Analysis of Network Data

A Statistical Analysis of the International Arms Trade Network from 1950-2013

Master's Thesis

Author: Christian Schmid
Supervisor: Prof. Dr. Göran Kauermann
Submission Date: May 19, 2015

Faculty for Mathematics, Informatics und Statistics of the
Ludwig-Maximilians University Munich

Acknowledgments

I would like to thank all the people who contributed in some capacity to the work described in this thesis.

First and foremost, I offer my sincerest gratitude to my supervisor, Prof. Dr. Göran Kauermann, who has supported me throughout my research with his patience and knowledge while allowing me to think independently as well. Without his encouragement and effort, this thesis would not have been possible.

I would also like to sincerely thank Prof. Dr. Paul W. Thurner for his guidance, understanding, patience, and most importantly, his friendship, during my graduate studies. His mentorship has been of paramount importance to me and I owe my involvement in the field of network analysis to his employing me as a research assistant and giving me the opportunity to participate at a network analysis workshop in Zurich. I also would like to thank him for the provision of all the data that were used in this thesis.

A very special thanks goes to Prof. Dr. Detlef Dürr, without whose motivation and encouragement I would not have considered a graduate career. It was under his tutelage that I became interested in mathematics and over the course of my studies I have come to consider him just as much a mentor and friend as a professor. I doubt that I will ever be able to convey my appreciation fully, but I owe him my eternal gratitude.

I want to thank my dear friend Christoph 'Rudi' Jansen, for all the endless hours we spent together in the library these past two years. I will never forget the lively discussions, the long working sessions, and the sometimes even longer computer game sessions. Christoph provided so much of the humor and entertainment that made the difficult and stressful times easier to endure.

I am deeply and forever indebted to my parents, Stefan and Jae-Sook, for their love, support and encouragement throughout my entire life. Their daily hard work made it possible for me to be as academically focused as I wanted to be. I am also very grateful to my sisters Jessica and Gloria and my brothers Maximilian, Sebastian, Amadeus and Godwin for all great memories!

Also, I thank my wife's parents, Jin and Jenny Kim. They, like Jessica and I, are a couple who endured and survived the experience of graduate school and have provided me with unending encouragement and support. I am grateful that I always

felt welcome and loved in their presence and that they gave me the opportunity and space to write significant parts of my thesis in their home.

Finally, and most importantly, I have saved the last word of acknowledgment to my wonderful wife, Jessica. Her support, encouragement, quiet patience, and unwavering love were undeniably the bedrock upon which the past years of my life have been built. Every day I am amazed by her beauty, which is not only visible on the surface, but also runs deep within her. I do not want to forget to mention that without her editing assistance, this thesis would still be full of spelling errors, incorrect comma placements, and phrases that are only intelligible to German-speaking readers.

Abstract

In this thesis we investigate the international arms trade network of major conventional weapons (MCW) between 1950 and 2013. After an introduction to the network theory and some descriptive analysis of the data we will model the arms trade network with the popular and well-known *exponential random graph model* (ERGM). However, we find that in order to guarantee a good model fit, the ERGM has to be extended into a *curved ERGM*. We are going to justify this step by introducing a method to model networks via *generalized additive models* (GAM), models which use smooth functions in order to include the effects of covariates. The estimated smooth functions will verify the use of *geometrically weighted degree* statistics in the ERGM. While discussing the use of GAMs for networks, we will also present a method based on a bootstrapping approach to model networks with logit models. Finally, we present and interpret the results of the fitted models.

Contents

0 Introduction

This thesis considers an approach on modeling the international trade of major conventional weapons (MCW) with statistical network models. Unlike the international trade for conventional products, the armament industry is usually referred to as being particularly unique (see Johannsen and Martinez-Zarzoso [31]). The reason lies in the fact that for the selling of weapons not only do economic factors matter, but political factors also play a crucial role. It is of fundamental importance, which nations are being equipped with another's weapons. Arming the wrong countries can not only endanger a nation's political interests, but also threaten its own national security. On the other hand, as discussed by Brzoska [6] and Moore [38], providing weapons to the right customers can induce economic as well as political advantages. Brauer [4] shows that the main reason is that despite many developing countries trying to establish a domestic military industry, the majority of these countries have not succeeded yet. As a consequence, these countries are still dependent on a few weapon suppliers, which are able to produce and distribute quality goods.

