model

The SAO model can be described in two different ways: using an intensity matrix Q, since the evolution of the network is modeled as a continuous-time Markov chain, or using a generalized linear model, since network evolution can also be considered as a sequence of micro steps. The latter perspective also allows to easily introduce some useful notations for the first formulation, so it is convenient to consider first the formulation given by the generalized linear model.
According to this point of view, the network process is decomposed into the change opportunity process and the change determination process, which are modeled by the rate function and the objective function, respectively.

3.2.1 The rate function

Actors may change their ties at different frequencies, according to the position they have in the network and their covariates. For instance, we can imagine that "younger individuals might change their ties more frequently than older individuals, or that more central actors might change their ties more frequently than peripheral actors" (Snijders, 2009a). In order to take into account this distinct aptitude for change, the rate function is defined. It describes the average frequency at which each actor has the opportunity to change.
It is well-known in statistics that a very adequate model to interpret the number of occurrences in a specific interval of time is the Poisson process (Ross, 1996), which supposes that the occurrences are independent from one another. One of the attractive properties of the Poisson process is that the waiting time between one occurrence to another can be modeled by an exponential distribution with parameter λ, where λ represents the rate at which an event occurs.
In a network context the occurrence is the opportunity to change. The independent assumption of the occurrences is also respected in our context since one of the assumptions of the SAO models requires that tie changes are not coordinated. The waiting time between one opportunity of change and another is well described by an exponential distribution with parameters given by the so-called *rate function*.
We denote by $\rho_i(\alpha, x)$ the rate of changes of the actor i, where x is the current state of the network, and α is a vector of parameter. This notation expresses the dependence of the rate function from structural effects and actor covariates. The waiting times until

the next opportunity for change by any actors follows an exponential distribution with parameter:

$$\rho(\alpha, x) = \sum_{i=1}^{g} \rho_i(\alpha, x)$$

Thus, given that an opportunity for change occurs, the probability $\pi(\alpha, x)$ that it is the actor i who has the opportunity to change is given by:

$$\pi_i(\alpha, x) = \frac{\rho_i(\alpha, x)}{\rho(\alpha, x)}$$

The simplest specification of a network rate of change is obtained assuming that all actors have the same rate of change λ between two consecutive observation moments. Thus, the waiting times until the next opportunity for change by any actors follows an exponential distribution with parameter $g\lambda$, and the probability that an actor i has the opportunity to make a change is equal to $1/g$.
A model with a constant rate function is usually easier to explain and can be simulated in a simpler and therefore quicker way. The latter is an advantage given the time-consuming algorithm for estimation, which will be described later. In the following pages it is assumed that the rate function is the same for all the actors.
In order to understand the following parts, it is not necessary to give a more in depth insight of the rate function and on the effects that determine its variation over the actors. Thus, no more will be said about the change opportunity process. More details can be found in Carrington *et al.* (2005) and in Snijders & Van Duijn (1997)[1].
Before describing the objective function some words should be spent about the interpretation of the rate parameter λ. It can assume non negative values, and the higher its value is, the greater the number of changes between two observation moments is. Thus, the effect of the rate of change on the observed network can be expressed in terms of the number of ties turned into its opposite, which is computed using the following formula:

$$s_m(x(t_{m+1}), x(t_m)) = \sum_{\substack{i,j=1 \\ i \neq j}}^{m} |x_{ij}(t_{m+1}) - x_{ij}(t_m)|$$

3.2.2 The objective function

The objective function plays a key role in the change determination process, i.e. it determines the precise tie that an actor made when he had the opportunity to change. The idea underlying the change determination process is that at a given time t, actor i has the opportunity to change and he can choose to modify one of his outgoing ties or maintain his outgoing ties as they are. The purpose of the change determination process is to probabilistically describe the choice of i.

[1] In this paper Snijders and van Duijn proved that the reciprocity model proposed by Wasserman in 1980 is a special case of the SAO model when the rate function is a linear combination of the in-degree, out-degree, and reciprocated degree.

Let us focus on an actor i, who can change one of his outgoing ties at a certain given moment, and analyze which are the possible choices of i. Given the current state of the network, i can decide not to change anything or to change one of his outgoing ties, for instance the tie x_{ij} directed to an actor j, into its opposite. Since we are considering simple digraphs, and a tie can assume values 1 or 0, according to the fact that it is present or it is absent, changing a tie into its opposite means that the tie variable changes from 1 to 0 or from 0 to 1. In the first case the tie is terminated, while in the second the tie is created. Thus, if a relation between i and j exists in the current state of the network ($x_{ij} = 1$) and i decides to change it, the considered tie is deleted ($x_{ij} = 0$). Vice versa the tie is created.
This suggests that the set of admissible choice has cardinality equal to g: $g - 1$ changes and 1 non-change. Consequently, the set of possible states for the considered network given the current state contains g elements: $g-1$ network which are equals to the current state except for the value assumed by the changed tie and 1 equal to the current state.
As previously said it turns out that each actor can choose between a discrete finite set of alternatives, which are mutually exclusive (since the selected actor can make only one change) and exhaustive, since he can decide among all the other actors having full knowledge of the entire network.
There is a wide econometric literature that deals with discrete choices (Hensher & Johnson, 1981; Agresti & Corporation, 1990; Green, 2000; Train, 2003). Among the models that were formulated, there are the random utility models. They are applied when a decision maker faces a choice between n alternatives. He would obtain a certain level of utility from each alternative, so that he chooses the alternative that provides him the greatest utility.
In a random utility model the utility function U_{ij} of an actor i facing the choice j is given by:

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{3.1}$$

where V_{ij} is the part of utility that a researcher can capture while ϵ_{ij} is a random term. Different models can be defined according to the distribution of ϵ_{ij}. In particular, if ϵ_{ij} is distributed as an extreme type I distribution (Gumbel)[2], then the resulting model is a multinomial logit model. Thus, the probability that an actor i faces the choice j is[3]:

$$p_{ij} = \frac{e^{V_{ij}}}{\sum_{j=1}^{n} e^{V_{ij}}} \tag{3.2}$$

The random utility model can be applied to model the change determination process, and the utility function corresponds to the so-called objective function. Thus, the ob-

[2]Different distributions were proposed to model extreme values. Among them, there is the Gumbel distribution, which presents the following distribution function (Johnson & Kotz, 1970):

$$P(X \leq x) = exp\{-exp\{(x-a)/b\}\} \quad -\infty < x < \infty,\ a \in \mathbb{R},\ b > 0$$

[3]The proof can be found in Maddala (1986).

jective function represents the utility function that an actor wants to maximize given the constraints in the current network structure. Informally speaking, it expresses the degree of satisfaction of an actor towards the current state of the network and how likely it is for the actor to change the current state in a particular way.
Formally, we will denote the objective function by $f_i(\beta, x(i \rightsquigarrow j))$, where the subscript i underlines that this is the utility function for the focal actor i, while the two quantities between parentheses express the idea that the objective function is a function of the state of the network obtained when i changes his outgoing tie towards j, $x(i \rightsquigarrow j)$, and of statistical parameters β. In more detail, the objective function is defined as

$$f_i(\beta, x(i \rightsquigarrow j)) = \sum_{i=1}^{k} \beta_k s_{ik}(x(i \rightsquigarrow j)) + U_i(t, x, j) \tag{3.3}$$

where β_k are statistical parameters, $s_{ik}(x(i \rightsquigarrow j))$ are the effects and $U_i(t, x, j)$ is a random utility term. Let us analyze each component of the linear combination in equation (3.3) in more detail.
The $s_{ik}(x(i \rightsquigarrow j))$ are called effects, and they are relevant functions of the digraphs which are supposed to play a key role in the network evolution. In other words they represent the leading forces of the underlying process that governs network changes from an observation moment to another. As will be shown later, examples of effects are reciprocity and transitivity, whose outcomes have already been encounter in Paragraph 1.3.3. It is fundamental to specify that network effects are aspects of the network as perceived by the focal actor i.
The strength of each effect is represented by the corresponding parameter β_k, which should be estimated on the basis of the longitudinal network data observed. β_k can assume any real values and can be interpreted as follows. If β_k is equal to 0, it means that the corresponding effect plays no role in the network dynamics. If it assumes positive values, then there is higher probability of moving into networks where the corresponding effect is higher. Vice versa if the parameter takes negative values there is higher probability of moving into networks where the corresponding effect is lower. Thus, for instance, if the parameter β_k is related to the reciprocity effect and it takes a positive value, this means that the number of reciprocal dyad increases between two observation moments, so that there is evidence towards reciprocity.
The last term $U_i(t, x, j)$ of the objective function is the random term, distributed as a Gumbel distribution. Thus, according to equation (3.2), the probability that an actor i changes his outgoing ties towards j or leaves his outgoing ties variables unchanged is:

$$p_{ij}(\beta; x(i \rightsquigarrow j)) = \frac{e^{f_i(\beta, x(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_i(\beta, x(i \rightsquigarrow h))}} \tag{3.4}$$

If $i = j$, then equation (3.4) represents the probability of not changing anything.
Having described the mathematical form of the objective function, we will discuss which effects can be included in the model.

The objective function effects

Several effects are proposed in order to specify the objective function. They can be determined endogenously or exogenously, according to the fact that they regard the structure of the network or actor attributes, i.e. covariates. A complete list of effects can be found in Ripley & Snijders (2010), but here only the basilar ones are considered. Each effect is characterized by a subscript i to remind that these network aspects are determined from the point of view of the actor i, who has the opportunity to change. Furthermore, we use a more compact notation, denoting by x the resulting network instead of $x(i \rightsquigarrow j)$.
Let us start considering endogenous effects which depends on network structure.

Endogenous effects
In this paragraph a short description of the main endogenous effect is given.

- The *outdegree effect* corresponds to the number of outgoing ties of actor i and is defined as:
$$s_{i1}(x) = \sum_j x_{ij}$$
Usually the corresponding parameter β_1 is negative if the network is sparse (i.e. very low density).

- The *reciprocity effect* represents the number of mutual dyads in which the actor i is involved:
$$s_{i2}(x) = \sum_j x_{ij}x_{ji}$$
Denoting by β_2 the associated parameter, β_2 often assumes positive values, since a lot of social relations show a tendency towards reciprocation.

- The *transitive effect* can be modeled according to different effects. Each of them interprets a different situation. The more common structure is represented by the *transitive triad*, whose related effect counts the number of transitive patterns in which an actor i is involved:
$$s_{i3}(x) = \sum_{j,h} x_{ij}x_{ih}x_{jh}$$

- The *three cycle-effect* is a sort of a generalized reciprocity which involves three actors i, j and h. It expresses the idea that it is not important if j reciprocates the tie from i, but it is important that he sends a tie to an actor h, who has an outgoing tie towards i. This means that the tie between i and j is reciprocated in an indirect way through actor h.
$$s_{i4}(x) = \sum_{j,h} x_{ij}x_{jh}x_{hi}$$

- The *in-degree related popularity effect* is defined as the sum of the in-degrees of whom i is related to:

$$s_{i5}(x) = \sum_j x_{ij} \sqrt{\sum_h x_{hj}}$$

and it expresses the idea that a very popular actor is more chosen. This means that popularity reinforce itself.

- The *out-degree related activity effect* represents the sum of the out-degrees of the others to whom i is related:

$$s_{i6}(x) = \sum_j x_{ij} \sqrt{\sum_h x_{jh}}$$

The interpretation of these effects is similar to the previous one but it regards the out-degree. In particular, the out-degree related activity effect suggests that an actor is more active since the more popular. In this case popularity reinforces expansivity.

- The *indirect ties effect* is the number of actors j to whom i is indirectly tied to, through at least one intermediary:

$$s_{i7}(x) = \#\{j : x_{ij} = 0, \max_h(x_{ih}, x_{hj}) > 0\}$$

- The *balance effect* may also be called *structural equivalence with respect to outgoing ties*. It expresses a preference of actors to have ties to those other actors who have a similar set of outgoing ties as they have. It is defined by the similarity between the outgoing ties of actor i and the outgoing ties of the other actors j to whom i is tied, to

$$s_{i8}(x) = \sum_{j=1}^{g} x_{ij} \sum_{\substack{h=1 \\ h \neq i,j}}^{g} (b_0 - \mid x_{ih} - x_{jh} \mid)$$

where b_0 is a constant included to reduce the correlation between this effect and the density effect.

There are many other effects that are not considered here which regard very specific structural network properties. One can wonder how to choose among them. From a practical point of view, the choice of effects that should be included in the model is guided by theory. In fact, according to the kind of network in analysis, different hypotheses about the leading forces of network dynamics can be formulated. In order to test them one should include the related effects in the objective function.
For instance, if we consider "friendship", we know that transitivity ("the friend of my friend is also my friend") is a well established network structure, and so we should include it in the model. Instead, if we are studying an advice network, we can imagine that a person to whom a lot of other people ask advice is chosen the most. Maybe because if many people ask this actor for advice, it means that he is reliable and one can trust him. Thus, other people decide to ask to him. In this case, the in-degree related popularity

effect plays a key role and should be tested.
Like all the other statistical models each effect is controlled for all the other effects included in the model. For this reason, one practical suggestion is to always include the density effect, since all the structural effects are related to the presence or absence of ties. Reciprocity is also fundamental in social relationships, and for this reason it should not be forgotten.

Exogenous effects
Let us now consider some effects related to actor attributes. Let Z be a covariate, such as gender, ethnicity, seniority rank, etc. We denote by z_i and z_j the values or the categories assumed by the covariate on the actor i and on the actor j, respectively. Some interesting effects are represented by:

- the *covariate-related popularity* that is defined by the sum of the covariate over all actors to whom i has a relation:
$$s_{i9}(x) = \sum_j x_{ij} z_j$$
- the *covariate-related activity* that is defined by $i's$ out-degree weighted by his covariate values
$$s_{i10}(x) = \sum_i x_{i+} z_i$$
- the *covariate-related similarity* that is the sum of measure of covariate similarity between i and j:
$$s_{i11}(x) = \sum_j x_{ij} \left(1 - \frac{|z_i - z_j|}{R_Z}\right)$$
where R_Z is the range of Z.

Again, the choice of exogenous covariates to be included in the objective function should be determined according to hypotheses derived from theory. If we consider friendship network data gathered in adolescent groups, sociological theory suggests that girls trust girls and boys trust boys, showing an evidence towards homophily with respect to this attribute. Thus, covariate-related similarity effect plays a key role and should be included in the objective function.
A positive parameter β_k, associated to popularity or activity effects will lead to association between the covariate and the receiver and sender tendency of an actor, respectively. In the same way a positive parameter related to the covariate-related similarity will lead to relations being formed particularly between actors who have similar values on the covariate.

3.2.3 An alternative formulation of the model

In the previous paragraph we described the network evolution process as the results of small changes, called micro-steps, and modeled by the rate and the objective functions. These two elements also define a continuous-time Markov chain, which can be

completely specified through the intensity matrix Q, whose generic element is the rate of change from one state to another. Let us define this matrix in our context.
The continuous time stochastic process is defined on the set $\mathcal{X}$ of all digraphs, or adjacency matrices. We denote by x the initial state of the network. At a certain time point an actor i has the opportunity to change his outgoing ties towards an actor j. If he decides to change, the next state of the network is $x(i \rightsquigarrow j)$, where the link x_{ij} between i and j is changed into its opposite. Thus, we can denote the rate of change from a state x to $x(i \rightsquigarrow j)$ using the following notation $q(x; x(i \rightsquigarrow j))$. We can compute it according to the definition of the rate of change given in (2.8):

$$q(i,j) = \lim_{dt \to 0} \frac{P\left(X(t+dt) = x(i \rightsquigarrow j) | X(t) = x\right)}{dt} = \lambda_i(x) p_{ij}(x)$$

where $\lambda_i(x)$ is the rate function and $p_{ij}(x)$ depends on the objective function. Since we assume that at each time point only one actor may change, all transition rates for matrices belonging to $\mathcal{X}$ and differing in more than one element are equal to 0. Consequently, the rate of change between two states, denoted by the networks x and $x^{'}$, is given by:

$$q(x, x^{'}) = \begin{cases} \lambda_i(x) p_{ij}(x) & \text{if } x \text{ and } x^{'} \text{ differ in the element (i,j)} \\ \\ 0 & \text{if } x \text{ and } x^{'} \text{ differ in more than 1 element} \end{cases} \tag{3.5}$$

If all the actors have the same opportunity to change then $\lambda_i(x) = \lambda$, and a simplest notation can be used.
The intensity matrix is a fundamental ingredient to prove that there is a relation between the ERGM and the SAO models. In more detail, focusing the attention on the generic element of Q and on the objective function defined in equation (3.3), one can prove that the ERGM is the stationary distribution for the continuous-time chain identified by Q (Snijders, 2001; Snijders *et al.*, 2010b), i.e. for the process that describes network evolution. Equation (2.5) is also the limiting distribution of the process that describes network dynamics, since all states communicate with one another[4]. Equation (3.5) shows that the network evolution process depends on the rate and on the objective functions. Both of them depend, as well, on parameters. In more detail, assuming that the rate function is constant between two consecutive observation moments, that each actor has the same opportunity to change, and that we have observed the network at Mdiscrete time points, the rate function depends on $M-1$ parameters λ_m $(m = 1, \cdots, M-1)$. Regarding the objective function, if we consider K effects then there are K parameters that are involved in the utility function. These parameters are denoted by β_k, $k = 1, \cdots, K$
If we are interested in determining which effects are relevant to explain network dynamics

[4]The ERGM model is not exactly the limiting distribution for the considered evolution stochastic process, since the SAO model is actor-oriented and the ERGM are not actor-oriented. Instead, the tie-based (Snijders, 2006, 2009a) version of the model described in this chapter has the ERGM as its limit distribution.

and their strength, then we must estimate the parameter which specify the SAO model and test if it is significant.
Let us now consider the procedure used and implemented in specific network analysis software to estimate the parameters of the SAO model.

3.3 The parameter estimation and testing

3.3.1 The estimation procedure

Let $\boldsymbol{\theta}$=($\boldsymbol{\lambda}$,$\boldsymbol{\beta}$) be the parameter of the SAO model, where $\boldsymbol{\lambda}$ is an $M-1$-dimensional parameter, according to the M observational time points considered, and $\boldsymbol{\beta}$ is a K-dimensional parameter, since there are K effects included in the objective function. Thus, the dimension for θ is $M-1+K$.
The proposed estimation technique is the Method of Moments (MoM) whose logic is quite straightforward. Let X be a random variable with distribution $\varphi(x;\boldsymbol{\delta})$, which depends on a p-dimensional vector of parameters $\boldsymbol{\delta} \in \boldsymbol{\Delta}$. To estimate $\boldsymbol{\delta}$, one can observe that the theoretical moments of a certain distribution usually depend on the statistical parameters which fully specify the distribution. Thus, the idea of the MoM is to estimate the parameter $\boldsymbol{\delta}$ with the values that assure that the theoretical expected values are equal to their sample counterparts.
To compute its estimate, the MoM requires to find a vector of p statistics $\boldsymbol{S}$ and to estimate $\boldsymbol{\delta}$ with the value that satisfies the following system of p equations:

$$E_{\boldsymbol{\delta}}[\boldsymbol{S}] = \boldsymbol{s}$$

where $\boldsymbol{s}$ is the vector of the sample values assumed by the statistics.
It follows that, if we want to estimate the parameter $\boldsymbol{\theta}$ of the SAO model, we must find $M-1+K$ statistics, set the theoretical expected value of each statistic equal to its sample counterpart, and solve it with respect to $\boldsymbol{\theta}$.
Since each parameter is related to a specific effect, the logic is to determine the statistics as functions of the corresponding effects. This idea is also reinforced by the fact that we cannot apply any formal method, such as reduction to sufficient statistics. Thus, relevance means that the expected values of the statistics should be sensitive to the parameter change. Thus, the relevant statistics for the parameters are:

- the total amount of change in the $m-th$ time period. It is an important statistic for the rate of change λ. It is defined as the number of different ties between two observation moments and is computed as follows:

$$S_{\lambda_m} = \sum_{\substack{i,j=1 \\ i \neq j}}^{g} |X_{ij}(t_{m+1}) - X_{ij}(t_m)|$$

  The choice of this statistic also relies on the property that it presents. In fact, if the parameters of the objective function are all equal to 0, then the model reduces to the trivial situation where the ties are randomly changing 0-1 variables, and S_{λ_m} is a sufficient statistics for λ_m.

- the sum over all actors i of the digraph statistics observed at time t_{m+1}

$$S_{mk} = \sum_{i=1}^{g} s_{ik}(X(t_{m+1}))$$

for the parameter β_k.

Having determined the statistics, the moment conditions should be defined. In order to do this one should keep in mind that the rate parameter λ is assumed to be constant within each time period, so that $M-1$ values must be estimated. Regarding the parameters β_k of the objective function, it is assumed that they are constant over the whole observation period[5], thus we must estimate K parameters related to the utility function of the model.
Consequently, the MoM estimator for θ is defined as the solution of the system of equation:

$$\begin{cases} E_\theta \left[S_{\lambda_m} \left(X(t_m), X(t_{m+1}) \right) | X(t_m) = x(t_m) \right] = s_{\lambda_m}(x(t_1), x(t_0)) & m = 1, \cdots, M-1 \\ \sum_{m=1}^{M-1} E_\theta \left[S_{mk} \left(X(t_{m+1}) \right) | X(t_m) = x(t_m) \right] = \sum_{m=1}^{M-1} s_{mk}(x(t_{m+1}), x(t_m)) & k = 1, \cdots, K \end{cases} \tag{3.6}$$

The system (3.6) does not always have a solution, even if we have sufficient information, i.e. we have as many moment conditions as parameters. In particular, the MoM is not suitable for observation $x(t_m)$ and $x(t_{m+1})$, which differ for a high number of ties, i.e. for couple of networks in which the statistic S_{λ_m} takes too high values. In practice, this happens when the observation moments are too far apart.
To express quantitatively the meaning of "too far apart", Snijders suggested to use the Jaccard index (Snijders, 2005), a similarity index that can also be applied to tie variables. Denoting by N_{11} the number of ties present at both waves, by N_{01} the number of ties newly created and by N_{10} the number of ties terminated, the Jaccard index is computed using the following formula:

$$Jacc = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{3.7}$$

Experience has shown that values between two consecutive observation moments should be higher than 0.3, otherwise one can doubt about the assumption that the change process is gradual.
At this point, the problem is how to solve the system (3.6) with respect to the parameter $\boldsymbol{\theta}$. Analytical and usual numerical procedures cannot be applied, since the theoretical expected value of the statistic cannot be calculated explicitly. For this reason a stochastic approximation of the solution is performed through the Robbins-Monro algorithm. This algorithm is implemented in the SIENA (Simulation Investigation for Empirical Network

[5]This hypothesis is quite restrictive. If it is not satisfied, estimation problems, such as bias and inefficiency, can arise (Lospinoso *et al.*, 2010).

Analysis) program, now also available as an R library called *RSiena*. This program allows to analyze cross-selectional data, longitudinal data and longitudinal data on networks and behaviors. Download, references and manual can be found at the website page *http://stat.gamma.rug.nl/siena.html*.