The reasons why two countries trade weapons are diverse. However, scientifically understanding the factors that influence the decisions to supply arms to different countries is fundamental for well-informed debates on the regulation of arms trading. As far as we know, little work has been done in this direction yet. Some first results are provided by the work of Akerman and Seim [1], Comola [7], Johannsen and Martinez-Zarzoso [31] and Willardson [54].

But why should the arms trade network not be analyzed with statistical standard methods and what are the advantages for an approach with statistical network models? The answer depends on a specific basic assumption, which most statistical standard models have in common: the *independence of the observations*. This particular assumption is crucial for the *maximum-likelihood estimation* (MLE) of the parameters in the models and as a consequence is absolutely essential. However, when examining certain network data, such as the international arms trade data, the basic independence assumption is not tenable anymore. Consider two countries, which are at war with each other. In fact, the weapon import of one country forces the other country to act as well. In this case it would be absolutely erroneous to assume that the weapon imports of both countries happen independently from each other. The dependency structure of the actors in the arms trade network is a systematic feature of the data, as opposed to an occasional coincidence. In this context we are talking about *relational data* (see Wasserman and Faust [52]).

Statistical network analysis allows the modeling of exactly these kinds of dependency structures by not treating dependency structures as inaccuracies or measuring errors, but rather including them as a central component of the network models.

This paper is structured as follows: In chapter 1 we introduce some basic definitions in network analysis. Chapter 2 introduces the data sets used in this paper, discusses the data structuring and provides some first descriptive results of the arms trade data set. In chapter 3 the *exponential random graph model* (ERGM) is introduced and some first model fits will be provided. However, we detect that some basic characteristics of the arms trade networks are captured insufficiently. This will result in a generalization of the ERGM, the so-called *curved ERGM* (CERGM), which is introduced in chapter 7.

We are going to justify the step from the regular ERGM to the CERGM by modeling the arms trade network with *generalized additive models* (GAM) and by taking a closer look at the estimated splines smoothers. This leads to the conclusion of a step-wise down weighting of the effect an actor's degree has on forming a new tie. Therefore, we introduce some basic smoothing techniques in chapter 4, deduce the GAM by discussing the generalized linear model (GLM) and the additive model (AM) in chapter 5 and finally introduce an approach on modeling networks with GAMs in chapter 6. In chapter 6 we will furthermore discuss an approach on modeling networks by using a logit model, which estimates the paramters by maximum pseudo-likelihood and circumvents the erroneous independency assumption by adjusting the biased parameter estimates with a bootstrapping technique. After having introduced the CERGM in chapter 7 we will present and interpret the results.

1 Introduction into Network Analysis

In this section we will give a short introduction into some basic terminology of network theory. Therefore, we have to identify networks with the mathematical structure of graphs. As a next step we are going to introduce some definitions for network properties. This chapter is mainly based on Diestel [12] Jansen [30] and Kolaczyk [32].

In order to be able to model networks with statistical methods, one has to identify networks with mathematical structures, the so called *graphs*. In doing so, we can differ between *directed* and *undirected* graphs. Since the international arms trade network will turn out to be a *directed* graph, we will narrow down most of the

definitions for directed networks. We begin with the formal definition of a graph:

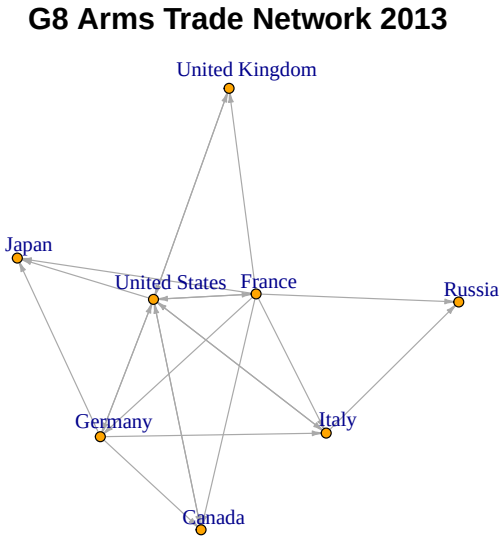
Definition 1. *Let V be a finite set and $E \subset V \times V$. Then, a finit directed graph is defined as the pair $G := (V, E)$. In this context, V is called the set of vertexes and E is denoted as the set of edges. The elements of V are called vertexes or nodes, while the elements of E are called edges or ties.*

The definition of a graph is the perfect tool to bridge from networks into mathematics. The elements of the set of vertexes V are symbolizing the actors in a network. Most of the time we are going to denote actors of the network $v_i, v_j \in V$ simply with their indices i and j. In our case the actors in the network are the countries in the world. The set of edges symbolizes the relation between two actors. This relation can be either directed or undirected. In context of the arms trade network, this relation indicates, whether a country i exports major conventional weapons to country j or not. This conveys that the arms trade network is a directed network: The case that i is selling weapons to j does not imply that j is also selling weapons to i. An example for an undirected network would be the network of the direct contiguity of countries. If country i shares a border with country j, this does also imply that country j is sharing a border with country i.

In the following paragraph we are going to define some terms for a graph $G = (V, E)$, which are crucial in network analysis and which are going to be used in this paper. For an edge $e_{ij} = (v_i, v_j) = (i, j)$, going from actor i to j, we are calling v_i the *tail* and v_j the *head* of edge e_{ij}. Since the networks of consideration are trade networks, we will also refer to v_i as the *sender* or *supplier* and to v_j as the *receiver* or *recipient*. For directed networks edge $e_{ij} = (v_i, v_j)$ has to be distinguished from the edge $e_{ji} = (v_j, v_i)$, since these edges are pointing the opposite direction even though they take place between the same actors i and j. A restriction we are going to make is that a graph has no *loops* $e_{ii} = (v_i, v_i)$, i.e. edges with tail and head on the same vertex. This means that for the arms trade network we are not paying attention to weapons produced for a nations own use. The number of actors $N_V = |V|$ in the network is usually called the *order* of the graph, while the number of edges $N_E = |E|$ is labeled as the *size* of the network. Furthermore, we refer to $N := N_V^2 - N_V = N_V(N_V - 1)$ as the number of possible edges in a directed network. This yields the following definition.

Definition 2. *Let $G = (V, E)$ be a finite ($N_V < \infty$), directed graph. The density $\rho(G)$ of G is defined as*

$$\rho(G) := \frac{N_E}{N}$$



	Ca	Fr	Ger	It	Jap	Rus	UK	USA
Canada	0	0	0	0	0	0	0	1
France	1	0	1	1	1	1	1	1
Germany	1	0	0	1	1	0	0	1
Italy	0	0	0	0	0	1	0	1
Japan	0	0	0	0	0	0	0	0
Russia	0	0	0	0	0	0	0	0
UK	0	0	0	0	0	0	0	1
USA	1	1	1	1	1	0	1	0

Figure 1: Graph and the corresponding adjacency matrix

The density of a network is the proportion between the actual number of edges and the possible number of edges. A *full graph*, i.e. a graph with every possible tie, has density $\rho = 1$, while an *empty graph* is defined as a graph without any edges, $\rho = 0$. However, something one has to be cautious with when comparing the density of two networks with different number of actors is that networks with smaller ρ do not necessarily have less edges. The number of possible ties increases as a quadratic function.