The Robbins-Monro algorithm

In this paragraph a brief sketch about the Robbins-Monro algorithm is given since the algorithm proposed later in Chapter 5 has a similar structure. An exhaustive description can be found in Snijders (2001).
The aim of the algorithm is to stochastically approximate the solution of equation (3.6) with the value $\widehat{\boldsymbol{\theta}}$ and to provide the corresponding variance-covariance matrix $\Sigma_{\widehat{\boldsymbol{\theta}}}$. Denoting by $\boldsymbol{S}$ the $M-1+K$-dimensional vector of statistics involved in the process estimation and by $\boldsymbol{s}$ its sample counterpart, this formula can be re-expressed in the following way:

$$E_{\boldsymbol{\theta}}[\mathbf{S}|X(t_0) = x(t_0)] = \mathbf{s}$$

The approximation $\Sigma_{\widehat{\boldsymbol{\theta}}}$ is given by:

$$cov(\widehat{\theta}) = D_{\theta}^{-1}\Sigma_{\theta}D_{\theta}^{-1}$$

where D is the first order derivative matrix of $\boldsymbol{S}$ with respect to θ:

$$D = \frac{\partial}{\partial\theta}E_{\boldsymbol{\theta}}[\mathbf{S}|X(t_0) = x(t_0)]$$

and $\Sigma_{\boldsymbol{\theta}}$ is the variance-covariance matrix of $\boldsymbol{S}$.
The Robbins-Monro algorithm proposed in this context is the multivariate version of the Robbins - Monro algorithm proposed in 1951 (Robbins & Monro, 1951) and is based on the iteration step:

$$\widehat{\boldsymbol{\theta}}_{N+1} = \widehat{\boldsymbol{\theta}}_N - a_N D^{-1}(\mathbf{s}_N - \mathbf{s})$$

where a_N is a sequence that converges slowly to 0, and $\mathbf{S}_N$ are simulated values according to the probability distribution defined by the parameter $\widehat{\boldsymbol{\theta}}_N$.
The computer algorithm is composed of three phases. The first provides a preliminary estimation of the first order derivative matrix D based on a small set of simulations from the starting values for $\widehat{\boldsymbol{\theta}}$; the second is the estimation phase related to the Robbins-Monro iteration step. This phase is divided into several sub-phases, each of them provides the estimation $\boldsymbol{\theta}_N$ which is the starting point for the following sub-phase. The estimate derived from the last sub-phase is the potential estimate for $\boldsymbol{\theta}$. The last phase is used for the estimation of the variance-covariance matrix of the estimator and to check the convergence for the algorithm.
The algorithm provides a stochastic approximation of the solution. Stochastic approximation methods are described in Chapter 4, since a similar approach is used to estimate the parameter of the SAO model according to the proposed new statistics.

Regarding the estimation process a Maximum Likelihood approach was recently proposed (Snijders *et al.*, 2010a) to estimate the parameter $\boldsymbol{\theta}$. The procedure requires the use of data augmentation and stochastic approximations. Simulation results seem to be promising since they show that the Maximum Likelihood estimator is more efficient than the regular MoM estimator for small data sets. Furthermore, the Maximum Likelihood approach can also be used to elaborate selection procedures and can be extended to more complex models.

3.3.2 Tests and goodness of fit

Once the parameters are estimated, statistical tests are performed to determine the significance of each parameter. Since, hypothesis testing is still a topic in working progress, just a brief review is given here, referring to the literature for more details.
Two possible tests are available according to the nature of the involved hypothesis. Let us start by considering the case in which the null hypothesis is simple, namely we want to test a single parameter of the model $H_0 : \theta_j = 0$. Having estimated the standard error of the estimators for the parameter θ, the natural way to test each parameter is represented by the usual statistical T-test. Denoting by $\widehat{\theta_j}$ the estimate of the generic component θ_j of the vector θ, by $s.e.(\widehat{\theta_j})$ the corresponding standard error, and by $H_0 : \theta_j = 0$ the null hypothesis, the test statistic is defined by:

$$\frac{\widehat{\theta_j}}{s.e.(\widehat{\theta_j})} \sim T$$

Under the null hypothesis, this ratio has a T distribution which can be approximated by the normal distribution as the number of simulation increases.
An alternative approach is to use a score-type test, which also allows to test composite hypotheses. The implemented version is a generalized Neyman-Rao score test (Schweinberger, 2005), and the generic null hypothesis regarding $j > 1$ parameters is expressed by:

$$H_0 : \boldsymbol{\theta}_0 = (\theta_1 = \cdots = \theta_j) = 0 \tag{3.8}$$

To test this hypothesis the unrestricted and the restricted model are considered, and denoting by $U(\widehat{\theta_0})$ and $I(\widehat{\theta_0})$ the score function and the information matrix computed in $\widehat{\theta_0}$, respectively, the test statistic is given by:

$$\eta_U = U(\widehat{\theta_0})^T I^{-1}(\widehat{\theta_0}) U(\widehat{\theta_0})$$

Testing multiple parameter can also be considered as testing the goodness of fit of the model (Schweinberger, 2005), since the test statistic is some function of the difference between the expected number of the considered effects and their observed value. Thus, having fixed the null hypothesis in (3.8), its rejection means that the model does not fit the data, and that the parameter are significant and must be included in the model.

3.4 Extensions of the SAO model

Since 1996, the year of the first proposal of the SAO model, several extensions were proposed by Snijders and colleagues (Snijders, 2001, 2007; Snijders *et al.*, 2007, 2010a). Some of them are related to the formulation of the model and other to its applications. Regarding the probabilistic structure of the model, one of the main developments relies on the idea that there can be different effects of endogenous and exogenous variables on the propensity of creating and terminating ties. In the previous lines it was assumed that terminating a tie is just the opposite of creating one, but this does not always represent reality. "It is conceivable, for example, that the loss when terminating a reciprocal tie is greater than the gain in creating one; or that transitive closure works especially for the creation of new ties, but hardly guards against termination of existing ties" (Snijders *et al.*, 2010b).
To model this difference a new definition of the objective function was given. The objective function is still defined as a linear combination of effects, statistical parameters and a random term, but it now depends on two components: the evaluation function $f_i(x(i \rightsquigarrow j); \beta)$ and the endowment function $g_i(x(i \rightsquigarrow j); \gamma)$. The former operates on the creation of ties, while the latter on the termination. Thus, the objective function is defined as:

$$\begin{aligned} & f_i(x(i \rightsquigarrow j); \beta) + g_i(x(i \rightsquigarrow j); \gamma) + U_i(t, x, j) = \\ & = \sum_{i=1}^{k} \beta_k s_{ik}(x(i \rightsquigarrow j) + \sum_{h=1}^{H} \beta_h s_{ih}(x(i \rightsquigarrow j) + U_i(t, x, j) \end{aligned} \tag{3.9}$$

From equation (3.9) it follows that the endowment function is also defined as a linear combination of effects, and that the evaluation and the endowment functions can include different effects.
A second extension of the SAO model has the aim to include and explicitly model social selection and influence, represented in Figure 1.8. In more detail, changes in a network can be due to endogenous or exogenous mechanisms, related to structural effects and to individual characteristics of network actors, respectively. If we think about a common relation, such as friendship, we can imagine that friends may influence one another, or friendship ties can rise because there is similarity between ego and the potential alter, with respect to some actor characteristics. Thus, the idea is to model network evolution distinguishing between the two processes.
Consequently, further assumptions should be postulated, mainly extending those for the network tie change to actor-level outcomes. First of all, we assume that each actor controls his/her outgoing ties as well as his/her own behavior, and that at any given time point t one actor can change one of his outgoing ties or one of his own behaviors. Furthermore, it is assumed that the changes that an actor applies to his/her outgoing ties are conditionally independent from the changes of his/her behavioral characteristics. These conditions allow to split the network evolution process into two sub-processes, one for influence and the other for selection. More detail about the modeling of co-evolution of networks and behavior can be found in Snijders *et al.* (2007) and Steglich *et al.* (n.d.).

Chapter 4

Generalized method of moments applied to parameter estimation of the Stochastic actor-oriented model

The vector of parameters $\boldsymbol{\theta}$ of the SAO model is estimated using the Method of Moments, whose logic is to find the value for $\boldsymbol{\theta}$ so that the sample moments are equal to their theoretical counterparts. Since the statistics involved in the Method of Moments estimation are computed considering only network configurations at time t_m, one can suspect that more information can be used exploiting the observation of the network at time t_{m-1}. For this reason new statistics, which are able to take into account both observational time points, are defined to estimate the parameter of the SAO model.
The number of new statistics exceeds the number of parameters. This led to an overdetermined system which requires the employment of the Generalized Method of Moments.

4.1 New statistics for parameter estimation

Let X be a network observed at two time points t_0 and t_1 on a set of n actors and $x(t_0)$ and $x(t_1)$ denote the observed networks. A very simple Stochastic actor-oriented model is considered. It includes rate, degree, reciprocity and transitivity effects so that the vector of parameters which characterize the model is $\boldsymbol{\theta}= (\lambda, \beta_1, \beta_2, \beta_3)$. Using the Method of Moments the estimation for $\boldsymbol{\theta}$ requires to set as many moment conditions as parameters, i.e. four moment conditions should be defined. From Chapter 3, it follows that an estimate for $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ that solves the system:

$$\begin{cases} E_\theta \left[S_R \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_R(x(t_0), x(t_1)) \\ E_\theta \left[S_1 \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_1(x(t_1)) \\ E_\theta \left[S_2 \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_2(x(t_1)) \\ E_\theta \left[S_3 \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_3(x(t_1)) \end{cases} \tag{4.1}$$

where S_R is the statistic for the rate parameter and S_1, S_2, S_3 are the statistics for degree, reciprocity and transitivity effects.
Looking at the last three equations of system (4.1), one can notice that the involved statistics are computed taking into account only the network $X(t_1)$, in fact the values they assume correspond to the total number of ties, reciprocal dyads and transitive triads observed at the last time points.
Since the network was also observed at time t_0 and $X(t_0)$ is only involved in the statistic for the rate of change, one can wonder if all the information deriving from the network observation at two consecutive time points is completely exploited. The answer is negative, since the observation of network structure at time t_0 allows to know from which mechanism a particular configuration originated, i.e. it also provides precious information for reciprocity and transitivity.
This aspect is very relevant for the purpose of this work, since it led to the idea of a definition of new statistics for the parameters estimation of the SAO model. For this reason it is necessary to clarify the meaning of the previous lines.
Let us start by considering reciprocity. As pointed out in Chapter 2 reciprocity is the simplest dependence structure in a network and it involves couples of actors. We consider the network $X(t_1)$, and we focus on a couple of actors i and j who constitute a reciprocal dyad. This situation is represented on the right side of Figure 4.1.
The mutual dyad noticed at t_1 could originate from different configurations existing between i and j at time t_0. In particular, three options are available from the point of view of actor i: it could be that at time t_0 no ties between i and j existed or j was in relationship with i or the mutual dyad was already present (left side of Figure 4.1). Knowing the nature of the dyad at the first observational time point allows us to distinguish from different mechanisms that gave rise to the mutual dyad found at time t_1.
In more detail, if at time t_0 a null dyad was observed we can infer that actors i and j started a reciprocal relation, but we do not exactly know how this happened: it could be that i sent a tie towards j and then j decided to reciprocate the relation or vice versa j was first the sender and then the receiver of ties[1]. This is a sort of "*ex-novo*" or "*new reciprocity*", which is created from nothing between the two time points considered.
If at time t_0 an asymmetric dyad was observed, for instance between j and i, we can

[1]The real sequence is unknown because the process evolves in continuous time but we observed it only at discrete time points.

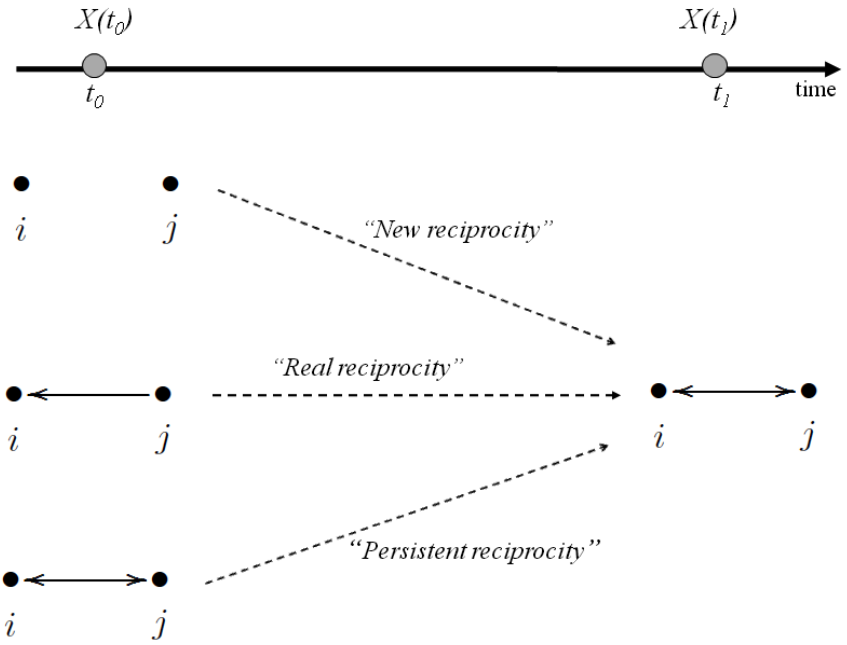


Figure 4.1: *Different situations from which a reciprocal dyad can arise*

infer that at a certain time between t_0 and t_1, i decided to send a tie towards j. Thus, we know the direction of reciprocity within the couple. In this case we will speak about "*real reciprocity*".
Finally, if a mutual dyad between i and j was also observed at time t_0, we can infer that reciprocal ties endure through time, and so we will speak about "*persistent reciprocity*". From a theoretical point of view, these three situations give a different contribution to reciprocity, so it seems reasonable to take into account both the final and the starting dyad configuration in the estimation process of the reciprocity parameter.
Practically this means that we introduced new moment conditions in the system (4.1) which do not only require that the expected values of the number of reciprocal dyad equals the corresponding observed values at time t_1, but also the equality for each case that leads to a mutual dyad, i.e. for "new", "real" and "persistent" reciprocal dyad. Since the interest is turned to reciprocity, one can suggest that the situation indicated with the terms "real reciprocity" is more relevant than the other two situations. For this theoretical reason and for statistical reasons which will be explained in the following pages, new and persistent reciprocity are jointly considered so that two moment conditions are added to the system (4.1).
The introduction of two moment conditions make it necessary to define two new statistics which were determined on the basis of the usual statistic S_2. The statistic related

to *new reciprocity* is

$$S_{21}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \sum_{i}^{g} S_{i21}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \\ = \sum_{i,j} X_{ij}(t_1) X_{ji}(t_1)(1 - X_{ij}(t_0)) X_{ji}(t_0) \tag{4.2}$$

where the sum over all of actors i assures that both of the cases considered in Figure 4.1 are counted. The statistic is built by taking into account the dichotomous nature of adjacency matrix cells. The product $X_{ij}(t_1)X_{ji}(t_1)(1 - X_{ji}(t_0))X_{ji}(t_0)$ takes value 1 if all the factors are equal to 1 and 0 otherwise. In terms of presence/absence of ties it should happen that:

- at time t_1 there should be a reciprocal dyad between i and j so that $X_{ij}(t_1) = 1$ and $X_{ji}(t_1) = 1$
- at time t_0 there should not be a tie between i and j so that $X_{ij}(t_0) = 0$, and consequently $1 - X_{ij}(t_0) = 1$
- at time t_0 there should be a tie between j and i so that $X_{ji}(t_0) = 1$

This proves that S_{21} counts only the reciprocal dyad originated from the reciprocation of a pre-existing tie.
Following the same statement, the corresponding statistic for new and persistent reciprocity is defined by:

$$S_{22}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \sum_{i}^{g} S_{i22}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \\ = \sum_{i,j} X_{ij}(t_1) X_{ji}(t_1)(1 - X_{ij}(t_0))(1 - X_{ji}(t_0)) + \sum_{i,j} X_{ij}(t_1) X_{ji}(t_1) X_{ij}(t_0) X_{ji}(t_0) \tag{4.3}$$

where the two addends compute the number of reciprocal dyads which originated respectively from null dyads and mutual dyads.
The new statistics for the reciprocity effect are related to the usual statistic S_2 by the relationship:

$$S_2 = S_{21} + S_{22}$$

In other words the new statistics decompose the original statistic according to the configuration existing at time t_0 for each couple of actors so that more information is considered.
A similar argument can be used for transitive triads. We now focus on three actors i, j and h who form a transitive triads at time t_1 (right side of Figure 4.2). Following the same statement applied for dyad configurations, we look for the possible situations from which a transitive triad can arise. There are eight different possibilities[2], but only some

[2]The number of possible configurations is given by the combination with repetition of 2 elements in class three since each tie can take value 0 or 1 and a transitive triad involves three ties.

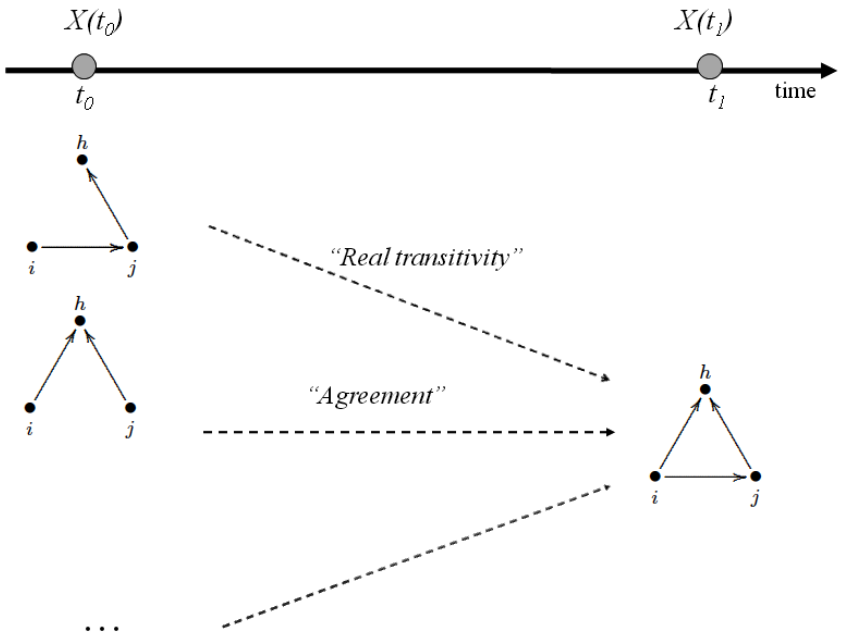


Figure 4.2: *Different situations from which a transitive triad can arise*

of them are considered because they are theoretically relevant and they are built from the point of view of the actor i (left side of Figure 4.2).
In particular, a first possibility is provided by the existence of two ties at time t_0, one between actors i and j and one between actors j and h. Then a relation from i to h was established between the two observational points. Thus, the transitive triad observed at time t_1 is really the result of transitivity. For this reason we refer to this first situation using the term "*real transitivity*".
A second possibility is that both i and j are related to h at time t_0 and then i decided to establish a tie with j. Since i and j are in relationship with h, we can say that they agree about the receiver choice, and so the resulting transitive triad at t_1 is the outcome of "*agreement*".
In Chapter 1 transitivity was explained through the sentence "The friend of my friend is also my friend", and the first circumstance depicted in Figure 4.2 exactly represent this process (*real transitivity*). While the second situation describes that both i and j were friends of h and then i chooses j as his friend (*agreement*).
We can suppose that the transitive triad observed arose for the same reasons, for instance because the three actors start going out together [3]. In each situation the disconnected actors get in touch going out together, but what is different is the mechanism that generates the third tie. In fact, for real transitivity a sort of chain effect "He (i) invited

[3]This is a big simplification of reality, since there are a lot of other factors which cause the birth of friendship relations, but for simplicity we can neglect all the other reasons

me (j), I (j) invite you (h)" is the engine. The flows of words from a friend to another enable people to meet and new friendship relations to arise.
Regarding agreement, we can imagine that if both i and j are friends of h, they go out together with him and so they meet, and they have the opportunity to become friends. Thus, the centrality of the actor i with respect to the other two considered actors plays a key role in this situation.
Again from a practical point of view, the two configurations from which a transitive triad arose give a different contribution to transitivity. Following the approach for dyads, the idea is to distinguish transitive triads according to initial configurations at time t_0 and to require that the expected number of each combination equals the observed values.
In order to define the new moment conditions (one for each of the cases is depicted in Figure 4.2), three new statistics are defined. Once more they are determined taking into account the dichotomous nature which describes the presence or the absence of a tie within a couple of actors. In more detail they are computed as a product of six factors, three related to $X(t_1)$ and three to $\mathrm{X}(t_0)$.
The statistics for the real transitivity is defined as:

$$S_{31} = S_{31}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \sum_{i=1}^{g} S_{i31}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) =$$

$$= \sum_{i,j,h} X_{ij}(t_1)X_{jh}(t_1)X_{ih}(t_1)X_{ij}(t_0)X_{jh}(t_0)(1 - X_{ih}(t_0)) \tag{4.4}$$

and it is easy to verify that each terms of the sum takes value 1 for each tern of actors i, j and h, if and only if a transitive triad is observed at time t_1 ($X_{ij}(t_1) = 1, X_{jh}(t_1) = 1, X_{ih}(t_1) = 1$) and ties between i and j and j and h are presented at time t_0 ($X_{ij}(t_0) = 1, X_{jh}(t_0) = 1, X_{ih}(t_0) = 0$) and 0 otherwise.
It is not difficult to verify that the following statistic is related to agreement:

$$S_{32} = S_{32}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = \sum_{i=1}^{g} S_{i32}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) =$$

$$= \sum_{i,j,h} X_{ij}(t_1)X_{jh}(t_1)X_{ih}(t_1)(1 - X_{ij}(t_0))X_{jh}(t_0)X_{ih}(t_0) \tag{4.5}$$

The statistic related to all the other possible configurations is determined as the difference among the statistic S_3 which interprets the number of transitive triads at time t_1 and those defined in the previous line:

$$S_{33}\left(X(t_0), X(t_1) \mid X(t_0) = x(t_0)\right) = S_3 - (S_{31} + S_{32}) \tag{4.6}$$

At this point we have created five new statistics which lead to five new moment conditions so that, applying the principle of Method of moments, system (4.1) can be rewritten in

the following way:

$$\begin{cases} E_\theta \left[S_R \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_R(x(t_1), x(t_0)) \\ E_\theta \left[S_1 \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_1(x(t_1), x(t_0)) \\ E_\theta \left[S_{21} \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_{21}(x(t_1), x(t_0)) \\ E_\theta \left[S_{22} \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_{22}(x(t_1), x(t_0)) \\ E_\theta \left[S_{31} \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_{31}(x(t_1), x(t_0)) \\ E_\theta \left[S_{32} \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_{32}(x(t_1), x(t_0)) \\ E_\theta \left[S_{33} \left(X(t_0), X(t_1) \right) | X(t_0) = x(t_0) \right] = s_{33}(x(t_1), x(t_0)) \end{cases} \tag{4.7}$$

The moment conditions related to the usual statistics for reciprocity and transitivity are replaced respectively by two and four moment conditions. The result is that now there are 7 moment conditions for the estimation of 4 parameters, and the system is overdetermined.
The regular method of moments (MoM) suggests to set as many moment conditions as parameters, so that there is sufficient information to estimate the parameter. When there are more moment conditions than parameters the regular MoM cannot be applied, but its generalization can be used. This generalization which deals with a bigger quantity of information is called Generalized Method of Moments.
From mathematical theory it follows that usually a system with more equations than unknowns has no solution unless there are linearly dependent equations that can be neglected so that the number of equations corresponds to the number of unknowns. If this does not happen, the usual methods for solving a system do not work and approximate solutions are the only alternative.
One of the methods for looking for approximate solutions is the least square method, which minimizes the norm of the difference between the two terms of each equations. If we refer to the matrix form, we can write the system as $Ax = b$, where A is the matrix of coefficients, x is the vector of unknowns and b is the vector of known terms. Then the solution is the value x' of x, so that $x' = \arg\min_x \|Ax - b\|$.
In system (4.7), equations are not linearly dependent and so to solve the system a logic similar to the least square can be applied. As a consequence, the estimate for $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which minimizes a specific metric between observed and theoretical values. This is the idea behind the Generalized Method of Moments.