If two vertexes v_i and v_j are connected by an edge $e_{ij} = (v_i, v_j)$ they are called *adjacent*. On the other hand, two edges can also be called adjacent if they share a common vertex, e.g., $e_{ij} = (v_i, v_j)$ and $e_{jk} = (v_j, v_k)$. The term *adjacent* yields the following pivotal definition in network theory, since it enables us to identify the

abstract structure of a graph with the more common and familiar structure of a matrix:

Definition 3. *Let $G = (V, E)$ be a finite, directed graph. Furthermore, let $V = (v_1, \ldots, v_{N_V})$ be an enumeration of the set of vertexes in G. We then define the matrix $A = (a_{ij}) \in \mathbb{R}^{N_V \times N_V}$ with*

$$a_{ij} = \begin{cases} 1 & , \text{ if } (v_i, v_j) \in E \\ 0 & , \text{ else} \end{cases}$$

$i, j \in \{1, \ldots, N_V\}$ as the adjacency matrix of graph G.

Note that a graph is completely specified by an adjacency matrix and vice versa. The graph and its corresponding adjacency matrix can be seen as a different perspective of the same object. This relationship between graph and adjacency matrix is illustrated in figure 1, which illustrates the binarized arms trade network of the G8 nations in 2013. The adjacency matrix can unproblematically be generated from the graph and one can easily draw the corresponding graph using the adjacency matrix.

When looking at the graph in figure 1 it stands to reason to count the ties connected to an actor in order to draw conclusions about the importance of that actor in the network. The definition introduced next will play a key role in the network models discussed later.

Definition 4. *Let $G = (V, E)$ be a finite, directed graph and $v \in V$. Then, the numbers*

$$deg^{in}(v) := |\{(v_i, v_j) \in E : v_j = v\}|$$

$$deg^{out}(v) := |\{(v_i, v_j) \in E : v_i = v\}|$$

are called the in-degree and out-degree of vertex v.

Thus, the in-degree of a node v is defined as the number of edge heads ending at v. On the other side, the out-degree is defined as the number of tails connected to v. Note that a node's in-degree can easily be calculated by adding up the node's column in the adjacency matrix. When looking at the adjacency matrix in figure 1 one can easily see that the UK has an in-degree of 2, purchasing weapons from France and the US. Similarly, one gets an actors out-degree by adding up the corresponding row.

The next term introduced is the definition of a *dyad*. A dyad is defined as a group of

two actors and their relation. Thus, a dyad can be seen as the smallest possible unit in network analysis, since it is a network consisting of only two actors. For directed networks we are going to differentiate between three kind of dyads: We are going to call a dyad (ij) *mutual* or *reciprocal* if there is an edge going from i to j and from j to i, i.e., $e_{ij}, e_{ji} \in E$. A dyad is called *asymmetric* or *one-sided* if there is only one edge between the two actors, i.e., $e_{ij} \in E \veebar e_{ji} \in E$, where $\veebar$ is defined as *exclusive 'or'*. Lastly, a dyad is called *null* if there is no edge between two actors i and j, i.e., $e_{ij}, e_{ji} \notin E$.

Finally, we are going to define the *geodesic distance* in terms of network analysis, which will be especially important for the *goodness-of-fit* considerations in chapter 3.6, 6.2, 6.3 and 7.2. In order to do so, we first have to define what we call a *path*.

Definition 5. *Let $G = (V, E)$ be a finite, directed graph. A path from v_0 to v_ℓ is defined as a sequence $(v_0, e_1, v_1, e_2, \ldots, v_{\ell-1}, e_\ell, v_\ell)$, where $v_i, v_j \in V$, $v_i \neq v_j$ and $e_i \in E$, $e_i := (v_{i-1}, v_i)$ for all $i, j \in \{1, \ldots, \ell\}$.*
The number ℓ is called the length of the path.

The definition of a path enables the definition of the geodesic distance.

Definition 6. *Let $G = (V, E)$ be a finite, directed graph and let $v_1, v_2 \in V$ be two nodes in the network. The geodesic distance $dist(v_1, v_2)$ between two nodes is the length ℓ of one of the shortest paths from v_1 to v_2. If there is no path between v_1 and v_2 define $dist(v_1, v_2) = \infty$.*

The geodesic distance is defined as the length of one of the shortest paths between two nodes v_1 and v_2. One has to be careful with the fact that for directed networks the geodesic distance between v_1, v_2 does not necessarily imply $dist(v_1, v_2) = dist(v_2, v_1)$. Furthermore, one can easily certify that the shortest path does not need to be unique. If one takes, for instance, the G8 trade network in figure (1) and determines the geodesic distance from the USA to Russia one recognizes easily that there are two paths with the length of 2 between these two countries. One through France and the other one via Italy.

2 Data Sources and Structuring

The international arms trade data for major conventional weapons was provided by the Stockholm International Peace Research Institute (SIPRI), a Swedish think tank specializing in research on international conflict, armaments, arms control and disarmament. SIPRI was established in 1966 on the basis of a decision by the Swedish

government. In addition to their headquarters in Stockholm they also have a presence in Beijing. See [49] for more information on SIPRI.

The initial data set is available for the time period 1950-2013 (as of 02/2015) and describes the international trade of major conventional weapons along a variety of attributes such as country and year. This means that weapons produced for a nations own personal use are not considered in the data. In order to measure the volume of international transfers of arms, SIPRI has developed a unique system. The data is listed in *trend indicator value* (TIV), a measure which is based on production costs (see Holtom et al. [24]).

One might ask the question why the data is not given in a monetary value, like constant USD or the like, which mirrors actual cash flow. According to SIPRI there are several reasons why measuring the cash flow would lead to distorted information. One main reason is certainly that for major conventional weapons there is not catalog price and, as a consequence, unit prices are usually negotiated individually. Furthermore, trading weapons often has underlying political or economic reasons. One can imagine that a supplier nation might be willing to send weapons in order to guarantee power equilibrium or to assert personal interests in a certain area. For instance, this effect can be seen when looking at the proxy wars which were happening during the Cold War between the Soviet Union and the United States. An example of economic motivations for sending weapons to a different nation is the deal between Germany and Saudi Arabia in 2014. Germany sent Leopard-2 tanks to the Arabian country and in return Germany received further oil supply benefits. Therefore, considering the cash flow would not mirror the actual value of the delivered tanks. Furthermore, according to SIPRI, the TIV has the other crucial advantage of being consistent over time, which makes it possible to compare the arms flow of different time periods. The aggregated trade volumes for each year are visualized in figure 2. One can see that from the beginning of the data acquisition the amount of weapon trades increased until it reached its peak in 1982. From this year until the mid-nineties the amount of yearly traded weapons decreased. In the twenty-first century we can again observe a clear increase in the volume of traded weapons.

In order to get a first rough impression for the values of the traded goods some examples shall be listed. For instance, a Leopard 2A4-tank is worth 4 million TIV, a Eurofighter has the value of 55 million TIV and a 209PN submarine is registered with 275 million TIV. Secondhand weapons get the value of two-fifths of the origi-