4.2 The Generalized Method of Moments

The Generalized Method of Moments (GMM) refers to a class of estimators introduced into the econometric literature by Hansen (Hansen, 1982) and also developed

contemporaneously and independently by Burguete et al. (Burguete *et al.*, 1982). Since then, GMM has become very popular both in applications and in theoretical analysis for some main reasons.
From a practical point of view the GMM provides some advantages with respect to the Maximum likelihood procedure. First of all the GMM supplies a way to estimate parameters of partially specified models where maximum likelihood estimation is not feasible. In fact, while the maximum likelihood method requires the specification of the full data generating process, Hansen's framework uses information in population moments as a basis for the estimation of any set of parameters of interest, so that it can also deal with partially specified models.
Second, the GMM can be applied under very weak conditions on the generation of the random variables, thereby allowing wide application in cross-sectional, panel, or time-series data.
Third the GMM offers computational advantages. Even in fully specified models, the MLE can create problems, since it is too cumbersome to implement; whereas the GMM provides a practical method of obtaining consistent, asymptotically normal estimators of parameters in potentially nonlinear dynamic models.
Finally, from a theoretical point of view the GMM is a very general framework for statistical analysis, since it includes least squares (LS), instrumental variables (IV) and maximum likelihood (ML) estimation techniques as special cases.

4.2.1 The estimation method

There are two alternative ways to specify GMM estimators, but in this context only one is considered[4].
Let $\mathbf{x} = (x_1, \cdots, x_n)$ be a realization of a stochastic process $\{x_t, t = 1, 2, \cdots\}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ an unknown p-dimensional vector of parameters, which index a statistical model. It is supposed that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$. In order to estimate $\boldsymbol{\theta}$ it is assumed that the parameter satisfy the moment conditions:

$$E\left[\boldsymbol{f}(x;\theta)\right] = 0 \tag{4.8}$$

where $\boldsymbol{f}(\mathbf{x};\theta)$ is a $q \times 1$ vector of continuous and differentiable functions of $\boldsymbol{\theta}$. Some other requests can be specified on $\boldsymbol{f}$ according to the model so that asymptotic results, such as the law of large numbers and the central limit theorems, are applicable.
Three conditions must hold so that the system of equations defined by (4.8) provides enough information to estimate $\boldsymbol{\theta}$.
The first is an identification condition, which assumes that the system (4.8) is satisfied only at the true value $\boldsymbol{\theta}_0$, i.e.

$$E\left[\boldsymbol{f}(x;\theta)\right] = 0 \;\Leftrightarrow\; \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

The second condition is based on the fact that each equation in (4.8) corresponds to a piece of information, and so there should be at least as many moment conditions as

[4]References about the two approaches are given by Hansen (Hansen, 2007).

parameters to be estimated, i.e. $q \geq p$. This condition, as will be shown later, is not sufficient because one must also require that every moment condition contains q unique pieces of information about θ_0 so that:

$$E\left[\frac{\partial \boldsymbol{f}(x,\theta_0)}{\partial \theta}\right]$$

is of full rank p.
If these conditions are satisfied, then a GMM estimator of $\boldsymbol{\theta}$ can be constructed as follows. Let $\boldsymbol{g}(x,\theta)$ denote the sample counterparts of (4.8) constructed from a sample of size n, namely,

$$\mathbf{g}(x,\theta) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(x,\theta)$$

and let

$$Q(\theta) = \mathbf{g}(x,\theta)' W \mathbf{g}(x,\theta)$$

where W is a $q \times q$ positive definite matrix which may depend on the data, but it converges in probability to a matrix of constants.
A GMM estimation for θ is the value $\widehat{\theta}$ which minimizes $Q(\theta)$:

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} Q(\theta) = \arg\min_{\theta\in\Theta} \mathbf{g}(x,\theta)' W \mathbf{g}(x,\theta)$$

Since W is a symmetric positive definite matrix, $Q(\theta)$ is a meaningful measure of distance. Consequently it clearly appears that the logic behind the GMM estimation is to minimize the distance between the observed and the expected values, a principle which is very closed to that of minimum Chi-square.
The symmetry and the positive definite nature of W also allow to apply the Cholesky decomposition, so that W can be decomposed into the product of a lower triangular matrix and its conjugate transpose. In this case, the conjugate corresponds to the lower triangular matrix since W is a real matrix. This suggests that W can be decomposed into matrix product $V'V$, where V is a lower triangular matrix. Consequently the function $Q(\theta)$ can be rewritten in a Euclidean norm form:

$$Q(\theta) = \|V\mathbf{g}(x,\theta)\|^2$$

so that a GMM estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero.
Some other words should be spent on the requirements of the weighting matrix. The positive definiteness of W ensures that both $Q(\theta) \geq 0$ for any $\boldsymbol{\theta}$, and also that $Q(\widehat{\theta}) = 0$ if $\mathbf{g}(x;\widehat{\theta}) = 0$. Thus, $Q(\theta)$ can be made exactly equal to 0 in the just identified case, but it is strictly positive in the over-identified case [5].

[5] A less restrictive request is that W is a semi-definite matrix. $Q(\theta)$ continues to be a distance measure, but the positive semi-definiteness leaves open the possibility that $Q(\theta)$ is zero at a value of $\widehat{\theta}$, which does not satisfy the sample moment conditions

Since the estimate for $\boldsymbol{\theta}$ is the value which minimizes $Q(\theta)$, it is the solution to the first-order conditions:

$$\underbrace{\Gamma(\theta)'}_{p\times q}\underbrace{W}_{q\times q}\underbrace{\boldsymbol{g}(\theta)}_{q\times 1} = 0 \tag{4.9}$$

where $\Gamma(\widehat{\theta})$ is the $(q \times p)$ matrix of the first order partial derivative of $\boldsymbol{f}(\theta)$ with respect to $\boldsymbol{\theta}$. Thus the ij-th element of this matrix is:

$$[\Gamma(\theta)]_{ij} = \frac{\partial \boldsymbol{g}_i(\theta)}{\partial \theta_j} \qquad i = 1, \cdots, q, \quad j = 1, \cdots, p$$

In this way we set p linear combinations of the q moment conditions equal to zero so that the usual techniques for solving a system of equations can be applied. This explains the assumption that the matrix of the first order derivative expected value should have full rank. In fact the system defined by (4.9) admits a solution if and only if the p equations are independent, since otherwise no sufficient information is provided.
Formula (4.9) can be written since one of the assumptions is that $\boldsymbol{f}$ is a differentiable function of $\boldsymbol{\theta}$. This assumption is not fundamental, in some cases it is useful but is unnecessary in nearly all the models, but it allows to give an intuitive interpretation for the GMM.
In particular (4.9) reveals that the estimator GMM based on $E[\boldsymbol{f}(\mathbf{x};\theta)] = 0$ corresponds to the regular MoM based on:

$$\Gamma(\theta)'(W^{1/2})'W^{1/2}E[\boldsymbol{f}(\mathbf{x};\theta)] = \Delta(\theta)'W^{1/2}E[\boldsymbol{f}(\mathbf{x},\theta)] = 0 \tag{4.10}$$

where $\Delta(\theta) = W^{1/2}\Gamma(\theta)$. If $p = q$ then (4.10) is equivalent to $E[\boldsymbol{f}(x;\theta)] = 0$ and the GMM reduced to the regular MoM estimator, so that the weighting matrix W plays no role in the analysis. If $q > p$, the population moment conditions can be decomposed into identifying and over-identifying restrictions associated with GMM estimation. In particular, the identifying restrictions represents the part of the population moment conditions which are used in estimation, and they are picked out by the least squares projection of $W^{1/2}E[\boldsymbol{f}(\mathbf{x};\theta)]$ onto the column space of $\Delta(\theta)$:

$$\Delta(\theta)\left[\Delta(\theta)\Delta(\theta)\right]^{-1}\Delta(\theta)^T W^{1/2}E[\boldsymbol{f}(\mathbf{x};\theta)] = 0$$

The overidentifying restrictions represent the remainder, therefore they are the least squares projection of $W^{1/2}E[\boldsymbol{f}(\mathbf{x};\theta)]$ onto the orthogonal space of $\Delta(\theta)$:

$$\left\{I_q - \Delta(\theta)\left[\Delta(\theta)\Delta(\theta)\right]^{-1}\Delta(\theta)^T\right\}W^{1/2}E[\boldsymbol{f}(\mathbf{x};\theta)] = 0$$

The role of the two sets of restrictions is reflected in their sample counterparts. Consequently, since the identifying restrictions represent the information used during the estimation process, their sample analogues are satisfied at $\widehat{\boldsymbol{\theta}}$ by construction. In contrast, the over-identifying restrictions are ignored in estimation and so their sample analogs are not satisfied.

From (4.9) it emerges that the coefficient of the p linear combinations are determined by the weighting matrix W, which plays two key roles: it allows to define a proper metrics, and it attributes a particular importance to each equation of (4.9). This suggests that the choice of W is quite relevant during the estimation process. Looking at the asymptotic distribution and properties of the GMM, Hansen determined the optimal choice for W, which is one of the subjects of the next paragraph.

4.2.2 Properties of GMM estimators

One of the main advantages of the GMM is that it yields estimators with very desirable large sample properties, such as consistency and asymptotically normal distribution, under very weak conditions. These properties were first demonstrated by Hansen in 1982 (Hansen, 1982) and then developed by Newey and McFadden in 1994 (Newey & McFadden, 1994). Their contribution to the growing number of applications is unquestionable. In more detail let us consider the assumptions that lead to the proof of consistency and asymptotic normality of the parameter estimators.
Let $\{x_t : t \geq 1\}$ be a stationary and ergodic process defined on a probability space $\Omega \subseteq \mathbb{R}^r$ and $\boldsymbol{f} \colon \Omega \times \mathbb{R}^q$ a function so that (i) is continuous on $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ for all $x \in \Omega$, (ii) $E[\boldsymbol{f}(x;\theta)]$ exists and is finite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and (iii) $E[\boldsymbol{f};\theta)]$ is continuous on $\boldsymbol{\Theta}$.
Then some regularity conditions must hold. It is assumed that $\boldsymbol{f}$ is a C^1 function on $\boldsymbol{\Theta}$, and that the expected value of its first derivative exists and is finite.
Finally, some assumptions about identification and parameter space should be considered. First, it is assumed that the moment conditions are satisfied at the true value $\boldsymbol{\theta}_0$. Second, global or local identification should be assumed. The former requires that the population moment condition only holds at one value in the entire parameter space, i.e.

$$E[\boldsymbol{f}(x;\theta)] \neq 0 \qquad \forall \theta \in \boldsymbol{\Theta} \qquad \text{such that} \qquad \theta \neq \theta_0$$

while the latter is a weak condition which requires that:

$$rank \left\{ E\left[\frac{\partial \boldsymbol{f}(x,\theta_0)}{\partial \theta} \right] \right\} = p$$

Local identification is necessary since global identification failure often arises because of the nature of $\boldsymbol{f}$ as a function of $\boldsymbol{\theta}$ or because it is difficult to find primitive conditions for global identification. Then the idea is to search an identification condition which holds for some suitably defined neighborhood of $\boldsymbol{\theta}_0$, instead of for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$[6]. Regarding the

[6]To derive the condition for local identification it is necessary to introduce the concept of $\epsilon -$ *neighborhood*. A $\epsilon -$ *neighborhood* of θ_0 is defined to be the set N_ϵ $N_\epsilon = \{\theta; \|\theta - \theta_0\| < \epsilon\}$. Considering sufficiently small ϵ so that $\mathbf{g}(\mathbf{x};\theta)$ is equal to the following first order Taylor series expansion in N_ϵ

$$\mathbf{g}(x;\theta) = \mathbf{g}(x;\theta_0) + \left[\frac{\partial}{\partial \theta'} \mathbf{g}(x;\theta_0) \right] (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

and taking expectation on both sides yields

$$E[\mathbf{g}(x;\theta)] = E\left[\frac{\partial}{\partial \theta'} \mathbf{g}(x;\theta_0) \right] (\theta - \theta_0)$$

parameters space we shall assume that $\boldsymbol{\Theta}$ is a compact space.
If all these conditions are satisfied, then the GMM estimator is consistent (Theorem 2.1 in Newey & McFadden (1994)).

Theorem 1. *If there is a function $Q(\theta_0)$ such that:*
i) $Q(\theta_0)$ is uniquely maximized at $\boldsymbol{\theta}_0$
ii) $\boldsymbol{\Theta}$ is compact
iii) $Q_n(\theta)$ converges uniformly in probability to $Q(\theta_0)$
then $\boldsymbol{\theta}_n$ converges in probability to $\boldsymbol{\theta}_0$.

Just to give a hint to the proof and to explain the role played by the previous assumptions, we can consider that the idea is to prove that if $\boldsymbol{\theta}_n$ minimizes $Q_n(\theta)$ and $Q_n(\theta)$ converges in probability to $Q(\theta_0)$ whose unique minimum is at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $\boldsymbol{\theta}_n$ must converge in probability to $\boldsymbol{\theta}_0$.
Uniform convergence and continuity together with the moment existent assumption are needed to use the law of large numbers and prove that $Q_n(\theta)$ converges in probability to $Q(\theta_0)$, while identification is fundamental to prove that $\boldsymbol{\theta}_n$ must converge in probability to $\boldsymbol{\theta}_0$.
Regarding compactness, it is substantive since it requires that the parameter bounds are known, but it is often ignored in application contexts, since it can be replaced by the condition that $\boldsymbol{\Theta}$ is a convex set, $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$ and $Q_n(\theta)$ is concave.
This is stated in the following theorem (Theorem 2.7 in Newey & McFadden (1994)) where the uniformly convergence in probability is also relaxed with a convergence in probability.

Theorem 2. *If there is a function $Q(\theta_0)$ such that:*
i) $Q(\theta_0)$ is uniquely maximized at $\boldsymbol{\theta}_0$
ii) $\boldsymbol{\theta}_0$ is an element of the interior of a convex set $\boldsymbol{\Theta}$ and $Q_n(\theta)$ is concave
iii) $Q_n(\theta)$ converges in probability to $Q(\theta_0)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
then exists with probability approaching 1 and $\boldsymbol{\theta}_n$ converges in probability to $\boldsymbol{\theta}_0$.

The consistency of the GMM estimator is the basis for its asymptotic normality distribution. Some additional assumptions are necessary in order to apply the central limit theorem and the Slutsky theorem and to infer the normal distribution. They are stated in the following theorem:

Theorem 3. *If $\widehat{\boldsymbol{\theta}}$ is a consistent estimator for $\boldsymbol{\theta}$ and:*
i) $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$
ii) $\mathbf{f}$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ with probability approaching one
iii) $E\left[\mathbf{f}(x;\theta)\right] = 0$ and $E\left[\|\mathbf{f}(x;\theta)\|\right]$ is finite
iv) $E\left[\sup_{\theta\in\mathcal{N}} \left\| \dfrac{\partial \boldsymbol{f}(x,\theta)}{\partial\theta} \right\| \right] < \infty$
v) $\Gamma' W \Gamma$ is nonsingular for $\Gamma = E\left[\dfrac{\partial \mathbf{f}(x,\theta_0)}{\partial\theta}\right]$

then for $\Omega = E\left[\mathbf{f}(x,\theta_0)\mathbf{f}(x,\theta_0)'\right]$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left[0, (\Gamma' W \Gamma)\Gamma' W \Omega W \Gamma (\Gamma' W \Gamma)^{-1}\right]$$

The first three assumptions are used to prove that the first order conditions computed in $\widehat{\boldsymbol{\theta}}$ are satisfied with probability approaching one and to approximate $\boldsymbol{f}(x,\widehat{\theta})$ using the Mean value theorem, so that after analytical rearrangement it is possible to write:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left[(\Gamma(\widehat{\theta})'\widehat{W}\Gamma(\overline{\theta})\right]^{-1}\Gamma(\widehat{\theta})'\widehat{W}\boldsymbol{f}(x,\widehat{\theta_0})$$

where $\overline{\boldsymbol{\theta}}$ is the mean value.
Assumption (iv) allows to prove that $\Gamma(\widehat{\theta})$ and $\Gamma(\overline{\theta})$ converge in probability to Γ so that $\left[\Gamma(\widehat{\theta})'\widehat{W}\Gamma(\overline{\theta})\right]^{-1}\Gamma(\widehat{\theta})'\widehat{W}$ converge in probability to $\left[\Gamma' W \Gamma\right]^{-1}\Gamma' W$.
Applying the central limit theorem to $\mathbf{g}(x,\theta)$ it follows that $\mathbf{g}(x,\theta)$ converges in distribution to a Normal distribution with mean 0 and variance-covariance matrix Ω. And then for the Slutsky theorem the thesis of Theorem 3 follows.
Some interesting observations for the efficiency of the GMM estimator can be deducted by its asymptotic distribution. In particular, it can be easily proved that an optimal choice for the weighting matrix is $W = \Omega^{-1}$; in fact, the asymptotic variance-covariance matrix becomes:

$$(\Gamma' W \Gamma)^{-1}\Gamma' W \Omega\ W'\Gamma(\Gamma' W \Gamma)^{-1} = (\Gamma' W \Gamma)^{-1}$$

and the GMM estimator is efficient.
Efficiency, consistency and asymptotic normality are desirable properties and are one of the strong points of the GMM estimators. Like all the other methods the GMM also presents some drawbacks. In particular, it suffers from the same problem of the regular MoM. The concern about GMM and MoM estimation is the choice of moment conditions since one in general will obtain a different estimator using a different set of moment conditions implied by a model. This was Fisher's criticism of the MoM which led him to propose the Maximum Likelihood. Since there is currently little guidance available upon which moment conditions to choose, the choice of moments remains a drawback to the method.

4.2.3 Computation of GMM estimators

The previous two paragraphs define the GMM estimators, describe its logic and its properties, but do not practically specify how to construct such an estimator. Two orders of problems arise.
The first is related to the minimization of the objective function $Q(\theta)$ because an analytical solution does not always exists even in simple cases. Let us considers an example which is also useful to underline the peculiarities of GMM estimators.

Example 1. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two i.i.d. samples obtained from two normal distributions with the same mean μ and different variance $\sigma_x^2 \neq \sigma_y^2$:

$$\boldsymbol{x}=(x_1,\cdots,x_n) \text{ i.i.d. from } X\sim N(\mu,\sigma_x^2)$$
$$\boldsymbol{y}=(y_1,\cdots,y_n) \text{ i.i.d. from } Y\sim N(\mu,\sigma_y^2)$$

It is supposed that the variance-covariance matrix is known:

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}=E\left\{\begin{bmatrix}x_i-\mu\\ y_i-\mu\end{bmatrix}\begin{bmatrix}x_i-\mu & y_i-\mu\end{bmatrix}\right\}$$

In order to estimate μ, it is natural to take a minimum-variance linear combination of the two sample means. This is also an example of the GMM estimator. To prove this let us define the function $\mathbf{g}$ as the difference between the sample mean and its theoretical counterpart both for $\mathbf{x}$ and $\mathbf{y}$.

$$\mathbf{g}(x,y;\mu)=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}x_i-\widehat{\mu}\\ \frac{1}{n}\sum_{i=1}^{n}y_i-\widehat{\mu}\end{bmatrix}=\begin{bmatrix}\bar{x}-\widehat{\mu}\\ \bar{y}-\widehat{\mu}\end{bmatrix}$$

and according to the GMM perspective an estimate for μ is the value $\widehat{\mu}$ that set the vector $\mathbf{g}$ as close to $\mathbf{0}$ as possible.
Since there are more conditions then parameters, we look for the value $\widehat{\mu}$ that minimizes the following quadratic form:

$$Q(\mu)=\mathbf{g}(x,y;\mu)'\Sigma^{-1}\mathbf{g}(x,y;\mu)=\begin{bmatrix}\bar{x}-\widehat{\mu}\\ \bar{y}-\widehat{\mu}\end{bmatrix}'\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}\bar{x}-\widehat{\mu}\\ \bar{y}-\widehat{\mu}\end{bmatrix}=$$

$$=\begin{bmatrix}\bar{x}-\widehat{\mu} & \bar{y}-\widehat{\mu}\end{bmatrix}\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}\bar{x}-\widehat{\mu}\\ \bar{y}-\widehat{\mu}\end{bmatrix}= \tag{4.11}$$

$$=\tfrac{1}{\sigma_x^2\sigma_y^2-\sigma_{xy}}\left[(\bar{x}-\mu)^2\sigma_y^2-2\sigma_{xy}(\bar{x}-\mu)(\bar{y}-\mu)+(\bar{y}-\mu)^2\sigma_x^2\right]$$

Equation (4.11) shows that GMM puts weights on the sample means proportional to the inverse of their variances, so that a higher weight is associated to the sample with the lower variance. This procedure should also recall the generalized least squares.
Computing the first order derivative of $Q(\mu)$ with respect to μ, setting the result equal to 0

$$\frac{\partial}{\partial\mu}Q(\mu)=-2(\bar{x}-\mu)\sigma_y^2+2\sigma_{xy}(\bar{y}-\mu)+2\sigma_{xy}(\bar{x}-\mu)-2(\bar{y}-\mu)\sigma_x^2=0$$

and solving the equation provides the GMM estimate for μ:

$$\widehat{\mu}=\tfrac{\bar{x}\sigma_y^2+\bar{y}\sigma_x^2-(\bar{x}+\bar{y})\sigma_{xy}}{\sigma_x^2+\sigma_y^2-2\sigma_{xy}}$$

In this case the computation of the GMM estimate is quite easy thanks to the fact that it was possible to apply an analytical approach and that the variance-covariance

matrix is known. In other cases these two facts do not happen and different strategies must be carried out.
In particular, numerical minimization routines represent the solution if an analytical approach is not feasible. Two possibilities are available. The first one is *grid search*, which requires to search over the entire parameters space. Since in most problems a grid search is computationally inefficient, it is not often used. An alternative is represented by *gradient methods* (Fletcher, 1980; Quandt, 1983; Gallant & Corporation, 1987; Bonnans *et al.*, 2006), which are deterministic numeric approaches to optimization problems.
Just to give an idea, we consider a function $h(\theta)$ depending on a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ which should be minimized. Minimizing h with respect to $\boldsymbol{\theta}$ is equivalent to solve the equation $\frac{\partial}{\partial\theta}h(\theta) = 0$. The gradient method produces a sequence of $\{\boldsymbol{\theta}_i\}$ defined by step:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left[\frac{\partial^2}{\partial\theta\partial\theta'}h(\theta_i)\right]^{-1}\frac{\partial}{\partial\theta}h(\theta_i)$$

that converges to the exact solution $\boldsymbol{\theta}$, when the domain $\boldsymbol{\theta}$ and the function $h(\theta)$ are both convex. This kind of algorithm is based on a series Taylor approximations of function to be minimized and are performed till the estimate for the parameter stabilizes. Usually this is measured by the size of each parameter increment relative to the previous parameter value.
In the context of the GMM estimator, Newton's step can be defined in the following way:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left[\frac{\partial^2}{\partial\theta\partial\theta'}Q(\theta_i)\right]^{-1}\frac{\partial}{\partial\theta}Q(\theta_i) = \theta_i - \left[\Gamma' W \Gamma\right]^{-1}\Gamma' W \boldsymbol{f}(x;\theta) \tag{4.12}$$

and the procedure is iterated till a convergence criterion is satisfied. There are different criteria which can be adopted. The first one is based on the fact that if $\widehat{\boldsymbol{\theta}}$ is the value which minimize $Q(\theta)$, then the updating routine should not move away from this point. This suggests that the minimum has been found if

$$\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\| < \epsilon \tag{4.13}$$

where ϵ is an arbitrarily small positive constant, usually equal to 10^{-6}. An alternative is to assume that the updating should not alter the objective function so that

$$|Q(\theta_i + 1) - Q(\theta_i)| < \epsilon \tag{4.14}$$

The main criticism of these criteria is that they indicate lack of progress rather then convergence. Nevertheless, in most cases lack of progress occurs because a minimum was encountered. Another issue relies on the fact that the encountered minimum can be a local one and not the global.
Alternatively, convergence can be assessed by examining the first order conditions, so that once the minimum is reached the following criterion should be satisfied

$$\left\|\frac{\partial}{\partial\theta}Q(\theta)\right\| < \epsilon \tag{4.15}$$

but the issue of finding a local minimum is still not solved.
From (4.12) it clearly emerges that the Newton-Raphson step depends on the weighting matrix W. Regarding the optimal choice for the weighting matrix W, it was proved that the GMM is efficient if W corresponds to the inverse of the variance covariance matrix of $\boldsymbol{g}$. In theory we can assume that the variance-covariance matrix is known, but in practice this never happens so that it is necessary to estimate W in a consistent way.
This leads to the well-known "two-step" estimator (Hansen, 2007) which works in this way: on the first step a sub-optimal choice of W (usually the identity matrix) is used to obtain a preliminary estimation for $\boldsymbol{\theta}$, which is denoted by $\widehat{\boldsymbol{\theta}_1}$; then $\widehat{\boldsymbol{\theta}_1}$ is used to obtain a consistent estimator for W, which is denoted by $\widehat{W_1}$. On the second step $\boldsymbol{\theta}$ is re-estimated assuming $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}_1}$. This process can be continued iteratively until the estimates converge or the number of two steps is large. The resulting estimator is known in the literature as the *iterated GMM estimator*.
An alternative to the *iterated GMM estimator* is given by the *Continuously Updating GMM* which contemporaneously updates the estimate for W and for $\boldsymbol{\theta}$ (Pakes & Pollard, 1989). This means that "instead of taking the weighting matrix as given in each step of the GMM estimation, we also consider an estimator in which the covariance matrix is continuously altered as $\boldsymbol{\theta}$ is changed in the minimization" (Hansen *et al.*, 1996). The peculiarity of this method is that the weighting matrix varies with $\boldsymbol{\theta}$ and this alters the shape of the criterion function that is minimized.
The second problem is related to the moment condition which requires to equal the theoretical moment with its sample counterpart. In some situations the analytical expression for the theoretical moment does not exist, since the solution of the involved integral is too difficult or does not exist. It is easy to imagine that this problem also arises with the regular MoM, and in this context the basic idea is to approximate the moment conditions using Monte Carlo (MC) Simulation or Importance sampling (IS). This method is known in the literature as the "Method of Simulated Moments" (MSM) (McFadden, 1989; Pakes & Pollard, 1989).
Just as the egular MoM can be generalized to the GMM, the MSM can be extended to the case where there are more moment conditions than parameters leading to the "Generalized Method of Simulated Moments" (GMSM) (Liesenfeld & Breitung, 1998; Gourieroux & Monfort, 1996, 1993).
Since the estimation now requires a set of simulations, the variance of the GMM estimator contains an additional component which is due to the variation in the MC or IS estimates of the moment restrictions. Nevertheless, the additional variance vanishes as the number of simulations sample size increases, and the GMSM estimator attains the efficiency of the corresponding GMM estimator.

4.3 The GMM estimation applied to the SAO model

The proposal of new statistics for the parameter estimation of the SAO model and the application of the principles of the regular MoM led to the system (4.7), described in Paragraph 4.1. Consequently the moment conditions to estimate the parameters of

the model are given by:

$$E_{\boldsymbol{\theta}}\left[S_i\left(X(t_0), X(t_1)\right) | X(t_0) = x(t_0) - s_i(x(t_1), x(t_0))\right] = 0 \tag{4.16}$$

where the quantity between square brackets plays the role of the function $\boldsymbol{f}$ and the subscript $\boldsymbol{\theta}$ indicates that this expected value is a function of the parameter $\boldsymbol{\theta}$. The subscript i is an index that identified the statistic involved.
A more convenient way to write (4.16) is to use a vectorial notation:

$$E_{\boldsymbol{\theta}}\left[\mathbf{S}\left(X(t_0), X(t_1)\right) | X(t_0) = x(t_0) - \mathbf{s}(x(t_1), x(t_0))\right] = 0 \tag{4.17}$$

where $\boldsymbol{S}$ is the vector of the statistics and $\boldsymbol{s}$ is the vector of the corresponding observed values.
Equation (4.16) (or equivalently equation (4.17)) defined an over-identified system of equations, so that an estimate for parameter $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which minimizes a specific metric between the observed and theoretical moments. This specific metric is defined by the GMM.
The application of the GMM in the econometric context refers to a sample-based inference. Therefore, it should be adjusted to a model based context which is typical of network modeling.
Let us start formalizing the estimation problem and defining the quantities involved. First of all, we must verify that the system in (4.17) provides enough information to estimate $\boldsymbol{\theta}$. We observed that we have more moment conditions than parameters and none of the moment conditions is a linear combinations of the previous moment conditions. Furthermore, we assume that the problem is identified, i.e. the system in (4.17) is satisfied only at the true value $\boldsymbol{\theta}_0$. These two conditions assure that the estimation problem is well defined, so that the system in (4.17) allows to find an estimate for the vector of parameters $\boldsymbol{\theta}$.
Let us look at the quantities involved in the estimation process. The central role is played by the objective function $Q(\theta)$ which depends on both the function $\boldsymbol{g}(x;\theta)$ and on the weighting matrix W. The function $\boldsymbol{g}(x;\theta)$ represents the difference between the theoretical expected values of the statistics and their sample counterparts. Thus, it follows from the system of equations (4.16) that:

$$\boldsymbol{g}(x(t_0), x(t_1); \theta) = E_\theta\left[\mathbf{S}\left(X(t_0), X(t_1)\right) | X(t_0) = x(t_0)\right] - \mathbf{s}(x(t_1), x(t_0))$$

Since we observed only one network, the sample counterpart of each statistic is exactly equal to the observed value. Then, we can write the objective function in the following manner:

$$Q(\theta) = \boldsymbol{g}(x(t_0), x(t_1); \theta)' W \boldsymbol{g}(x(t_0), x(t_1); \theta)$$

To obtain the GMM estimation for $\boldsymbol{\theta}$, one should compute the objective function $Q(\theta)$, and then he/she should minimize it. This procedure is not so straightforward in a network context since two orders of problems arise.
The first is related to the lack of an analytical form for $Q(\theta)$. In fact, the theoretical expected values of the statistics cannot be computed explicitly except for some special

and rather trivial cases (Snijders & Van Duijn, 1997), as already mentioned in Chapter 3.
The second problem concerns finding a consistent estimator for the weighting matrix W, since we observed just one network at each time point, and the classical estimation methods for variance and covariance cannot be applied.
The solution to both of these issues relies on simulations. Since the underlying network evolution process is a continuous-time Markov chain, it is rather straightforward to simulate the process evolution according to a desired distribution [7]. Therefore, a stochastic approximation method can be used to approximate the objective function $Q(\theta)$.
Stochastic approximation methods are an alternative to the stochastic search methods when the interest lies on optimizing an objective function, $h(\theta)$. In more detail, let us consider the optimization problem[8]:

$$\max_{\theta \in \Theta} h(\theta) \tag{4.18}$$

There are different approaches to solve the optimization problem in (4.18): numerical optimization, stochastic search and stochastic approximation methods (Fletcher, 1980; Spall, 2003; Bonnans *et al.*, 2006).
Numerical optimization methods are usually used when the analytical form of the function h is known so that a deterministic optimization algorithms can be applied to solve the optimization problem in (4.18). In the literature there are many deterministic algorithms. Some of them rely on gradient-based techniques, such as the Newton-Raphson method, which was described briefly in Paragraph 4.2.3.
Numerical optimization techniques are deterministic and suffer from the presence of local maxima or saddlepoints, when the function h is highly nonlinear, and the domain Θ is irregular. To deal with this drawback, stochastic methods can be applied. There are two different stochastic approaches which can be performed, namely stochastic search techniques and stochastic approximations.
Stochastic search techniques are rudimentary ways of using simulations to obtain an approximation to the solution of (4.18) based on exploration methods. The idea is to maximize the given function h by considering a random sequence of points. Specifically, the idea is to simulate a sequence of points $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n)$ over $\boldsymbol{\Theta}$ according to an arbitrary distribution ϕ and then to approximate the solution of the optimization problem (4.18) with $h(\widehat{\theta}) = \max\{h(\theta_1), \cdots, h(\theta_n)\}$. This solution may be very inefficient if ϕ is not chosen in connection with h, but under some regularity requirements, including the compactness of $\boldsymbol{\Theta}$, it is bound to converge. When $\boldsymbol{\Theta}$ is compact, one can simulate from a uniform distribution. A more efficient way is to choose ϕ so that the probability of simulating in regions where h is large is higher than the probability of simulating in regions where it is small. Thus, the distribution of ϕ is related to h in a non-linear way.
Stochastic search approaches quickly become impractical as the dimension of $\boldsymbol{\Theta}$ or the

[7] The description of the simulation algorithm will be given in Paragraph 5.1.1

[8] The optimization problem consists in finding a maximum of a function h. This case included also the case of a minimum search, since a minimization problem can be handled as a maximization problem when substituting $-h$ to h.

complexity of the problems increases, so that different stochastic searches were proposed in the literature. Among these approaches, there is a stochastic version of gradient methods, which consists in exploring the surface of h in a local manner, namely by defining a sequence $\{\boldsymbol{\theta}_j\}$ by moving from $\boldsymbol{\theta}_j$ to $\boldsymbol{\theta}_{j+1}$ in a dependent step. In contrast to the deterministic approach, the update does not proceed along the steepest slope of h in θ_j, but at each time the direction is picked at random. This may avoid the algorithm from being trapped in a local maxima or in saddlepoints of h.
The stochastic search methods concerned the exploration of the domain $\boldsymbol{\Theta}$, and it requires the knowledge of the functional form of the objective function which should be optimized. Therefore, these approaches are not useful in our context, since there is no way to write the objective function $Q(\theta)$ in an analytical form. An alternative to stochastic search approaches is represented by stochastic approximation, whose logic is first to approximate the objective function and then to maximize it. Obviously, the approximation introduces an additional level of error with respect to the stochastic search techniques, and this represents a drawback for the method. Furthermore, the approach is reliable when the objective function approximation converges almost surely to the exact objective function as the number of simulations increases. This nice property assures the adequacy of the stochastic approximation and it should be verified before performing such a method.
From the previous lines it follows that a stochastic approximation should be implemented to find the GMM estimator for the parameter of the Stochastic actor-oriented model. This suggests that one should proceed in two phases: the first aims to reliably approximate the objective function $Q(\theta)$, while the second to minimize the approximation.
Regarding the first step, the approximation of $Q(\theta)$ requires the evaluation of the expected values $E_{\boldsymbol{\theta}}\left[\mathbf{S}\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)\right]$ and the weighting matrix W, which are described in the next paragraphs.
Regarding the second step, the minimization involves a numerical approach, which should be determined so that computational efficiency will be assured. Since this choice is related to computational aspects, it will be one of the topics of the next chapter.
Before starting to describe how to approximate the quantities involved in the objective function, we note that the stochastic approximation requires a set of simulations, so that the estimate for $\boldsymbol{\theta}$ is not the GMM estimator but the GMSM estimator. Indeed, before describing the algorithm and the computational aspects, it is necessary to show that the stochastic approximation $\widehat{Q}(\theta)$ converges almost surely to $Q(\theta)$, so that the GMSM estimator converges to the GMM estimate.

4.3.1 The approximation of the conditional expected values

The computation of the theoretical expected values of the statistics involved in the estimation process can be solved noting that an expected value can be reduced to an integration problem. In the literature there are some useful ways to deal with integration problems. Among them, Monte Carlo (MC) and Importance sampling (IS) methods have a special role, since they are quite easy to apply and present very nice properties. Both procedures are based on simulations, and they differ in the distribution from which values

are generated.
In more detail, let us consider a probability density $\varphi(x)$ of a random variable X, assume we are interested in estimating the expected value:

$$E[h(X)] = \int_{\mathcal{X}} h(x)\varphi(x)dx \tag{4.19}$$

where $\mathcal{X}$ is the support of X and $h : \mathcal{X} \to \mathbb{R}$. We assume that the expected values and the variance of the function $h(X)$ exist and are finite, namely:

$$E[h(X)] = \overline{h} < \infty \qquad Var[h(X)] = \sigma^2 < \infty$$

The principle of the Monte Carlo method (Liu, 2003; Robert & Casella, 2004, 2010) is quite simple. It suggests generating a sample $(X_1, \cdots, X_n)$ from the density $\varphi(x)$ and to approximate the expected value with the empirical average, i.e.:

$$\widehat{E}[h(X)] = \frac{1}{n}\sum_{j=1}^{n} h(x_j) \tag{4.20}$$

The MC estimator defined by the (4.20) shows desirable properties. In particular, using the Strong Law of Large Numbers, it can be proved that it almost surely converges to the true value $E[h(X)]$. Thus, the MC estimator is consistent. Furthermore, it is unbiased and normally distributed.
The MC method is applicable only when it is possible (and reasonably simple) to generate from the distribution $\varphi(x)$. When this does not happen, the Importance sampling (IS) method can be applied. The IS technique consists in simulating samples under an instrumental distribution $\psi(x)$, and then approximating the target distribution $\varphi(x)$ by weighting these samples using appropriately defined importance weights.
Let $\psi(x)$ be an arbitrary density, such that $h(x)\varphi(x)$ is dominated by $\psi(x)$ (that is $\psi(x) = 0$ implies that $h(x)\varphi(x) = 0$). Then an alternative way to write equation (4.19) is represented by the *fundamental IS identity*:

$$E_\varphi[h(X)] = \int_{\mathcal{X}} h(x)\varphi(x)dx = \int_{\mathcal{X}} h(x)\frac{\varphi(x)}{\psi(x)}\psi(x)dx = E_\psi\left[h(x)\frac{\varphi(x)}{\psi(x)}\right]$$

that is, equation (4.19) can be expressed as an expectation under the density $\psi(x)$. A requirement on the instrumental distribution $\psi(x)$ is that the support of $\psi(x)$ must include the support of $h \times \varphi$, i.e. $supp(\psi) \supseteq supp(h \times f)$. Since the expected value is still computed on the support of X, choosing a smaller support truncates the integral in (4.19), so that a biased result is produced.
The principle of IS method suggests generating a sample $(X_1, \cdots, X_n)$ from $\psi(x)$, and to compute the estimator for the considered expected value through the empirical weighted average:

$$\widehat{E}_\varphi[h(X)] = \frac{1}{n}\sum_{j=1}^{n} h(x_j)\frac{\varphi(x_j)}{\psi(x_j)} = \frac{1}{n}\sum_{j=1}^{n} h(x_j)\omega(x_j) \tag{4.21}$$

where the quantities $\omega(x_j)$ are called IS weights or sometimes are referred to likelihood ratios, so that the IS estimator is also known as the likelihood ratio estimator. These weights are used to correct the bias, introduced in the IS procedure by simulating from the instrumental distribution $\psi(x)$, and they provide a relative assessment of the adequacy of the generated sample to the target density. In particular, it is necessary to point out that each weight $\omega(x_j)$ can take any positive value, but this is not an "absolute" measure of how much more likely x_j is generated from $\varphi(x)$. In fact, if $\omega(x_j)$ assumes a high value, it does not mean that x_j is very likely to be generated by $\varphi(x)$, but simply that it is more likely than the other simulated values, so that the IS weights should be interpreted as a "relative" measure.
It is straightforward to observe that when the weights are all equal to one, the IS estimator in equation (4.21) corresponds to the classical MC estimator defined by equation (4.20), so that the IS method includes the MC estimator as a special case.
The estimator in (4.21) is unbiased and consistent, since it converges almost surely to $E[h(X)]$ by the Strong Law of Large Numbers. Regarding the variance of the IS estimator some problems arise. In more detail, if the instrumental distribution has tails lighter than those of the target distribution, the variance of the corresponding estimator will be infinite for many functions h. Thus, the instrumental distribution $\psi(x)$ must have thicker tails than those of the target distribution $\varphi(x)$.
An alternative estimator to (4.21) is the *self-normalized IS estimator*, which is useful when the target distribution or the instrumental distribution are known up to a constant. The self-normalized IS estimator is computed dividing the weights by their sum, so that the multiplicative constant can be reduced:

$$\widehat{E}_{\varphi}[h(X)] = \frac{\sum_{j=1}^{n} h(x_j)\omega(x_j)}{\sum_{j=1}^{n} \omega(x_j)} \tag{4.22}$$

It can be proved that even the estimator defined by equation (4.22) is consistent by the Strong Law of Large Numbers and although it is biased, the improvement in variance makes it a preferred alternative to (4.21) (Cappé *et al.*, 2005; Robert & Casella, 2004).
The IS estimator presents some advantages with respect to the MC method when the choice of the instrumental distribution is adequate. Although there are few requirements on the instrumental distribution (it should dominate the target distribution, and it should have thicker tails), the IS estimator suffers from the choice of $\psi(x)$. In fact, a wrong choice of the instrumental distribution can cause degeneracy problems (Rubinstein & Kroese, 2008).
In particular, the distribution $\omega(X)$ of the IS weights under the importance distribution $\psi(x)$ may become increasingly skewed as the dimensions of the generated sample $(X_1, \cdots, X_n)$ increased, i.e. $\omega(X)$ may take values close to 0 with high probability and large values with small but significant probability. This means that in the generated sample $(X_1, \cdots, X_n)$ there are few values that are more likely to be generated from $\varphi(x)$ than the other simulated values, and this introduces a bias in the IS estimator.

Since the weight degeneracy phenomenon has serious consequences, it is of great utility to find a way to diagnose and to prevent it. In the literature there are several indexes that allow to detect the presence of degeneracy (Cappé *et al.*, 2005). One of them is the "Shannon Entropy" or "perplexity" defined by the following expression:

$$Ent = -\sum_{j=1}^{n} \frac{\omega(x_j)}{\sum_{j=1}^{n} \omega(x_j)} \; log \left(\frac{\omega(x_j)}{\sum_{j=1}^{n} \omega(x_j)} \right) \tag{4.23}$$

The Shannon Entropy is maximal when the normalized weights are all equal to $1/n$, and then $Ent = log(n)$. The minimal value of the Shannon Entropy is 0, and this happens when one of the normalized weights takes the value 1 and all the others are null.
The idea to use Shannon entropy index to judge the adequacy of the instrumental distribution $\psi(x)$ derives from the relationship between this index and the Kullback-Leibler distance (also called relative entropy) (Cappé *et al.*, 2008). Specifically, the Kullback-Leibler distance between the target and the instrumental distributions is defined by the integral:

$$D_{KL} = \int log \frac{\varphi(x)}{\psi(x)} \varphi(x) dx$$

and it is related to the Entropy index through the following equation:

$$\frac{exp[Ent]}{n} = exp[-D_{KL}]$$

which allows to show that the distance is null when the index assumes its maximal value.
Since the weight degeneracy is detrimental, different methods to choose the instrumental distribution $\psi(x)$ were proposed in the literature. Two optimal choices, based on the minimization of a criterion function, were proposed by Rubinstein (Rubinstein & Kroese, 2008). In more detail, the involved criterion functions are the variance of the resulting estimator and the Kullback-Leibler distance; the resulting methods are called the "variance minimization method" and the "cross-entropy method" respectively. In both cases when the probability distribution function $\varphi(x)$ belongs to some parametric family of distribution, it is often convenient to choose the importance sampling distribution from the same family.
The MC and the IS methods are a solution to the lack of an analytical form of the expected values involved in the GMM estimation of the parameter of the SAO model. In fact, the definition of the model given in Chapter 3, can be used directly to simulate data from the probability distribution, which describes network evolution conditional on an initial state.
Let us consider the generic statistic $S_k\left(X(t_0), X(t_1) | X(t_0) = x(t_0)\right)$. We are interested in estimating its expected value, which can be expressed by the following integral:

$$E_{\boldsymbol{\theta}_a}[S_k\left(X(t_0), X(t_1)\right) | X(t_0) = x(t_0)] = \int_{\mathcal{X}} s_k\left(x(t_0), x(t_1) | X(t_0) = x(t_0)\right) \; \varphi(x(t_1)) dx(t_1) \tag{4.24}$$

where $\mathcal{X}$ is the set of all possible digraphs of g actors, φ is the probability distribution associated to the network evolution process, and $s_k : \mathcal{X} \to \mathbb{R}$ is a function that allows to compute the considered statistic. The equality in (4.19) is guaranteed by the definition of the conditional expected value and by the fact that we are conditioning with respect to $x(t_0)$. The conditions related to the finiteness of the expected values and the variances of the statistics are satisfied since the outcome space is finite.
The principle of the Monte Carlo method requires generating a sample $(X_1, \cdots, X_n)$ from the density φ and to approximate the expected value with the empirical average:

$$\widehat{E}_{\boldsymbol{\theta}}[S_k\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)] = \frac{1}{n}\sum_{j=1}^{n} s_k\left(x(t_0), x_j(t_1)|x(t_0)\right) \tag{4.25}$$

while the corresponding IS estimator is computed using the following weighted mean:

$$\widehat{E}_{\boldsymbol{\theta}}[S_k\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)] = \frac{1}{n}\sum_{j=1}^{n} s_k\left(x(t_0), x_j(t_1)|x(t_0)\right)\frac{\varphi(x_j(t_1))}{\psi(x_j(t_1))} \tag{4.26}$$

Since the distribution of the continuous-time process $\{X(t), t_1 < t < t_M\}$ for the SAO model belongs to the exponential family distributions, it is convenient to choose the instrumental distribution from the same exponential family. Thus, assuming that θ_0 is the true value of the parameter θ, and $\widetilde{\theta}$ is the value which adequately defines the instrumental distribution, the IS defines estimator in (4.26) can be written as:

$$\widehat{E}_{\boldsymbol{\theta}}[S_k\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)] = \frac{1}{n}\sum_{j=1}^{n} s_k\left(x(t_0), x_j(t_1)|x(t_0)\right)\frac{\varphi(x_j(t_1);\theta_0)}{\varphi(x_j(t_1);\widetilde{\theta})} \tag{4.27}$$

where $\varphi(x_j(t_1);\theta_0)$ and $\varphi(x_j(t_1);\widetilde{\theta})$ are the target and the instrumental distributions, respectively. Denoting the importance weights by $\omega(x_j(t_1)) = \dfrac{\varphi(x_j(t_1);\theta_0)}{\varphi(x_j(t_1);\widetilde{\theta})}$, the corresponding self-normalized estimator is:

$$\widehat{E}_{\boldsymbol{\theta}}[S_k\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)] = \frac{\sum_{j=1}^{n} s_k\left(x(t_0), x_j(t_1)|x(t_0)\right)\omega(x_j(t_1))}{\sum_{j=1}^{n}\omega(x_j(t_1))} \tag{4.28}$$

Equations (4.25), (4.27) and (4.28) identify three possible estimators for the expected value of the statistic S_k, but the IS estimator will be preferred to the MC estimator mainly for two reasons. The first one is that it usually provides smaller standard errors than those of the MC estimator. But this is not a sufficient criterion to choose between the two procedures, since computational efficiency should be also taken into account when simulation techniques are used. This explains the existence of the second reason. The IS method also saves computational time, since the same sample (generated from

$\varphi(x_j(t_1);\theta^*))$ can be used repeatedly, not only for different functions s_k but also for different densities $\varphi(x_j(t_1);\theta)$, a feature which is quite attractive for stochastic simulation algorithms. Furthermore, the self normalized IS estimator will be preferred to the naive IS estimator because it shows a lower variance.
Having defined how to approximate the expected values involved in the estimation process, we can now focus our attention on the estimation of the weighting matrix W.

4.3.2 The estimation of the weighting matrix W

The other quantity which plays an important part in the determination of the objective function $Q(\theta)$ is the weighting matrix W. The role of this matrix was described in the previous paragraph. W is a positive definite matrix which determines the identifying and the over-identifying restrictions, namely it attributes importance to each moment condition implied in the estimation process. It is also involved in the computation of the standard error of the GMM estimator, therefore the choice of W can influence the efficiency of the corresponding estimator.
The last remark suggests the importance of determining the variance of the GMM estimator in order to establish the structure of the weighting matrix W, so that efficiency is assured. Then, the estimation method should be defined subsequently.
In order to determine the efficient weighting matrix W a similar approach to that used by Hall (Hall, 2005) or by Gourieroux and Monfort (Gourieroux & Monfort, 1996) in the context of the GMSM can be exploited, taking into account the peculiarities of network data. In particular, we define the asymptotic normality distribution of the GMSM estimator, and then we look for the estimate of W, which minimizes the asymptotic variance of the GMSM estimator.
To determine the asymptotic distribution of the estimator, let $\widehat{\boldsymbol{\theta}}_{GMSM}$ and $\boldsymbol{\theta}_0$ be the GMSM estimator and the true value for θ, respectively. We focus on the quantity $\boldsymbol{g}(x(t_0), x(t_1); \widehat{\boldsymbol{\theta}}_{GMSM})$, which is estimated using the IS method on a sample size of n simulations, as described in the previous paragraph. We can approximate this quantity using an expansion of the first order conditions around $\boldsymbol{\theta}_0$:

$$\mathbf{g}(x(t_0), x(t_1); \widehat{\theta}_{GMSM}) = \mathbf{g}(x(t_0), x(t_1); \theta_0) + \Gamma\left(\widehat{\theta}_{GMSM}\right)(\widehat{\boldsymbol{\theta}}_{GMSM} - \boldsymbol{\theta}_0) \tag{4.29}$$

The premultiplication of (4.29) by $\sqrt{n}\Gamma(\widehat{\theta}_{GMSM})^T W$ yields:

$$\sqrt{n}\Gamma(\widehat{\boldsymbol{\theta}}_{GMSM})^T W \mathbf{g}(x(t_0), x(t_1); \widehat{\theta}_{GMSM} =$$

$$= \Gamma(\widehat{\theta}_{GMSM})^T W \sqrt{n}\mathbf{g}(x(t_0), x(t_1); \theta_0) + \Gamma(\widehat{\theta}_{GMSM})^T W \Gamma(\theta_{GMSM})\sqrt{n}(\widehat{\boldsymbol{\theta}}_{GMSM} - \boldsymbol{\theta}_0) \tag{4.30}$$

Since $\boldsymbol{\theta_{GMSM}}$ is the $GMSM$ estimator, the first order conditions in (4.16) imply that the left hand site of (4.30) is 0. So with some rearrangement it follows that:

$$0 = \Gamma(\widehat{\theta}_{GMSM})^T W \sqrt{n}\mathbf{g}(x(t_0), x(t_1); \theta_0) + \Gamma(\widehat{\theta}_{GMSM})^T W \Gamma(\widehat{\theta}_{GMSM})\sqrt{n}(\widehat{\boldsymbol{\theta}}_{GMSM} - \boldsymbol{\theta}_0) \tag{4.31}$$

or equivalently:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{GMSM} - \boldsymbol{\theta}_0) = \left[\Gamma(\widehat{\theta}_{GMSM})^T W \Gamma(\widehat{\theta}_{GMSM})\right]^{-1} \Gamma(\widehat{\theta}_{GMSM})^T W \sqrt{n}\mathbf{g}(x(t_0), x(t_1); \theta_0) \tag{4.32}$$

Equation (4.32) suggests that the distribution of $(\widehat{\theta}_{GMSM} - \theta_0)$ is related to the distribution of $\mathbf{g}(x(t_0), x(t_1); \theta_0)$. This quantity is the sample counterpart of $\boldsymbol{f}$, and it is approximated using the IS method. The IS estimator is asymptotically normal (Geweke, 2005; Cappé *et al.*, 2005), i.e.

$$\sqrt{n}\left(\widehat{E}_{\boldsymbol{\theta}}[\mathbf{S}] - \mathbf{s}\right) \xrightarrow{d} N(0, \Sigma_g)$$

where Σ_g is the variance-covariance matrix of the IS estimator and it is defined as:

$$\Sigma_g = \int \omega^2(x_j(t-1))\ [\mathbf{S}_i - \mathbf{s}_i]^2\ \varphi(x_j(t-1))\ dx_j(t-1)$$

Let us assume that $\Gamma(\widehat{\theta}_{ISn})$ is a consistent estimator for Γ as n increases, so that,

$$\left[\Gamma(\widehat{\theta}_{GMSM})^T W \Gamma(\widehat{\theta}_{GMSM})\right]^{-1} \Gamma(\widehat{\theta}_{GMSM})^T W \xrightarrow{p} \left[\Gamma^T W \Gamma\right]^{-1} \Gamma^T W$$

The Slutsky theorem [9] combined with the properties related to the linear combination of the Gaussian distribution yields:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{GMSM} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma)$$

where

$$\Sigma = \left(\Gamma^T W \Gamma\right)^{-1} \Gamma^T W \Sigma_g \left[\left(\Gamma^T W \Gamma\right)^{-1} \Gamma^T W\right]^T \tag{4.33}$$

From equation (4.33) it follows that the optimal choice for the weighting matrix W is the inverse of the variance-covariance matrix of $\boldsymbol{g}$, i.e. $W = (\Sigma_g)^{-1}$, so that:

$$\Sigma = \left(\Gamma^T W \Gamma\right)^{-1} \tag{4.34}$$

Consequently, finding a consistent estimator for W is equivalent to determining a consistent estimator for Σ_g. This estimator was proposed by Geweke (Geweke, 2005) and Cappé (Cappé *et al.*, 2005) and it is given by:

$$\widehat{\Sigma}_g = \frac{n \sum_{j=1}^{n} \omega^2(x_j(t_1))\ [\mathbf{s}(x_j(t_1)) - \overline{\mathbf{s}}]^2}{\left(\sum_{j=1}^{n} \omega(x_j(t_1))\right)^2} \tag{4.35}$$

[9] The Slutsky theorem claims that if $\{X_n\}$ is a sequence of random variables which convergences in distribution to the random variable X, and $\{Y_n\}$ is another sequence of random variables which converges in probability to the constant $c \in \mathbb{R}$, then the sequence $\{X_n, Y_n\}$ converges in distribution to cX. The theorem can also be generalized to multivariate random variables.

This implies that W can be estimated as the inverse of $\widehat{\Sigma}_g$. Thus, a consistent estimator for W is:

$$\widehat{W} = \left[\frac{n \sum_{j=1}^{n} \omega^2(x_j(t_1)) \left[\mathbf{s}(x_j(t_1)) - \bar{\mathbf{s}}\right]^2}{\left(\sum_{j=1}^{n} \omega(x_j(t_1)) \right)^2} \right]^{-1} \tag{4.36}$$

At this point we can approximate the objective function $Q(\theta)$ with $\widehat{Q}(\theta)$ using the IS estimator for the expected values $E_{\boldsymbol{\theta}}\left[\mathbf{S}\left(X(t_0), X(t_1)\right) | X(t_0) = x(t_0)\right]$ and the consistent estimator of the weighting matrix W. Then, an estimate for $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}_{GMSM}$ such that:

$$\widehat{\boldsymbol{\theta}}_{GMSM} = \arg\min_{\theta \in \Theta} \widehat{Q}(\theta)$$

4.3.3 The logic behind the GMSM estimator

The principle of stochastic approximation method consists in approximating the objective function $Q(\theta)$ through simulations and then in optimizing it. This approximation will be denoted by $\widehat{Q}_n(\theta)$, where the subscript n indicates that the objective function is approximated using a sample of n simulations.
Stochastic approximation idea works well if the approximated objective function $\widehat{Q}(\theta)$ converges to the real objective function $Q(\theta)$, so that the approximated estimator, which was denoted by $\widehat{\boldsymbol{\theta}}_{GMSM}$ converges almost surely to the GMM estimate $\boldsymbol{\theta}_{GMM}$.
In our case it is not difficult to show that $\widehat{Q}_n(\theta)$ converges almost surely to $Q(\theta)$. Let us consider the approximates objective function:

$$\widehat{Q}_n(\theta) = \left[\frac{\sum_{j=1}^{n} \mathbf{s}(x(t_0), x_j(t_1)|x(t_0))\omega(x_j(t_1))}{\sum_{j=1}^{n} \omega(x_j(t_1))} - \mathbf{s}\left(x(t_0), x(t_1)\right) \right]^T \widehat{W}$$

$$\left[\frac{\sum_{j=1}^{n} \mathbf{s}(x(t_0), x_j(t_1)|x(t_0))\omega(x_j(t_1))}{\sum_{j=1}^{n} \omega(x_j(t_1))} - \mathbf{s}(x(t_0), x(t_1)) \right]$$

Which is the product of the approximations of $\boldsymbol{g}(x(t_0), x(t_1))$ and of the weighting matrix W, defined in (4.27) and (4.36). When the number n of simulations increases, the IS estimator converges almost surely to the real expected values [10], i.e.

$$\frac{\sum_{j=1}^{n} \mathbf{s}\left(x(t_0), x_j(t_1)|x(t_0)\right) \omega(x_j(t_1))}{\sum_{j=1}^{n} \omega(x_j(t_1))} \overset{a.s.}{\to} E_\theta[\mathbf{S}\left(X(t_0), X(t_1)|X(t_0) = x(t_0)\right)]$$

[10] This result follows from the Strong Law of Large Numbers, and the proof can be found in the literature (Robert & Casella, 2004)

Thus:

$$\widehat{\mathbf{g}}_n(X(t_1), X(t_0)) \xrightarrow{a.s.} \mathbf{g}(X(t_1), X(t_0); \theta)$$

Regarding the weighting matrix W, Geweke (Geweke, 2005) proved that the quantity in equation (4.35) converges almost surely to Σ_g. This yields

$$\widehat{W}_n \xrightarrow{a.s.} W$$

At this point the Continuous Mapping Theorem can be used to prove the almost sure convergence of the approximated objective function.

Theorem 4. *(**Continuous Mapping Theorem**) Let X_n a sequence of random variables, $f : \mathbb{R} \to \mathbb{R}$ be a continuous function, and X be a random variable.*

i) if $X_n \xrightarrow{a.s.} X$ then $f(X_n) \xrightarrow{a.s.} f(X)$
ii) if $X_n \xrightarrow{p} X$ then $f(X_n) \xrightarrow{p} f(X)$
iii) if $X_n \xrightarrow{d} X$ then $f(X_n) \xrightarrow{d} f(X)$

The continuous mapping theorem can be easily generalized for the sequence of multivariate random variables, so that it can be also applied in this case. In particular, if we consider the two sequences $\widehat{\mathbf{g}}_n(X(t_1), X(t_0); \theta)$ and W_n and we define the function f in the following way:

$$\boldsymbol{f}(\widehat{\mathbf{g}}_n(X(t_1), X(t_0); \theta), \widehat{W}_n) = \widehat{\mathbf{g}}_n(X(t_1), X(t_0))^T \ \widehat{W}_n \ \widehat{\mathbf{g}}_n(X(t_1), X(t_0))$$

from the continuous mapping theorem it follows that

$$\widehat{\mathbf{g}}_n(X(t_1), X(t_0))^T \ \widehat{W}_n \ \widehat{\mathbf{g}}_n(X(t_1), X(t_0)) \xrightarrow{a.s.} \mathbf{g}(X(t_1), X(t_0))^T \ W \ \mathbf{g}(X(t_1), X(t_0))$$

Now we can prove that $\widehat{\boldsymbol{\theta}}_{GMSM} \xrightarrow{a.s.} \boldsymbol{\theta}_{GMM}$, which follows directly from the convexity of $\widehat{Q}_n(\theta)$ and $Q(\theta)$. These two functions are positive definite quadratic forms, since W is a positive definite matrix[11]. Therefore, $\widehat{Q}_n(\theta)$ and $Q(\theta)$ are strictly convex functions, and they admit a unique minimum. The almost sure convergence of $\widehat{Q}_n(\theta)$ to $Q(\theta)$ implies the almost sure convergence of their minimum, so that:

$$\widehat{\boldsymbol{\theta}}_{GMSM} \xrightarrow{a.s.} \boldsymbol{\theta}_{GMM}$$

Thus, the approximation $\widehat{Q}_n(\theta)$ proposed for the objective function $Q(\theta)$ is adequate and can be used in the stochastic approximation.
Before describing how we can practically obtain the $\widehat{\boldsymbol{\theta}}_{GMSM}$ estimator, i.e. the stochastic algorithm, one remark should be emphasized. In this paragraph we proved that $\widehat{\boldsymbol{\theta}}_{GMSM} \xrightarrow{a.s.} \boldsymbol{\theta}_{GMM}$. This does not mean that $\boldsymbol{\theta}_{GMSM}$ converges almost surely to the true value θ_0 of the parameter, but that it converges almost surely to the GMM estimate $\boldsymbol{\theta}_{GMM}$. Consistency is a desirable asymptotic property which describes the behavior of the estimator when the sample size increases. In our case we observed only one network over time, thus proving consistency is not particularly useful.

[11] Every positive definite matrix is invertible, and its inverse is also a positive definite (Horn & Johnson, 1990). Thus, W is a positive definite matrix, since it is the inverse of a variance-covariance matrix, which by definition is positive definite.

Chapter 5

Stochastic approximation algorithm

In Chapter 4 the definition of new statistics and the estimation method were defined, also specifying the need for a stochastic approximation in order to compute the estimate for the parameter of the SAO model.
In this chapter practical aspects, such as the implemented algorithm and the obtained results, are presented focusing on the main aspects. [1].

5.1 The algorithm

The aim of the algorithm is to approximate the solution of the system (4.17). The solution is represented by the true value θ_0 for the parameter θ, and its approximation will be denoted by $\widehat{\theta}$, neglecting the subscript $GMSM$, which refers to the estimation method used. From Chapter 4, it follows that

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \widehat{Q}(\theta)$$

where the objective function $Q(\theta)$ is approximated by stochastic simulations, which leads to a Monte Carlo optimization problem.
Several algorithms for Monte Carlo optimization are proposed in the literature (Cappé *et al.*, 2005) as alternative techniques to naive grid search methods. Most of them are based on the optimization of the likelihood function but can be generalized to any objective function. In particular, Geyer and Thompson in 1992 (Geyer & Thompson, 1992) introduced the first gradient-based algorithm which uses Monte Carlo simulations in order to optimize the likelihood function of an exponential family distribution parametrized by the parameter θ. They suggested approximating the likelihood function $\ell(\theta)$ using Monte Carlo methods. The estimation is obtained through an iterative algorithm which required to choose an initial value $\widetilde{\theta}$, to generate a sample from the instrumental distribution parametrized by $\widetilde{\theta}$ and to maximize the following likelihood function. The result

[1] To simplify notation the bold character for denoting vectors is neglected in this chapter

is a new point θ^* in the parameter space, which can be used to iterate the procedure till convergence.
It can be proved that the Monte Carlo approximation converges almost surely to the exact maximum likelihood estimation as the number of Monte Carlo simulations increases. This is a nice property that assures the adequacy of the algorithm.
The implementation of the algorithm proposed by Geyer and Thompson is not confined to the optimization of a likelihood function deriving from a distribution belonging to the exponential family, but it can be generalized to the optimization of any function which involves an integral. Gelfand and Carlin (Gelfand & Carlin, 1993) extended this procedure to a broad class of constrained or missing-data models and suggested that IS can be used instead of the MC approach to reduce computational efficiency.
Nearly at the same time Gelman (Gelman, 1995) proposed a similar algorithm for the regular MoM in which importance sampling is used as the Monte Carlo method. He proposed "an iterative approach to matching moments using a numerical equation-solving algorithm applied to Monte Carlo estimates of moments and their derivatives", which deals well with our problem.
Let $\varphi(x;\theta)$ be a family of distributions, whose support is $\mathcal{X}$, and we are interested in estimating the p-dimensional parameter $\theta \in \Theta$ using the MoM:

$$\mu(\theta) - \overline{\mu} = 0 \tag{5.1}$$

where $\mu(\theta) = E[f(x;\theta)]$ is the p-dimensional vector of the theoretical moments, $\overline{\mu} = \overline{f}(x;\theta)$ is its sample counterpart, and f is a real valued function defined on $\mathcal{X}$. If $\mu(\theta)$ can be expressed in a closed form and the system in (5.1) is complex to solve, the MoM estimation θ can be computed using the Newton-Raphson (NR) algorithm, as follow: the initial value θ_t is iteratively updated by the step:

$$\theta_{t+1} = \theta_t + [\mu^{'}(\theta_t)]^{-1}(\mu(\theta) - \mu(\theta_t))$$

where $\mu^{'}(\theta)$ is the matrix of derivative of $\mu(\theta)$.
When $\mu(\theta)$ cannot be expressed in an analytical form, Gelman proposed a stochastic approximation. In particular, he suggested approximating the theoretical expected values using Monte Carlo simulation. If $(x_1, \cdots, x_n)$ is a set of n simulated values drawn from the distribution of $\varphi(x;\theta)$, then from Chapter 4, it follows that the MC estimator can be approximated by:

$$\widehat{\mu}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \tag{5.2}$$

and the corresponding first order derivatives matrix by [2]:

$$\widehat{\mu}^{'}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{\partial}{\partial\theta}\ log\varphi(x_i;\theta) \tag{5.3}$$

Having defined these quantities, one can specify the MC algorithm. First a set of N_t values of X is drawn from the distribution $\varphi(x;\theta_t)$, given the current guess θ_t. Then the NR optimization step is performed using (5.2) and (5.3). At each step the number of simulations should be increased, so that the MC error approaches 0 as the NR algorithm approaches convergence.
The drawback of this algorithm relies on the fact that a new set of simulations must be drawn at each simulation step, a procedure too cumbersome in terms of computational time.
For this reason, Gelman proposed to use importance sampling. Thus, having drawn a set of simulated values from an instrumental distribution $\psi(x)$, the expected values and their first order derivatives can be approximated by:

$$\begin{aligned}\widehat{\mu}(\theta) &= \tfrac{1}{n}\sum_{i=1}^{n} f(x_i)\tfrac{\varphi(x)}{\psi(x)} \\ \widehat{\mu}^{'}(\theta) &= \tfrac{1}{n}\sum_{i=1}^{n} f(x_i)\tfrac{\partial}{\partial\theta}\ log\varphi(x_i;\theta)\tfrac{\varphi(x)}{\psi(x)}\end{aligned} \tag{5.4}$$

These estimates will be useful as long as $\psi(x)$ is close to $\varphi(x)$.
As already mentioned, IS techniques present the advantage that the same sample can be used repeatedly so that computational burden can be reduced. Exploiting this property of the IS methods, Gelman defined a less time consuming algorithm which is based on *optimization steps* and *simulation steps*. In more detail, an optimization step is a single step of the NR algorithm based on a single set of simulations, while a simulation step is defined as a new set of simulated values from the current guess of θ.
Indeed, the idea is to perform several optimization steps after each simulation step. "A single optimization step will generally not be enough, since if the initial guess is far from the actual moments estimate, the importance ratios in the estimates (5.4) will become too variable as θ_t moves from its initial guess. However, few simulation steps, each followed by several optimization steps, should bring the estimate θ_t close enough to the goal that further importance ratios will be well behaved. At this point one can sample a large number n draws from $psi(x;\theta_t)$ and iterate NR to approximate convergence"

[2] Under regularity conditions it is easy to prove:

$$\mu^{'}(\theta) = \tfrac{\partial}{\partial\theta}E[f(x;\theta)] = \tfrac{\partial}{\partial\theta}\int f(x)\varphi(x;\theta)dx = \int f(x)\tfrac{\partial}{\partial\theta}\varphi(x;\theta)dx = \int f(x)\varphi(x;\theta)\underbrace{\frac{1}{\varphi(x;\theta)}\frac{\partial}{\partial\theta}\varphi(x;\theta)}_{\frac{\partial}{\partial\theta}log\varphi(x;\theta)}dx =$$

$$= \int f(x)\tfrac{\partial}{\partial\theta}log\varphi(x;\theta)\varphi(x;\theta)dx = E[f(x)\tfrac{\partial}{\partial\theta}log\varphi(x;\theta)]$$

(Gelman, 1995).
The idea of performing several optimization steps after each simulation step can be applied in our context to determine the estimation for the parameter θ and to minimize the objective function $Q(\theta)$. This is the approach used to compute the GMSM estimator introduced in Chapter 4.
In particular, the proposed algorithm has a similar structure of the Robbins-Monro algorithm briefly sketched in paragraph 3.3.1. It is divided into three phases: the first phase provides the initial values of the algorithm and the initial estimate for the weighting matrix W; the second phase implements Gelman's algorithm and is divided into two sub-phases. The former iteratively approximates the objective function $Q(\theta)$, using an exiguous number of simulations, and performing several NR steps after a simulation step. The latter is based on a large number of simulations and on the iteration of the optimization steps till a convergence criterion is satisfied; the third phase evaluates the goodness of fit of the model and the standard error of the estimates.
In the following pages the algorithm will be described in detail, defining first the simulation process and then the three phases.

5.1.1 Simulation

The stochastic approximation requires simulating network evolution through two consecutive time points. It is convenient to construct the continuous-time Markov chain underlying the network evolution process as the combination of its holding times, which defines the change opportunity process and its jump process, regarding the change determination process (Snijders & Van Duijn, 1997). As already mentioned in Chapter 3, the g actors are acting independently given the current state of the network. Assuming that each actor has the same individual change rate λ, the holding times between consecutive changes have a negative exponential distribution with parameter $g\lambda$, so that the expected weighting time between two consecutive time changes is $1/g\lambda$. Thus, the probability that an actor i has the opportunity to change his outgoing ties at a given time points is equal to $1/g$, i.e. each actor has the same probability to change under the assumption that each actor has the same rate of change.
The probability that node i changes his outgoing tie towards node j is given by the multinomial logit expression:

$$p_{ij}(\beta; x(i \rightsquigarrow j)) = \frac{e^{f_i(\beta, x(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_i(\beta, x(i \rightsquigarrow h))}} \tag{5.5}$$

Thus, the simulation algorithm can be described as follows:

1. Set the time $t = 0$ and $x = x(t_0)$

2. Generate the holding time dt according to a negative exponential distribution with parameter $g\lambda$

3. Select randomly the actor $i \in \mathcal{G}$, who makes the changes, with constant probability $\frac{1}{g}$
4. Select randomly the actor $j \in \mathcal{G}$, to whom i changes his outgoing tie, with the probability given by (5.5).
5. Set the time $t = t + dt$ and $x = x(t_0)(i \rightsquigarrow j)$
6. Repeat step 2. to 5. until $t = 1$

5.1.2 Phase 1

The stochastic approximation $\widehat{Q}(\theta)$ of the objective function requires the generation of a set of simulations from the process that describes network evolution. This implies that an initial value $\widetilde{\theta}$ for the parameter θ should be established. There are many studies in the algorithm literature that prove the importance of the role played by the initial value (Spall, 2003; Bonnans *et al.*, 2006; Antoniou & Lu, 2007). The choice should be made very carefully because it influences the convergence and the computational burden of the algorithm.
In our context the initial point is given by the parameter estimation obtained from the Robbins-Monro algorithm, already implemented in RSiena, using only one or two sub-phases during phase 2. This value, denoted by $\widetilde{\theta}$, allows to simulate network evolution and to find a first approximation of the weighting matrix W.
From Paragraph 4.3.2, it follows that a consistent estimator for W is given by formula (4.36) that requires the computation of the simulated values for the considered statistics and the IS weights. Once we have simulated the network process, it is straightforward to compute the simulated values for the statistics using the formulae described in the previous two chapters. The major problem in this context is related to the computation of the IS weights.
In more detail, the IS weights are defined as the ratio between the likelihood function based on the target distribution and that of the instrumental distribution. Thus, the likelihood function $L(\theta)$ should be computed. We suppose that it is known what exactly happens between two consecutive observational points, t_0 and t_1, i.e. each micro-step. Denoting the number of micro-steps between two consecutive observational points by R, we can express the likelihood function as the product of the probability of exactly having R changes times the probability of exactly observing the sequence of changes which led to the network observed at time t_1. Under the assumption that the rate of changes are constant, the number of micro-steps has a Poisson distribution with parameter $g\lambda$. Thus, the probability of exactly recording R micro-steps between t_0 and t_1 is

$$P(\#micro - steps = R) = e^{-g\lambda}\frac{(g\lambda)^R}{R!}$$

In Chapter 3 the following assumptions were stated: the changing network is the outcome of a Markov process, only one tie can change at any moment, and ties depend only sequentially via the changing configurations of the whole network. According to them,

the probability of observing the sequence of changes that lead to the network observed at time t_1 can be computed as the product of the probabilities given by (5.5). Thus, the likelihood function can be computed as follows:

$$L(\theta) = e^{(-g\lambda)} \frac{(g\lambda)^R}{R!} \prod_{r=1}^{R} \frac{1}{g} \frac{e^{f_{ri}(\beta, x_{(r-1)}(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_{ri}(\beta, x_{(r-1)}(i \rightsquigarrow h))}} \tag{5.6}$$

We can now determine the likelihood ratio. The denominator of the ratio is given by the likelihood function computed according to the parameter value $\widetilde{\theta}$ of the instrumental distribution, while the numerator by the likelihood function is computed according to the parameter value θ^* of the target distribution, i.e.

$$\begin{aligned} \omega(x(i \rightsquigarrow j)) = \frac{L(\theta^*)}{L(\widetilde{\theta})} &= \frac{e^{(-g\lambda^*)} \frac{(g\lambda^*)^R}{R!} \prod_{r=1}^{R} \frac{1}{g} \frac{e^{f_{ri}(\beta^*, x_{(r-1)}(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_{ri}(\beta^*, x_{(r-1)}(i \rightsquigarrow h))}}}{e^{(-g\widetilde{\lambda})} \frac{(g\widetilde{\lambda})^R}{R!} \prod_{r=1}^{R} \frac{1}{g} \frac{e^{f_{ri}(\widetilde{\beta}, x_{(r-1)}(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_{ri}(\widetilde{\beta}, x_{(r-1)}(i \rightsquigarrow h))}}} \\ &= \frac{e^{(-g\lambda^*)} \prod_{r=1}^{R} \frac{e^{f_{ri}(\beta^*, x_{(r-1)}(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_{ri}(\beta^*, x_{(r-1)}(i \rightsquigarrow h))}}}{e^{(-g\widetilde{\lambda})} \prod_{r=1}^{R} \frac{e^{f_{ri}(\widetilde{\beta}, x_{(r-1)}(i \rightsquigarrow j))}}{\sum_{h=1}^{g} e^{f_{ri}(\widetilde{\beta}, x_{(r-1)}(i \rightsquigarrow h))}}} \end{aligned} \tag{5.7}$$

where f_{ri} is the objective function computed at the $r-th$micro-step and $x_{(r-1)}$ is the network deriving from the $(r-1)-th$ micro-step.
It clearly appears that in Phase 1 even the value for the target distribution should be initialized carefully. In order for the likelihood ratio $\omega(x)$ never to be extremely large, which would make the standard deviation of the IS estimator too large, the two parameters θ^* and $\widetilde{\theta}$ must be close to each other, and this can be mathematically described by saying that their difference expressed in standard errors should be small. This could be best expressed by the Mahalanobis distance (Mahalanobis, 1936; De Maesschalck *et al.*, 2000), since the correlation between the estimators for the elements within θ can be important.
In our context, the Mahalanobis distance (MD) is defined by the following formula:

$$MD(\widetilde{\theta}, \theta^*) = \sqrt{(\widetilde{\theta} - \theta^*)\Sigma_{\widetilde{\theta}}(\widetilde{\theta} - \theta^*)^T} \tag{5.8}$$

where $\Sigma_{\widetilde{\theta}}$ is the variance-covariance matrix that takes into account the correlation in the data, since it is computed using the inverse of the variance-covariance matrix of

$\widetilde{\theta}$. Equation (5.8) shows that the MD takes into account the correlation among the estimators of the components of the vector of parameter θ, since it is calculated using the inverse of the variance-covariance matrix of the initial value $\widetilde{\theta}$ related to the instrumental distribution.
In practice, this means that we look for the value θ^* which is close to $\widetilde{\theta}$ according to the metric given by the Mahalanobis distance. Different trials suggest that the distance between the two vectors $\widetilde{\theta}$ and θ^* should be less than 0.1, otherwise the distribution of the normalize IS weights tends to degenerate. This also assures that the Shannon Entropy index is close to its maximum.
Figure 5.1 shows the histogram of normalized IS weights computed on a set of 300 simulations and based on two values $\widetilde{\theta}$ and θ^* whose Mahalanobis distance is less than 0.1. It clearly appears that the distribution of the normalized IS weights is centered on the value $1/300 = 0.0333$ and it is symmetric around this value, suggesting that there is no degeneracy. Furthermore, the Entropy index assume the value 5.699 which is very close to its maximum ($\log(300) = 5.704$).
The determination of θ^* can be described in the following steps:

1.1) Generate the components of θ^* according to a Uniform distribution whose extremes are defined by half the standard error of each estimators :

$$\lambda^* \sim U(\widetilde{\lambda} - s.e.(\widetilde{\lambda})/2, \widetilde{\lambda} + s.e.(\widetilde{\lambda})/2)$$

$$\beta_1^* \sim U(\widetilde{\beta_1} - s.e.(\widetilde{\beta_1})/2, \widetilde{\beta_1} + s.e.(\widetilde{\beta_1})/2)$$

$$\beta_2^* \sim U(\widetilde{\beta_2} - s.e.(\widetilde{\beta_2})/2, \widetilde{\beta_2} + s.e.(\widetilde{\beta_2})/2)$$

$$\beta_3^* \sim U(\widetilde{\beta_3} - s.e.(\widetilde{\beta_3})/2, \widetilde{\beta_3} + s.e.(\widetilde{\beta_3})/2)$$

1.2) If

$$MD(\widetilde{\theta}, \theta^*) = \sqrt{(\widetilde{\theta} - \theta^*)\Sigma_{\widetilde{\theta}}(\widetilde{\theta} - \theta^*)^T} < 0.1$$

then $\theta^* = (\lambda^*, \beta_1^*, \beta_2^*, \beta_3^*)$ otherwise repeat point 1.

Having defined these quantities, a small set of $n = 200$ simulations can be performed. Since the process is simulated using the algorithm in Paragraph 5.1.1 and the micro-steps are properly stored in an R object, we have all the necessary information to compute the importance weights and the weighting matrix W according to (5.7) and (4.36), respectively.

5.1.3 Phase 2

The aim of the second phase is to approximate the estimate θ_{GMM} using the approach described by Gelman in 1995 (Gelman, 1995). The algorithm suggests to use a small set of n_1 simulations to bring the estimate $\widetilde{\theta}$ close enough to the goal that the IS weights will be well behaved.

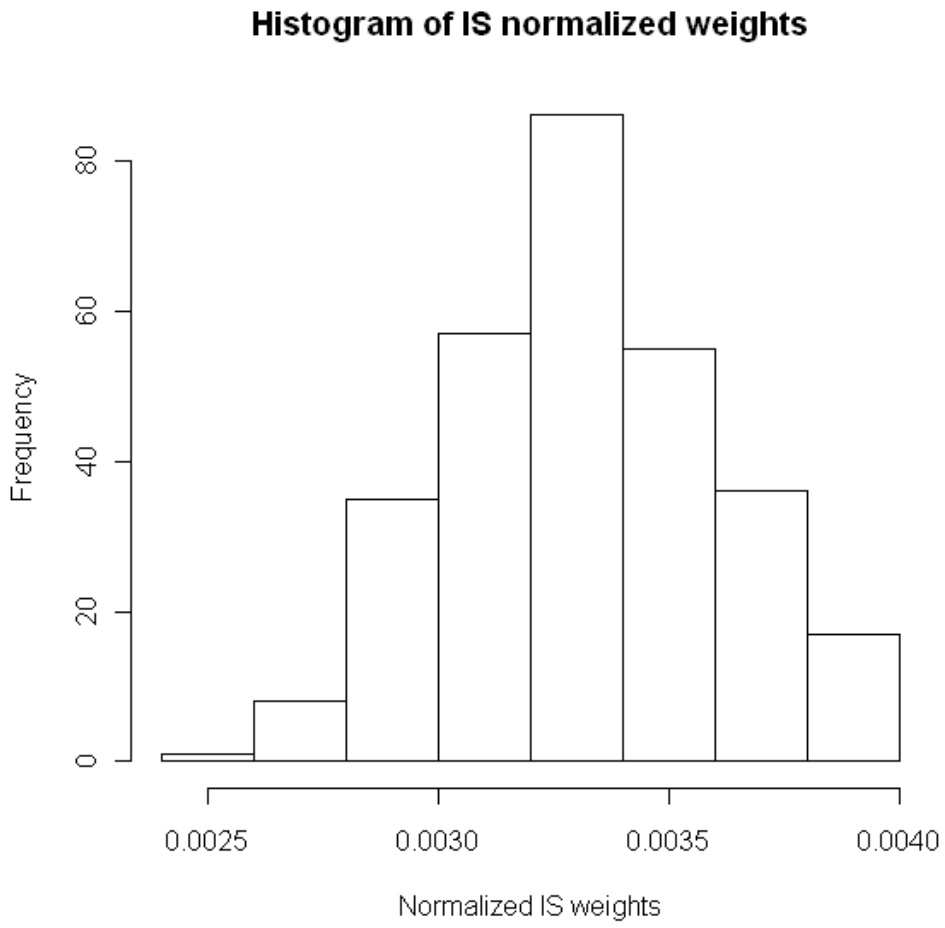


Figure 5.1: *Normalized IS weights computed on a set of 300 simulations from the values* $\widetilde{\theta} = (2.004, -2.623, 2.454, 0.222)$ *and* $\theta^* = (2.010, -2.636, 2.448, 0.233)$ *which have a Mahalanobis distance less than 0.1 (0.085)*

In more detail, using the IS technique to approximate the theoretical expected values of the considered statistics, we have to face the problem of IS weights degeneracy. Intuitively, if we are far from the goal θ_{GMM}, we should expect that the IS weights vary too much, even if the number of NR steps is small. Thus, the idea is to use a small set of simulations followed by few NR steps, saying `nstep`, to bring our estimate close to the goal. Then, a large number of simulations n_2 is drawn so that NR steps approximate convergence.
Therefore, the second phase of the algorithm is divided into two sub-phases that trace out the scheme previously described.

Sub-phase 1

In this sub-phase a small set of n_1 simulations is performed to bring the estimate $\widetilde{\theta}$ close enough to the goal so that the IS weights well-behaved and the NR steps can be performed several times to reach convergence. In order to do that, the set of simulations is also used to update the weighting matrix and the expected values estimation. The result is that the objective function Q is approximated step by step and the value $\widetilde{\theta}$ approaches the GMM estimate θ_{GMM} so that a bigger set of simulations can be drawn. To achieve this purpose the following procedure is used.

2.1.1) *Approximate the expected values.*
The expected values involved in the estimation process are approximated using the self-normalized IS estimator, whose IS weights are defined in equation (5.7). In more detail, given the current value of the parameters θ^* and $\widetilde{\theta}$, and applying formula (4.28), the theoretical expected value of each statistic is computed in the following way:

$$\widehat{E}[S_k\left(X(t_0), X(t_1)\right)|X(t_0) = x(t_0)] = \frac{\sum_{j=1}^{n_1} s_k\left(x(t_0), x_j(t_1)|x(t_0)\right) \frac{L(\theta^*, x_j(t_1))}{L(\widetilde{\theta}, x_j(t_1))}}{\sum_{j=1}^{n_1} \frac{L(\theta^*, x_j(t_1))}{L(\widetilde{\theta}, x_j(t_1))}}$$

2.1.2) *Estimate the first and the second order derivatives of th objective function Q.* The minimization problem that defines the GMM estimator requires to set the first order derivatives $Q^{'}(\theta)$ of the objective function $Q(\theta)$ equal to 0, i.e.

$$Q^{'}(\theta) = \Gamma^T W f(x;\theta) = 0$$

As already mentioned in the previous chapter, one of the methods for finding roots of an equation $h(x) = 0$ is a numerical approach, such as the Newton Raphson method, whose result is a sequence of values defined by the step

$$x_{i+1} = x_i + \frac{h(x_i)}{h^{'}(x_i)}$$

Thus, in order to apply the NR algorithm to minimize the objective function $Q(\theta)$ one must estimate the first and the second order derivative matrices of $Q(\theta)$, since we are interested in setting the first order derivative as close to zero as possible. According to Gelman's article the first order derivatives are estimated using the IS methods.
In more detail, we denote by Γ the first order derivative matrix of the function $g(x;\theta)$ with respect to parameter θ. An approximation for Γ is provided by the formula (5.4). Focusing the attention on the statistics S_k and on the parameter θ_l, the generic cell of the first derivative matrix can be approximated in the following way:

$$\widehat{\gamma}_{kl} = \frac{1}{n_1}\sum_{j=1}^{n_1} s_k(x_j(t))\frac{\partial}{\partial\theta_l}\ log\ L(x_j;\widetilde{\theta}_l)\omega(x_j(t)) \tag{5.9}$$

where $\frac{\partial}{\partial\theta_l}\ log\ L(x_j;\theta_l^*)$ is the score function computed in θ_l^*. Knowing the expression for the likelihood of the model (equation 5.6), it is not difficult to determine the corresponding log-likelihood $\ell(\theta)$.

$$\ell(\theta) = ln\left(e^{(-g\lambda)}\frac{(g\lambda)^R}{R!}\prod_{r=1}^{R}\frac{1}{g}\frac{e^{f_{ri}(\beta,x_{(r-1)}(i\rightsquigarrow j))}}{\sum\limits_h e^{f_{ri}(\beta,x_{(r-1)}(i\rightsquigarrow h))}}\right) =$$

$$= -g\lambda + Rlng + Rln\lambda - lnR! - Rlng + \sum_{r=1}^{R} f_{ri}(\beta, x_{(r-1)}(i\rightsquigarrow j)) - \sum_{r=1}^{R} ln\sum e^{f_{ri}(\beta,x_{(r-1)}(i\rightsquigarrow j))}$$

Now, deriving with respect to the parameters, we obtain the corresponding score function:

$\frac{\partial\ell(\theta)}{\partial\lambda} = -n + \frac{R}{\lambda}$

$\frac{\partial\ell(\theta)}{\partial\beta_k} =$

$$= \sum_{r=1}^{R}\left(s_{rik}(x_{(r-1)}(i\rightsquigarrow j)) - \frac{1}{\sum\limits_h e^{\sum_k \beta_k s_{rik}(x_{(r-1)}(i\rightsquigarrow j))}}\sum_{h\neq i} s_{rik}(x_{(r-1)}(i\rightsquigarrow j))e^{\sum_k \beta_k s_{rik}(x_{(r-1)}(i\rightsquigarrow j))}\right)$$

Thus, denoting by $\widehat{\Gamma}(\widetilde{\theta})$ the approximation of the matrix Γ, we can approximate the first order derivative of Q, in the following way:

$$\widehat{Q}'(\theta) = \widehat{\Gamma}(\widetilde{\theta})^T\widehat{W}\widehat{g}(x;\widetilde{\theta})$$

Reasoning in a similar manner, the approximation of the second order derivative matrix can be approximate by:

$$\widehat{Q}''(\theta) = \widehat{\Gamma}(\widetilde{\theta})^T\widehat{W}\widehat{\Gamma}(\widetilde{\theta})$$

2.1.3) *Perform a Newton-Raphson step.*
The NR step is performed `nstep` times to find the new value for the instrumental distribution:

$$\widetilde{\theta}_{nstep+1} = \widetilde{\theta}_{nstep} + (\widehat{\Gamma}(\widetilde{\theta}_{nstep})^T\widehat{W}\widehat{\Gamma}(\widetilde{\theta}_{nstep}))^{-1}\widehat{\Gamma}(\widetilde{\theta}_{nstep})^T\widehat{W}\widehat{g}(x;\widetilde{\theta}_{nstep})$$

2.1.4) *Iterate steps from point 1 to 3*
To decide the number of NR iterations one can fix a priori the number `nstep` of NR steps, following some guidelines deriving from empirical evidence. The risk is that the IS weights can degenerate more or less frequently according to the initial point of the algorithm. In more detail, if between two consecutive NR steps, the variation in the parameter estimates is too large, the IS weights degenerate leading to all the consequences described in paragraph 4.3.1. This can happen at any NR step when we are still far from the goal.
Thus, the idea is to control the number of steps through the Shannon Entropy index described in equation (4.23). This means that during the steps from 1. to 3. the index is computed and if its value is too small, the current value of θ^* is assumed as the new value for $\widetilde{\theta}$ otherwise the NR steps are performed until the number of NR steps reaches a fix number `nstep`.
Therefore, two quantities must be defined: the a priori maximum number `nstep` of NR steps associated to each simulation step, and the threshold value for the Shannon Entropy Index.
The latter is defined so that the value assumed by the Shannon index $Ent(\theta^*, \widetilde{\theta})$ at the current value of the parameters θ^* and $\widetilde{\theta}$ is nearly equal to the 80% of the maximum value of the index given by $\log(n_1)$. The threshold 80% was established in an empirical way, in fact it was observed that for a lower value the weight distribution becomes too skewed, showing a huge amount of weights close to 0 and few higher values.
The choice of the maximum number of NR steps is related to the Shannon entropy threshold. In fact, it was observed that an optimal choice for `nstep` is given by 3 or 5 steps otherwise the normalized IS weights degenerate.

2.1.5) *Set* $\widetilde{\theta} = \theta^*$.

2.1.6) *Estimate the variance-covariance matrix* Σ_θ. The estimate of $\Sigma_{\widetilde{\theta}}$ is used to evaluate the standard error associated to the estimates and to generate the new value for θ^* (see step 9 below). According to (4.34), Σ_θ is approximated by:

$$\Sigma_{\widetilde{\theta}} = \left(\widehat{\Gamma}^T \, \widehat{W} \, \widehat{\Gamma}\right)^{-1}$$

where $\widehat{\Gamma}$ is the approximation first order derivatives matrix, computed in the current value of $\widetilde{\theta}_t$.

2.1.7) *Test algorithm convergence.*
This point of the algorithm is not properly a check for convergence, but it determines the end of sub-phase 1, i.e. if the new estimation for θ^* is close to the goal. If this happens then the sequence of simulation and optimization steps stops and a bigger set of simulations can be drawn.
The stopping rule is suggested by Gelman. He proposed to increase the number of simulations "when the variability step by step gets larger than the systematic

movement towards convergence" (Gelman, 1995). Denoting by $\{\widetilde{\theta}_t\}$ the successive values for the parameter θ determined by the $(t+1)$-th simulation-optimization step, we can defined the systematic movement using the norm of the two subsequent values $\widetilde{\theta}_t$ and $\widetilde{\theta}_{t+1}$, i.e:

$$\left\|\widetilde{\theta}_{t+1} - \widetilde{\theta}_t\right\| = \sqrt{(\widetilde{\theta}_{t+1} - \widetilde{\theta}_t)(\widetilde{\theta}_{t+1} - \widetilde{\theta}_t)^T}$$

The variability of the sequence $\{\widetilde{\theta}_t\}$ is defined using the standard error $s.e.(\{\widetilde{\theta}_t\})$ of the norms between the values of two consecutive simulation-optimization steps. Intuitively, if the variability of the norms step by step becomes larger than the norms of the current simulation-optimization step, it means that the estimate for the value $\widetilde{\theta}_{t+1}$ does not move far apart from the previous one and so sub-phase 1 can be stopped.
Thus:
if

$$s.e.(\{\widetilde{\theta}_t\}) > \left\|\widetilde{\theta}_{t+1} - \widetilde{\theta}_t\right\|$$

sub-phase 1 ends. Otherwise points from 8 to 10 are performed.

2.1.8) *Perform a new simulation step.*
Simulate n_1 draws from the new $\widetilde{\theta}$.

2.1.9) *Compute the new value for the target distribution.*
Using $\widetilde{\theta}$ and $\Sigma_{\widetilde{\theta}}$ the value θ^* is determined applying the Mahalanobis distance using the same procedure of points 1.1 and 1.2.

2.1.10) *Update the weighting matrix W.*
The weighting matrix W is updated according to the new value of $\widetilde{\theta}$ using formula (4.36)

2.1.11) *Repeat points 1 to 5 until the criterion of convergence is satisfied.*

Sub-phase 2

Sub-phase 2 aims to improve the convergence of the algorithm deriving from sub-phase 1. The idea is to draw a bigger number n_2 of simulations from the last estimate $\widetilde{\theta}$ of Sub-phase 1, to update the weighting matrix W and to perform several NR steps till convergence. In more detail, Sub-phase 2 consists in the following points:

2.2.1) *Generate the initial values.*
The last estimate deriving from Sub-phase 1 is the new value $\widetilde{\theta}$. Thus, knowing the estimate of its variance-covariance matrix the value θ^* is generated as in points 1.1 and 1.2.

2.2.2) *Draw n_2 simulations from the current values of $\widetilde{\theta}$*

2.2.3) *Update W according to formula (4.36)*

2.2.4) *Perform several NR steps till convergence*
Since the IS weights are now well behaving NR steps can be performed till convergence. Different convergence criteria can be applied. Three possibilities were described in paragraph 4.2.3 (equations (4.13), (4.14) and (4.15)) but they can be dangerous in view of the random nature of the updates for two main reasons. The first is that an update might be close to 0 by chance. The second relies on the fact that, if ϵ is small compared to the standard deviation of the updates, then when you are close to the optimum, the process still has a very small probability of stopping.
Thus a different criterion was applied. The stopping rule was proposed by Andrews (Andrews, 1997), and it is based on the asymptotic distribution of $Q(\theta)$. In more detail, Hansen (Hansen, 1982) proved that the minimized GMM criterion function has an asymptotic χ^2_{q-p} distribution, where p is the dimension of the parameter space and q is the number of moment conditions.
Thus, Andrews suggested to stop the NR steps when:

$$Q(\theta^*) \ \leq \ \frac{c_{q-p}}{n} \tag{5.10}$$

where c_{q-p} is the quantile of a χ^2_{q-p} distribution that leaves on the left usually a probability α equal to 0.05. Consequently, the NR steps are performed till the approximation of $Q(\theta)$ satisfies equation (5.10)

2.2.5) *Set* $\widehat{\theta} = \theta^*$

5.1.4 Phase3

The last phase is similar to the Robbins-Monro algorithm one. In fact a large set of simulations is drawn from the final estimates $\widehat{\theta}$. This big simulation is used to estimate the variance-covariance matrix of $\widehat{\theta}$:

$$\Sigma_{\widehat{\theta}} = \Gamma(\widehat{\theta})^T \ W(\widehat{\theta}) \ \Gamma(\widehat{\theta})$$

and to check the goodness of fit of the model. If the model well fits the data, then we should expect that the simulated values from $\widehat{\theta}$ assume values close to the observed ones. Thus, to measure the discrepancy among simulated and observed value a t-test is performed. In more detail, let $s_k^{sim}(x_j, \widehat{\theta})$ the number of configurations $s_k(x)$ deriving from the j-th simulation and by $s_k(x, \theta_0)$ the corresponding observed values. Denoting by $\overline{s}_k = \frac{1}{n}\left[s_k^{sim}(x_j, \widehat{\theta})\right]$ the average of the values assumed by the statistics according to the simulations and by $s.e.(\overline{s}_k)$, the ratio:

$$\frac{\overline{s}_k - s_k(x, \theta_0)}{s.e.(\overline{s}_k)}$$

is the realization of a T random variable as the number of simulations increases. If this value is close to 0 (i.e. less than 0.1) then the model well fits the data, otherwise it is not adequate.

5.2 Simulation results

In this section results deriving from several simulations are reported. A small simulation study is presented as an exploration of the reliability of the GMSM estimators and of relative efficiency of the GMSM and the MoM estimators. Intuition suggests that new statistics work well when the observation times are close, while the regular statistics can be applied for large waves. This is equivalent to saying that the Jaccard index assumes high values and low values, respectively.
In more detail, if we assume to observe the network at close observation moments, then we expect to observe a small value of changes, thus a small Jaccard index. In this situation, it is easy to follow the change that give rise to a reciprocal dyad or a transitive triad, and the new statistics can include this information in the estimation process clearly distinguishing between the origins of an observed effect. Thus, we believe that the GMSM estimator performs better than the MoM estimator.
On the other hand, if the observation times are far apart, then we expect a high number of changes and the additional information included in the new statistics cannot be so relevant. For instance, if we consider a network with a high reciprocity effect, and we observe it at two distant observation moments, we can imagine that null dyads and the asymmetric dyads become reciprocal, so that they have nearly the same weight in the process estimation. Then, the regular MoM estimator should be more efficient then the GMSM, since it does not involve the approximation of the weighting matrix W.
We consider two observational points and a very simple model which includes only the rate, the out-degree, the reciprocity and the transitivity effects. Therefore $\theta = (\lambda, \beta_1, \beta_2, \beta_3)$ is the four dimensional vector of parameters that should be estimated, where λ is the rate parameter, and β_1, β_2 and β_3 are related to out-degree, reciprocity and transitivity. Thus, the simulations are performed fixing different values of θ.
The choice of the values for simulating should be taken carefully for two main reasons. The former is related to the assumption of the SAO model which states that the process between two consecutive observational points is gradual. In more detail, if the combination of the four parameters leads to higher values of changes, then the assumption is not verified and the initial point of the stochastic approximation given by the Robbins-Monro algorithm is unreliable. So, the Jaccard index defined in equation (3.7) plays a key role for the choice of the parameter value for the simulations.
The latter is related to out-degree. If network density is too high and the combination of the parameters determine an increase of the ties present in the network, i.e. all the changes between the first and second networks are upwards, then the resulting out-degree effects cannot be estimated. In fact, all the actors show a tendency to create ties and this makes the effect singular. Thus, the degree effect is not provided by the Robbins-Monro algorithm.
These two points were the guidelines for the choice of the simulation values of the parameters.
The results are based on 500 simulations. The algorithm did not reach convergence in less than the 3% of simulations in all the cases. Two different applications of the GMSM

were considered. The GMSM approach was first implemented defining new statistics only for the reciprocity effects, i.e. using the two statistics S_{21} and S_{22} defined in the equations (4.2) and (4.3).
The first set of simulations was based on networks of 50 actors, generated from the values $\beta_1 = -2.5$, $\beta_2 = 2$ and $\beta_3 = 0.25$. They differ for the rate parameter λ which determines the Jaccard index variation among the networks.
Table 5.1 shows the main characteristics of the considered networks, that is the Jaccard index and the values assumed by the statistics. It clearly appears that the rate parameter is related to the number of changes between two observational points of time, in fact the higher the value of the rate parameter is, the lower the value of the Jaccard index is.
Table 5.2 reports the average estimates and the standard errors of the considered estimators, according to the different value of the rate parameter (in row) and the methods used (in columns).
If we look at the average estimates we can observe that they are quite close to the true values and the GMSM estimates are not significantly different from those obtained through the regular MoM. Thus, the algorithm provides reliable estimates. We can now focus the attention on the standard errors.
When the Jaccard index assumes the values 0.842 the standard errors related to $\widehat{\theta}_{MoM}$ are higher than those of both the GMSM estimators. This is true except for the out-degree, for which no new statistics were defined and for the rate in the case of $\widehat{\theta}_{GMSM}$. Taking into account only the two GMSM estimators and comparing them with respect to the standard error related to β_2 and β_3, the more efficient estimator is given by $\widehat{\theta}_{GMSM}$. As the Jaccard index increases, this tendency changes. In particular given a Jaccard index of 0.673, we observed that the standard errors of $\widehat{\theta}_{MoM}$ are still higher except for the reciprocity parameter. Furthermore, $\widehat{\theta}^*_{GMSM}$ seems to be more efficient than $\widehat{\theta}_{GMSM}$. Finally, when the Jaccard index takes value equal to 0.530, simulations suggest that $\widehat{\theta}_{MoM}$ is the more efficient estimator, even if $\widehat{\theta}_{GMSM}$ presents a smaller standard error for the out-degree parameter. Varying the number of actors the trend is still respected, as Tables 5.4 and 5.6 show. Some other considerations can be expressed. In fact, looking at the simulations characterized by a low Jaccard index, it seems that $\widehat{\theta}_{GMSM}$ is more efficient than $\widehat{\theta}^*_{GMSM}$, but this trend is not confirmed in all the cases.
Furthermore, the standard values of the considered estimators related to the rate and out-degree parameters are quite similar. A bigger variability is registered for the standard error of the reciprocity and the transitivity parameters. This confirm what intuition suggests. Having observed that the number of actors does not influence the performance of the considered estimators, we can analyze what happens if the values assumed by the reciprocity and transitivity parameters vary, considering a network of 50 actors. This choice is due to algorithm computational time. As it will be shown later, the greater the size of the network is, the higher the computational effort is. Since from the previous simulations it follows that the number of changes has a main role and the reciprocity and transitivity parameters do not influence the changes but determine only the tie that is changed, we expect that these two parameters will not play a key role in the efficiency

comparison of the MoM estimators and of the GMSM estimators.

	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Jaccard	0.842	0.673	0.530
Changes	53	79	137
Degree	225	213	187
Reciprocity	62	34	58
Real	32	14	28
New+Persitent	30	20	30
Transitiviy	103	88	69
Real	5	10	9
Agreement	9	15	7

Table 5.1: *Jaccard Index and values assumed by the Statistics for different values of the rate parameter ($\beta_1 = -2.5$, $\beta_2 = 2$ and $\beta_3 = 0.25$, $g = 50$ nodes).*

	Est_{MoM}	$S.E._{MoM}$	Est^*_{GMSM}	$S.E.^*_{GMSM}$	Est_{GMSM}	$S.E._{GMSM}$
Jacc=0.842						
λ=1						
Rate function						
λ	1.025	0.150	1.859	**0.147**	1.100	0.154
Objective function						
β_1	-2.508	**0.312**	-2.382	0.314	-2.457	0.315
β_2	1.752	0.556	1.831	**0.526**	1.846	**0.501**
β_3	0.251	0.226	0.249	**0.222**	0.247	**0.206**
Jacc=0.673						
Rate function						
$\lambda = 2$						
λ	1.919	0.236	1.894	**0.209**	1.935	**0.198**
Objective function						
β_1	-2.347	0.240	-2.385	**0.225**	-2.511	0.274
β_2	1.817	**0.396**	1.776	0.487	1.896	0.471
β_3	0.219	0.202	0.251	**0.189**	0.244	**0.201**
Jacc=0.530						
λ=3						
Rate function						
λ	2.806	**0.174**	2.859	0.209	3.083	0.294
Objective function						
β_1	-2.314	0.210	-2.382	0.225	-2.313	**0.199**
β_2	1.945	**0.337**	2.032	0.487	1.998	0.556
β_3	0.272	**0.187**	0.22	0.189	0.234	0.191

Table 5.2: *Simulation results based on the networks described in Table 5.1*

	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$
Jaccard	0.876	0.652	0.489
Changes	86	220	321
Degree	480	466	369
Reciprocity	44	116	106
Real	10	42	48
New+Persitent	34	74	58
Transitivity	284	288	162
Real	15	25	18
Agreement	13	26	23

Table 5.3: *Jaccard Index and values assumed by the Statistics for different values of the rate parameter ($\beta_1 = -2.5$, $\beta_2 = 2$ and $\beta_3 = 0.25$, $g = 75$ nodes).*

	Est_{MoM}	$S.E._{MoM}$	Est^*_{GMSM}	$S.E._{GMSM}^*$	Est_{GMSM}	$S.E._{GMSM}$
Jacc=0.876						
λ=1						
Rate function						
λ	1.192	0.127	1.201	**0.126**	1.198	**0.125**
Objective function						
β_1	-2.461	0.248	-2.512	**0.247**	-2.501	**0.247**
β_2	1.817	0.452	1.914	0.412	1.927	**0.401**
β_3	0.301	0.148	0.285	0.150	0.284	**0.134**
Jacc=0.652						
λ=3						
Rate function						
λ=3						
λ	3.138	**0.219**	3.120	0.220	3.124	0.210
Objective function						
β_1	-2.616	**0.169**	-2.556	0.170	-2.561	0.185
β_2	2.336	0.250	2.254	0.241	2.152	**0.231**
β_3	0.361	0.089	0.321	0.104	0.324	**0.088**
Jacc=0.489						
λ=5						
Rate function						
λ	4.998	**0.335**	4.985	0.441	4.987	0.443
Objective function						
β_1	-2.439	**0.123**	-2.325	0.201	-2.421	0.205
β_2	2.009	**0.210**	1.999	**0.210**	1.997	0.301
β_3	0.21	**0.078**	0.262	0.080	0.241	0.090

Table 5.4: *Simulation results on the networks described in 5.3*

	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$
Jaccard	0.848	0.640	0.565
Changes	156	390	485
Degree	899	785	732
Reciprocity	132	180	186
Real	22	94	90
New+Persitent	110	86	96
Transitivity	788	545	545
Real	21	58	67
Agreement	23	44	51

Table 5.5: *Jaccard Index and values assumed by the Statistics for different values of the rate parameter ($\beta_1 = -2.5$, $\beta_2 = 2$ and $\beta_3 = 0.25$, $g = 100$ nodes).*

	Est_{MoM}	$S.E._{MoM}$	Est^*_{GMSM}	$S.E._{GMSM}^*$	Est_{GMSM}	$S.E._{GMSM}$
Jacc=0.878						
λ=2						
Rate function						
λ	1.917	**0.129**	2.034	0.130	2.099	**0.129**
Objective function						
β_1	-2.596	**0.193**	-2.587	**0.193**	-2.512	0.195
β_2	2.052	0.315	2.074	0.302	2.084	**0.301**
β_3	0.227	0.085	0.231	0.090	0.244	**0.084**
Jacc=0.640						
λ=4						
Rate function						
λ	4.237	0.229	4.235	**0.228**	4.321	0.230
Objective function						
β_1	-2.391	**0.118**	-2.454	0.119	-2.424	0.119
β_2	2.310	**0.184**	2.241	0.190	2.329	0.254
β_3	0.174	0.174	0.194	0.145	0.201	**0.143**
Jacc=0.565						
λ=6 Rate function						
λ	5.487	**0.253**	5.678	0.257	5.987	0.256
Objective function						
β_1	-2.533	0.113	-2.578	**0.110**	-2.487	0.12
β_2	2.171	**0.172**	2.236	0.197	2.147	0.195
β_3	0.240	**0.047**	0.221	0.098	0.235	0.102

Table 5.6: *Simulation results on the networks described in 5.5*

	$\beta_2 = 1$	$\beta_2 = 2$	$\beta_2 = 3$
Jaccard	0.719	0.673	0.700
Changes	74	79	97
Degree	196	213	275
Reciprocity	24	34	120
Real	4	14	84
New+Persitent	20	20	36
Transitiviy	71	88	27
Real	3	10	32
Agreement	0	15	161

Table 5.7: *Jaccard Index and values assumed by the Statistics for different values of the reciprocity parameter ($\lambda = 2$, $\beta_1 = -2.5$ and $\beta_3 = 0.25$, $g = 50$ nodes).*

	Est_{MoM}	$S.E._{MoM}$	Est_{GMSM}	$S.E._{GMSM}$
Jacc = 0.719				
$\beta_2 = 1$				
Rate function				
λ	1.748	0.207	1.804	**0.199**
Objective function				
β_1	-2.464	**0.392**	-2.501	0.395
β_2	1.108	0.383	0.987	**0.384**
β_3	0.238	0.104	0.251	**0.999**
Jacc = 0.673				
$\beta_2 = 2$				
Rate function				
λ	1.919	0.236	1.935	**0.198**
Objective function				
β_1	-2.347	**0.240**	-2.511	0.274
β_2	1.817	**0.396**	1.896	0.471
β_3	0.219	0.202	0.244	**0.201**
Jacc =0.700				
$\beta_2 = 3$				
Rate function				
λ	2.013	0.396	2.030	**0.216**
Objective function				
β_1	-2.218	**0.236**	-2.376	**0.236**
β_2	3.804	0.385	3.927	**0.384**
β_3	0.210	0.094	0.253	**0.081**

Table 5.8: *Simulation results on the networks described in 5.7*

	$\beta_3 = 0.25$	$\beta_3 = 0.5$	$\beta_3 = 0.75$
Jaccard	0.673	0.616	0.816
Changes	79	102	201
Degree	213	450	1033
Reciprocity	34	39	212
Real	14	6	100
New+Persitent	20	30	112
Transitiviy	88	265	1543
Real	10	35	232
Agreement	15	60	252

Table 5.9: *Jaccard Index and values assumed by the Statistics for different values of the transitivity parameter ($\lambda = 2$, $\beta_1 = -2.5$ and $\beta_2 = 2$, $g = 50$ nodes).*

	Est_{MoM}	$S.E._{MoM}$	Est_{GMSM}	$S.E._{GMSM}$
Jacc = 0.673				
$\beta_3 = 0.25$				
Rate function				
λ	1.919	0.236	1.935	**0.198**
Objective function				
β_1	-2.347	**0.240**	-2.511	0.274
β_2	1.817	**0.396**	1.896	0.471
β_3	0.219	0.202	0.244	**0.201**
Jacc = 0.662				
$\beta_3 = 0.5$	2.207	**0.239**	2.183	0.245
Rate function				
λ				
Objective function				
β_1	-2.240	**0.216**	-2.325	0.351
β_2	1.449	**0.365**	1.328	0.366
β_3	0.447	0.207	0.478	**0.205**
Jacc = 0.816				
$\beta_3 = 0.75$				
Rate function				
λ	2.058	0.149	2.030	**0.132**
Objective function				
β_1	-2.956	**0.165**	-2.376	**0.165**
β_2	1.956	0.199	3.927	**0.195**
β_3	0.800	0.074	0.253	**0.069**

Table 5.10: *Simulation results on the networks described in 5.9*

The results are reported in the Tables 5.8 and 5.10 but they are not particular meaningful. In fact, the values assumed by the parameters do not affect the Jaccard index and the GMSM estimator perform better than the MoM estimator when the simulated network are close in time. Regarding the asymptotic distribution of the $GMSM$ esti-

mator, results suggest that it is asymptotically normal. Just to give an idea, Figure 5.2, reports the histogram relative to the estimates deriving from a simulation of 600 simulations from the following parameter values: $\lambda = 2$, $\beta_1 = -2.5$, $\beta_2 = 4$ and $\beta_3 = 0.25$. It appears that the histograms assume the usual form of a normal density with mean usually close to the real values of the parameters except for the transitivity effect. Its histogram shows an irregular left tail, but probably a higher number of simulations can make the histogram similar to that of a gaussian distribution.

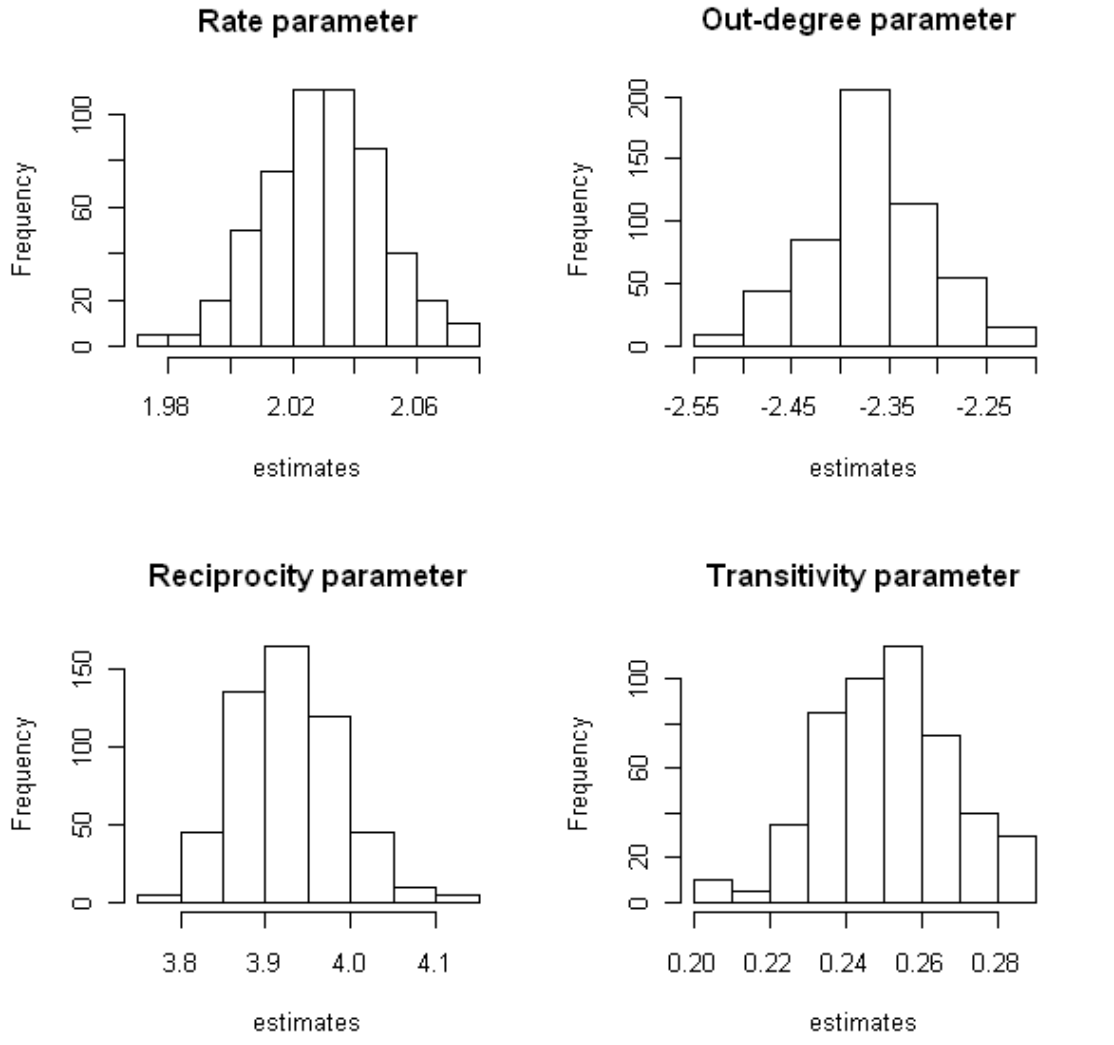


Figure 5.2: *The asymptotic distribution of the GMSM estimator*

5.3 Computational aspects

Different aspects should be evaluated when using stochastic approximations. In the previous paragraph the estimates obtained using the GMM approach and the MoM were compared, since efficiency is a relevant property from a statistical point of view. Time is important as well, because we must not wait for a long time in order to obtain the estimates. Thus, some words should be spent about computational time.
Few guidelines characterize the code used to implement the algorithm. The first one is to avoid cycles when it is possible. Cycles are quite expensive in terms of time and memory, so they reduce computational efficiency. For this reason, the computation of the new statistics is obtained through matrix operations.
The second guideline is to use parallel computing. Roughly speaking, the idea behind this strategy is that "a job is most efficiently performed if it can be divided into smaller tasks that can be calculated almost independently, without communication among them. Such problems are called *embarrassingly parallel* computing" (Yamamoto *et al.*, 2010). Many statistical methods have embarrassingly parallel nature, first of all simulations. Thus, since the algorithm simulates values several times, parallel computing can be used to reduce computational time and improve the efficiency of the algorithm.
Different libraries allow to exploit parallel computing in R, among them the library `snow`, `foreach` and `multicore`[3]. In this case, the `snow` package was applied, since it allows to create cross-platform cluster (i.e. the nodes of the cluster may be on different platforms) very easily. The simulation process was split on a different number of nodes in order to investigate the gain in efficiency.
Table 5.11 shows the mean time taken by the algorithm in order to perform some steps. In particular, the time:

- for finding the parameter value for the target distribution through the Mahalanobis distance (*target*)
- for simulating network evolution $n_1 = 100$ times (*Simulation*)
- for doing $nstep = 5$ NR steps (*Optimization*)
- for simulating network evolution $n_2 = 500$ times (*Simulation big*)
- for reaching convergence (*Final optimization*)

The results are quite evident. The parallel computing significantly reduces simulations time, which is the more expensive part of the algorithm. The mean total time to reach convergence is nearly about 5 minutes if one core is used. It decreases to about 3 minutes and to less than 2 minutes when the parallel computing is based on two and four cores, respectively. Varying the number of parameters, there is no significantly computational time variation, while increasing the number of actors and the number of statistics has some effects. Table 5.12 shows the computational time and its variation according to the number of actors.

[3]Details can be found at the web page http://cran.r-project.org/

	1 cores	2 cores	4 cores
Target	0.00	0.00	0.00
Simulation ($n_1 = 100$)	20.21	12.11	7.54
Optimization	15.22	14.12	13.96
Simulation big ($n_2 = 500$)	58.12	31.24	19.52
Final optimization	19.04	18.27	17.01

Table 5.11: *Computational time according to the number of cores used ($g = 50$, $\boldsymbol{\theta}$= $(2, -2.5, 2, 0.25)$, number of statistics=7)*

	75 actors	100 actors
Target	0.00	0.00
Simulation ($n_1 = 100$)	10.33	11.04
Optimization	15.43	18.21
Simulation big ($n_2 = 500$)	26.78	28.54
Final optimization	22.02	27.96

Table 5.12: *Computational time according to the number of actors and the use of 4 cores ($\boldsymbol{\theta}$= $(2, -2.5, 2, 0.25)$,number of statistics=7)*

Increasing the number of actors has consequences on both the simulation step and on the optimization step, since the adjacency matrix is involved in both steps. The mean total time for reaching convergence is less than 3 minutes and about 3.30 minutes for a network of 75 and 100 actors, respectively. This increase in time is quite relevant if compared to the nearly 2 minutes computed on 50 actors.
Regarding the number of statistics, its increase is significant only in terms of the optimization steps because the simulation does not involve the number of statistics (Table 5.13). The gain of using 5 statistics instead of 7 statistics is nearly about 15 seconds, but we know that in particular circumstances using all the statistics can be more appropriate for a statistical point of view and 15 seconds become not so relevant.

	5 Statistics	7 Statistics
Target	0.00	0.00
Simulation ($n_1 = 100$)	8.02	8.12
Optimization	11.98	13.96
Simulation big ($n_2 = 500$)	20.04	19.52
Final optimization	14.01	17.01

Table 5.13: *Computational time according to the number of cores used ($g = 50$, $\boldsymbol{\theta}$= $(2, -2.5, 2, 0.25)$)*

Conclusions and further developments

In the previous pages the proposal of new statistics for estimating the parameters of the SAO models was explained, underlying the relevant statistical matters that arise in this context and the need of using a different estimation method, represented by the GMM approach.
The introduction of new statistics in the estimation process seems to give promising results when the number of changes between consecutive points of time is small, i.e. when the observational points of time are close. This is consistent with what intuition suggests and leads to further developments and deeper analysis. More work should be done, since there are a lot of open-questions and different directions that can be investigated.
The first point is that more simulations are necessary in order to support the nice results. We have just observed that the time between two observation moments plays a key role, but probably intrinsic characteristics of networks, such as density, can say something more. Thus, one of the future developments is to perform more simulations, with the aim to define a more precise set of conditions for which it is useful apply the GMM.
In this context it can be also useful to perform a test on the over-identifying conditions, so that the relative importance of the involved statistics can be significantly proved. So, the test can be useful in application context in order to decide which statistics should be included in the estimation process. This is also important since it seems that increasing the number of the statistics leads to an increase of computational efficiency, even if the lost in computational time is not so relevant.
Simulations also underline that in few cases the algorithm does not reach convergence. Therefore, one can investigate if this is due merely to the simulated values or also to the starting point of the algorithm. The latter requires a sensitivity analysis of the starting values.
Another interesting question is to define the GMM estimator when the number of network observations is greater than two, since panel data are usually sequence of more than two observations. One of the assumptions of the SAO model during the estimation process using the Robbins- Monro algorithm is that the parameter related to the objective function are constant during all the observation periods. Therefore, the statistics to estimate the parameters are built as an unweighted sums of statistics over the periods. For the Hamming distances between consecutive observations, used to estimate the rate

parameters, this is optimal for simple models because of sufficiency considerations. For other statistics this can be also close to optimal. However, the extent to which this is indeed optimal could be investigated using the GMM.
Since the Markov property implies conditional independence between the periods, we know that the W matrix should be block diagonal, with 0 elements for pairs of statistics belonging to different periods. Under this restriction, the GMM that allows differential weights for different periods could be investigated.
Thus, the promising results arising from simulations lead to this further improvement which will be the topic of further research.

Bibliography

Agresti, A., & Corporation, Ebooks. 1990. *Categorical data analysis*. Vol. 5. Wiley Online Library.

Andrews, D.W.K. 1997. A stopping rule for the computation of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 913–931.

Antoniou, A., & Lu, W.S. 2007. *Practical optimization: algorithms and engineering applications*. Springer-Verlag New York Inc.

Baerveldt, C., Van Duijn, M.A.J., Vermeij, L., & Van Hemert, D.A. 2004. Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. *Social Networks*, **26**(1), 55–74.

Barnes, J.A. 1954. Class and committees in a Norwegian island parish. *Human relations*, **7**(1), 39.

Batagelj, V., & Mrvar, A. 2001. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social networks*, **23**(3), 237–243.

Bearman, P.S., Moody, J., & Stovel, K. 2004. Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks1. *ajs*, **110**(1), 44–91.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., & Sagastizábal, C.A. 2006. *Numerical optimization: theoretical and practical aspects*. Springer-Verlag New York Inc.

Borgatti, S.P., Everett, M.G., & Freeman, L.C. 2002. UCINET for Windows: Software for social network analysis. *Harvard Analytic Technologies*, **6**.

Brandes, U., & Wagner, D. 2003. Visone-analysis and visualization of social networks. *In: Graph drawing software*. Citeseer.

Brandes, U., Lerner, J., & Snijders, T.A.B. 2009. Networks evolving step by step: statistical analysis of dyadic event data. *Pages 200–205 of: Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*. IEEE.

Brueckner, J.K. 2005. Internalization of airport congestion: A network analysis. *International Journal of Industrial Organization*, **23**(7-8), 599–614.

Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., *et al.* 2003. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, **31**(9), 2443.

Burguete, J.F., Gallant, A.R., & Souza, G. 1982. On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews*, **1**(2), 151–190.

Cappé, O., Moulines, E., & Rydén, T. 2005. *Inference in hidden Markov models*. Springer Verlag.

Cappé, O., Douc, R., Guillin, A., Marin, J.M., & Robert, C.P. 2008. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, **18**(4), 447–459.

Carrington, P.J., Scott, J., & Wasserman, S. 2005. *Models and methods in social network analysis*. Cambridge Univ Pr.

Cartwright, D., & Zander, A.F. 1968. *Group dynamics: Research and theory*. Harper & Row.

Contractor, N.S., Wasserman, S., & Faust, K. 2006. Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Academy of Management Review*, **31**(3), 681.

Corander, J., Dahmström, K., & Dahmström, P. 1998. Maximum likelihood estimation for Markov graphs. *Research report*, **8**.

Davis, J.A. 1979. The davis/holland/leinhardt studies: An overview. *Perspectives on social network research*, 51–62.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, DL. 2000. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, **50**(1), 1–18.

Doreian, P., & Stokman, F.N. 1997. *Evolution of social networks*. Routledge.

Doreian, P., Teuter, K., & Wang, C.H. 1984. Network autocorrelation models. *Sociological Methods & Research*, **13**(2), 155.

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), 234–254.

Dunne, J.A., Williams, R.J., & Martinez, N.D. 2002. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(20), 12917.

Durrett, R., & Durrett, R. 1999. *Essentials of stochastic processes*. Springer New York.

Erdős, P., & Rényi, A. 1959. On random graphs I. *Publ. Math. Debrecen*, **6**(290-297), 156.

Erdős, P., & Rényi, A. 1960. *On the evolution of random graphs*. Citeseer.

Fienberg, S.E., & Wasserman, S.S. 1981. Categorical data analysis of single sociometric relations. *Sociological methodology*, **12**, 156–192.

Fletcher, R. 1980. *Practical Methods of Optimization, Vol. 1 and 2*.

Frank, O. 1977. Survey sampling in graphs. *Journal of Statistical Planning and Inference*, **1**(3), 235–264.

Frank, O. 1988. Random sampling and social networks: a survey of various approaches. *Mathematiques, Informatique, et Sciences Humaines*, **26**, 19–33.

Frank, O., & Strauss, D. 1986. Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

Freeman, L.C. 1984. The impact of computer based communication on the social structure of an emerging scientific specialty* 1. *Social Networks*, **6**(3), 201–221.

Gallant, A.R., & Corporation, Ebooks. 1987. Nonlinear statistical models.

Garfield, E., Sher, I.H., & Torpie, R.J. 1964. *The use of citation data in writing the history of science*.

Gelfand, A.E., & Carlin, B.P. 1993. Maximum-likelihood estimation for constrained-or missing-data models. *Canadian Journal of Statistics*, **21**(3), 303–311.

Gelman, A. 1995. Method of moments using Monte Carlo simulation. *Journal of Computational and Graphical Statistics*, **4**(1), 36–54.

Geweke, J. 2005. *Contemporary Bayesian econometrics and statistics*. Wiley-Interscience.

Geyer, C.J., & Thompson, E.A. 1992. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 657–699.

Goulder, A.W. 1960. The norm of reciprocity: A preliminary statement. *American Sociological Review*, **25**(2), 161–178.

Gourieroux, C., & Monfort, A. 1993. Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, **59**(1-2), 5–33.

Gourieroux, C., & Monfort, A. 1996. *Simulation-based econometric methods*. Oxford University Press, USA.

Granovetter, M. 1983. The strength of weak ties: A network theory revisited. *Sociological theory*, **1**(1), 201–233.

Granovetter, M.S. 1995. *Getting a job: A study of contacts and careers*. University of Chicago Press.

Green, W. 2000. *Econometric Analysis. 4-th edition*.

Grossman, J.W., Ion, P., & De Castro, R. 2003. *Erdos Number Project*.

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L.A.N. 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(22), 7794.

Hall, A.R. 2005. *Generalized method of moments*. Oxford University Press, USA.

Hallinan, M.T. 1979. The process of friendship formation. *Social Networks*, **1**(2), 193–210.

Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., & Morris, M. 2008. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, **24**(1), 1548.

Hanneman, R.A., & Riddle, M. 2005. *Introduction to social network methods*. University of California Riverside, CA.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hansen, L.P. 2007. Generalized Method of Moments Estimation. *The New Palgrave Dictionary of Economics, ed. by Stephen Durlauf, and Lawrence Blume. Macmillan, forthcoming*.

Hansen, L.P., Heaton, J., & Yaron, A. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, **14**(3), 262–280.

Harary, F., Norman, R.Z., & Cartwright, D. 1965. *Structural models*. Wiley New York.

Harrigan, N. 2007. PNet for Dummies: An introduction to estimating exponential random graph (p*) models with PNet. http://www.sna.unimelb.edu.au/pnet/pnet.html.

Heil, G.H., & White, H.C. 1976. An algorithm for finding simultaneous homomorphic correspondences between graphs and their image graphs. *Behavioral Science*, **21**(1), 26–35.

Hensher, D.A., & Johnson, L.W. 1981. *Applied discrete-choice modelling*. Halsted Press.

Hoff, P.D., Raftery, A.E., & Handcock, M.S. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Holland, P.W., & Leinhardt, S. 1972. *Holland and Leinhardt Reply: Some Evidence on the Transitivity of Positive Interpersonal Sentiment.*

Holland, P.W., & Leinhardt, S. 1977. A dynamic model for social networks. *The Journal of Mathematical Sociology*, **5**(1), 5–20.

Holland, P.W., & Leinhardt, S. 1981. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**(373), 33–50.

Horn, R.A., & Johnson, C.R. 1990. *Matrix analysis.* Cambridge Univ Pr.

Huisman, M. 2009. Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, **10**(1).

Huisman, M., & Van Duijn, M.A.J. 2003. StOCNET: Software for the statistical analysis of social networks. *Connections*, **25**(1), 7–26.

Huisman, M., & van Duijn, M.A.J. 2005. Software for statistical analysis of social networks. *Models and methods in social network analysis*, 270–316.

Hummon, N.P., Doreian, P., & Freeman, L.C. 1990. Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. *Science Communication*, **11**(4), 459.

Johnson, NL, & Kotz, S. 1970. Distributions in statistics: Continuous univariate distributions. Vol. 1, 2.

Jordan, F. 2008. Predicting target selection by terrorists: a network analysis of the 2005 London underground attacks. *International Journal of Critical Infrastructures*, **4**(1), 206–214.

Katz, L., & Powell, J.H. 1955. Measurement of the tendency toward reciprocation of choice. *Sociometry*, **18**(4), 403–409.

Knoke, D., & Yang, S. 2008. *Social network analysis.* Sage Publications, Inc.

Kolaczyk, E.D. 2009. *Statistical analysis of network data: methods and models.* Springer Verlag.

Krackhardt, D. 1990. Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly*, **35**(2), 342–369.

Krackhardt, D., & Hanson, J.R. 1997. Informal networks: the company. *Knowledge in organizations*, 37–50.

Krebs, V.E. 2002. Mapping networks of terrorist cells. *Connections*, **24**(3), 43–52.

Lazega, E., & Van Duijn, M. 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data* 1. *Social Networks*, **19**(4), 375–397.

Leenders, R.T.A.J. 2002. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, **24**(1), 21–47.

Liesenfeld, R., & Breitung, J. 1998. *Simulation based methods of moments in empirical finance*. Wirtschaftswiss. Fak. der Eberhard-Karls-Univ.

Liu, J.S. 2003. *Monte Carlo strategies in scientific computing*. Springer.

Liu, X., Bollen, J., Nelson, M.L., & Van de Sompel, H. 2005. Co-authorship networks in the digital library research community. *Information Processing & Management*, **41**(6), 1462–1480.

Lospinoso, J.A., Schweinberger, M., Snijders, T.A.B., & Ripley, R.M. 2010. Assessing and accounting for time heterogeneity in stochastic actor oriented models. *Advances in Data Analysis and Computation, vol. Special Issue on Social Networks (Submitted)*.

Maddala, G.S. 1986. *Limited-dependent and qualitative variables in econometrics*. Cambridge Univ Pr.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. *Page 49 of: Proceedings of the National Institute of Science, Calcutta*, vol. 12.

Marsden, P.V. 1990. Network data and measurement. *Annual review of sociology*, **16**(1), 435–463.

Marsden, P.V., & Friedkin, N.E. 1993. Network studies of social influence. *Sociological Methods & Research*, **22**(1), 127.

Mayo, E. 1977. *The social problems of an industrial civilization*. Ayer Co Pub.

McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, **57**(5), 995–1026.

Milgram, S. 1967. The small world problem. *Psychology today*, **2**(1), 60–67.

Mitchell, J.C. 1969. *Social networks in urban situations*. Univ. Press.

Mizruchi, M.S. 2003. What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates.

Moody, J. 1998. Matrix methods for calculating the triad census. *Social Networks*, **20**(4), 291–299.

Moody, J. 2001. Race, School Integration, and Friendship Segregation in America1. *ajs*, **107**(3), 679–716.

Moreno, J.L. 1953. *Who shall survive?: foundations of sociometry, group psychotherapy, and sociodrama*. Beacon House Beacon, NY.

Nemeth, R., & Smith, D.A. 1985. International trade and world-system structure: a multiple network analysis. *Review (Fernand Braudel Center)*, **8**(4), 517–560.

Newey, W.K., & McFadden, D. 1994. Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4**, 2111–2245.

Newman, M. 2004a. Who is the best connected scientist? A study of scientific coauthorship networks. *Complex networks*, 337–370.

Newman, MEJ. 2003. The structure and function of complex networks. *Arxiv preprint cond-mat/0303516*.

Newman, M.E.J. 2004b. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, **101**(Suppl 1), 5200.

Nooy, W., Mrvar, A., & Batagelj, V. 2005. Exploratory social network analysis with Pajek.

Norris, J.R. 1998. *Markov chains*. Cambridge Univ Pr.

Padgett, J.F., & Ansell, C.K. 1993. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, **98**(6), 1259–1319.

Pakes, A., & Pollard, D. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, **57**(5), 1027–1057.

Pattison, P., & Wasserman, S. 1999. Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, **52**(2), 169–193.

Pellegrini, M., Haynor, D., & Johnson, J.M. 2004. Protein interaction networks. *Expert Review of Proteomics*, **1**(2), 239–249.

Politianus, A., & Adimari, G. 1769. *Coniurationis Pactianae anni 1478 Commentarium*.

Quandt, R.E. 1983. Computational problems and methods. *Handbook of econometrics*, **1**, 699–764.

Ripley, R., & Snijders, T.A.B. 2010. Manual for SIENA version 4.0. *University of Oxford*.

Robbins, H., & Monro, S. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, **22**(3), 400–407.

Robert, C., & Casella, G. 2010. *Introducing Monte Carlo Methods with R*. Springer Verlag.

Robert, C.P., & Casella, G. 2004. *Monte Carlo statistical methods*. Springer Verlag.

Robins, G., & Pattison, P. 2001. Random graph models for temporal processes in social networks. *The Journal of Mathematical Sociology*, **25**(1), 5–41.

Robins, G., Pattison, P., & Wasserman, S. 1999. Logit models and logistic regressions for social networks: III. Valued relations. *Psychometrika*, **64**(3), 371–394.

Robins, G., Pattison, P., & Elliott, P. 2001a. Network models for social influence processes. *Psychometrika*, **66**(2), 161–189.

Robins, G., Elliott, P., & Pattison, P. 2001b. Network models for social selection processes. *Social Networks*, **23**(1), 1–30.

Robins, G., Pattison, P., & Woolcock, J. 2004. Missing data in networks: Exponential random graph (p*) models for networks with non-respondents. *Social Networks*, **26**(3), 257–283.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. 2007. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), 173–191.

Rodriguez, M.A., & Pepe, A. 2008. On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, **2**(3), 195–201.

Ross, S.M. 1996. *Stochastic processes*. Wiley New York.

Rothenberg, R. 2001. From whole cloth: Making up the terrorist network. *New York Times*.

Rubinstein, R.Y., & Kroese, D.P. 2008. *Simulation and the Monte Carlo method*. Wiley-interscience.

Schank, T., & Wagner, D. 2005. Finding, counting and listing all triangles in large graphs, an experimental study. *Experimental and Efficient Algorithms*, 606–609.

Schweinberger, M. 2005. Statistical modeling of network panel data: Goodness-of-fit. *Submitted, draft version downloadable from http://ppswmm. ppsw. rug. nl/schweinb/*.

Schweinberger, M., & Snijders, T.A.B. 2003. Settings in social networks: A measurement model. *Sociological Methodology*, **33**(1), 307–341.

Scott, J. 1988. Social network analysis. *Sociology*, **22**(1), 109.

Scotti, M., Podani, J., & Jordán, F. 2007. Weighting, scale dependence and indirect effects in ecological networks: a comparative study. *ecological complexity*, **4**(3), 148–159.

Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, PA, Mukherjee, G., & Manna, SS. 2003. Small-world properties of the Indian railway network. *Physical Review E*, **67**(3), 36106.

Shortreed, S., Handcock, M.S., & Hoff, P. 2006. Positional estimation within a latent space model for networks. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **2**(1), 24–33.

Small, H. 1999. Visualizing science by citation mapping. *Journal of the American society for Information Science*, **50**(9), 799–813.

Smith, D.A., & White, D.R. 1992. Structure and dynamics of the global economy: Network analysis of international trade 1965-1980. *Social Forces*, **70**(4), 857–893.

Snijders, T. 2009a. Longitudinal methods of network analysis. *Meyers B.(ed).*

Snijders, T., & Van Duijn, M. 1997. Simulation for statistical inference in dynamic network models. *LECTURE NOTES IN ECONOMICS AND MATHEMATICAL SYSTEMS*, 493–512.

Snijders, T., Koskinen, J., & Schweinberger, M. 2010a. Maximum likelihood estimation for social network dynamics. *Annals*, **4**(2), 567–588.

Snijders, T.A.B. 1996. Stochastic actor-oriented models for network change. *The Journal of Mathematical Sociology*, **21**(1), 149–172.

Snijders, T.A.B. 2001. The statistical evaluation of social network dynamics. *Sociological methodology*, 361–395.

Snijders, T.A.B. 2002. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**(2), 1–40.

Snijders, T.A.B. 2005. Models for Longitudinal Network Data. *Models and methods in social network analysis*, 215.

Snijders, T.A.B. 2006. Statistical methods for network dynamics. *Pages 281–296 of: Proceedings of the XLIII Scientific Meeting, Italian Statistical Society.* Citeseer.

Snijders, T.A.B. 2007. Analysing dynamics of non-directed social networks. *preparation. Transparencies available at internet.*

Snijders, T.A.B. 2009b. Specification and estimation of exponential random graph models for social (and other) networks.

Snijders, T.A.B., Pattison, P.E., Robins, G.L., & Handcock, M.S. 2006. New specifications for exponential random graph models. *Sociological Methodology*, **36**(1), 99–153.

Snijders, T.A.B., Steglich, C.E.G., & Schweinberger, M. 2007. Modeling the co-evolution of networks and behavior. *Longitudinal models in the behavioral and related sciences*, 41–71.

Snijders, T.A.B., Van de Bunt, G.G., & Steglich, C.E.G. 2010b. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, **32**(1), 44–60.

Solomonoff, R., & Rapoport, A. 1951. Connectivity of random nets. *Bulletin of Mathematical Biology*, **13**(2), 107–117.

Spall, J.C. 2003. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley and Sons.

Steglich, C., Snijders, T.A.B., & Pearson, M. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*.

Strauss, D., & Ikeda, M. 1990. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, **85**(409), 204–212.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge Univ Pr.

Travers, J., & Milgram, S. 1969. An experimental study of the small world problem. *Sociometry*, **32**(4), 425–443.

Udehn, L. 2002. The changing face of methodological individualism. *Annual Review of Sociology*, 479–508.

Valente, T.W. 2005. Network models and methods for studying the diffusion of innovations. *Models and methods in social network analysis*, 98–116.

Vermeij, L., van Duijn, M.A.J., & Baerveldt, C. 2009. Ethnic segregation in context: Social discrimination among native Dutch pupils and their ethnic minority classmates. *Social Networks*, **31**(4), 230–239.

Wang, P., Robins, G., & Pattison, P. 2006. PNet: Program for the estimation and simulation of p* exponential random graph models, User Manual. *Department of Psychology, University of Melbourne*.

Wasserman, S. 1980a. Analyzing social networks as stochastic processes. *Journal of the American statistical association*, **75**(370), 280–294.

Wasserman, S., & Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge Univ Pr.

Wasserman, S., & Iacobucci, D. 1988. Sequential social network data. *Psychometrika*, **53**(2), 261–282.

Wasserman, S., & Pattison, P. 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, **61**(3), 401–425.

Wasserman, S., & Robins, G. 2005. An introduction to random graphs, dependence graphs, and p*. *Models and methods in social network analysis*, 148–161.

Wasserman, S.S. 1980b. A stochastic model for directed graphs with transition rates determined by reciprocity. *Sociological methodology*, **11**, 392–412.

Wellman, B. 1981. Applying network analysis to the study of support. *Social networks and social support*, 171–200.

Wellman, B. 1988. Structural analysis: From method and metaphor to theory and substance. *Social structures: A network approach*, 19–61.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. 2003. Computer networks as social networks: Collaborative work, telework, and virtual community.

White, J.G., Southgate, E., Thomson, JN, & Brenner, S. 1986. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **314**(1165), 1.

Yamamoto, Y., Nakano, J., & Fujiwara, T. 2010. Parallel computing in the statistical system Jasp. *Computational Statistics*, **25**(2), 291–298.

Young, H.P. 2006. The diffusion of innovations in social networks. *Economy as an evolving complex system 3*, 267.

Zijlstra, B.J.H., & van Duijn, M.A.J. 2003. Manual p2. *Version 2.0*.

Zijlstra, B.J.H., van Duijn, M.A.J., & Snijders, T.A.B. 2006. The Multilevel p 2 Model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **2**(1), 42–47.