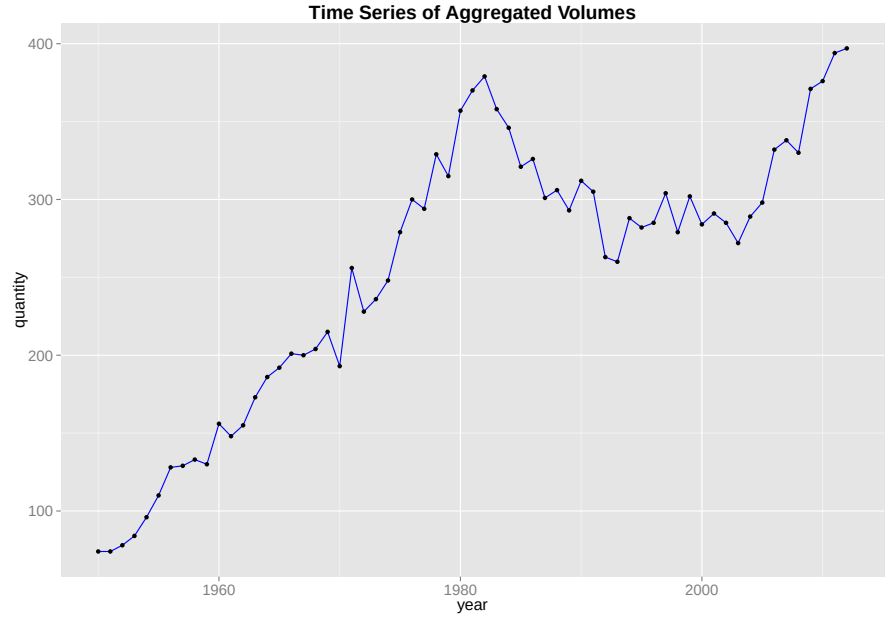


Figure 2: The aggregated trade volumes in million TIV for the time period 1950-2013

nal price, while used units which were significantly refurbished or modified by the supplier nation before delivery is given two-thirds of the original price. As a consequence, this measure enables the possibility of calculating trends and comparing the arms trade activities of different nations. In table 1 the top 10 supplier and the top 10 recipient nations are listed for the time period 2009-2013. One can easily discern that even twenty years after the end of the Cold War the international arms trade is characterized by two main actors: the United States and Russia. On the other hand, India as the main recipient of major conventional weapons strikes the attraction, which seems to be a result of the still-ongoing Kashmir conflict with Pakistan, a country ranked third in the recipient table, and consistently arising border conflicts with China, a country ranked second in the recipient table. In general one can ascertain by taking a look at figure 3 that the majority of arms imports are delivered to countries in Asia/Middle East.

When examining the international arms trade one will recognize that not only countries are involved in the network. In fact, international organizations like the UN and NATO, extremist groups like Al Quaida, Hamas, Hezbollah and embattled areas like Chechnia, Darfur or even Eastern Ukraine can actively be involved in the network. However, according to Akerman et al. [1] these trade flows are negligible

	Recipients	
	Land	TIV
1	India	18563.80
2	China	6581.37
3	Pakistan	6425.93
4	United Arab Emirates	5774.80
5	Saudi Arabia	5229.99
6	United States	5072.71
7	Australia	4792.76
8	South Korea	4752.28
9	Singapore	4438.57
10	Algeria	4226.95

	Suppliers	
	Land	TIV
1	United States	37660.46
2	Russia	36243.01
3	Germany	8619.34
4	China	7379.65
5	France	7195.38
6	United Kingdom	5510.32
7	Spain	3886.92
8	Ukraine	3502.08
9	Italy	3456.58
10	Israel	3156.09

Table 1: The left table lists the top 10 recipients and the right the top 10 supplier nations for the time period 2009-2013

and as a consequence, we decided not to consider them in this thesis. A list of all possible arms trading actors and a list of all excluded embattled areas are given in appendix 9.3 and 9.4.

After having defined the set of actors in the networks, one has to face the fact that some actors did not exists during the whole time period of consideration. For example, the German Democratic Republic disappears from the scene in 1990, while other countries like Estonia and Kazakhstan (re-)gain their independence in 1991. Later, we are going to model the arms trade network on an annual basis for the time period $1950 - 2013$. In order to adequately model these networks, we implement a function in R, which excludes every country from the list of all actors (appendix 9.3) that did not exist in the particular year of consideration. The time period through which a country is included into the models can also be found in appendix 9.3.

The left plot in figure 4 shows the number of actors in the network for each year. There is a conspicuous constant growth of actors from the 1960s until 1980 due to decolonization and a big jump from 1990 to 1991 as a consequence of the break-up of the Soviet Union. On the right side we have visualized the time series of the network's density. Even as we are careful while interpreting this plot, since the number of actors changes over time, one can observe similarities to the time series of the aggregated traded volumes. Just as in figure 2, we can see a peak in 1982, followed by a decrease until the nineties and a rise in the past ten years.

The data in their initial form are not suitable for network analysis, since they are not in adjacency form, and, as a consequence, have to be transformed. An adjacency

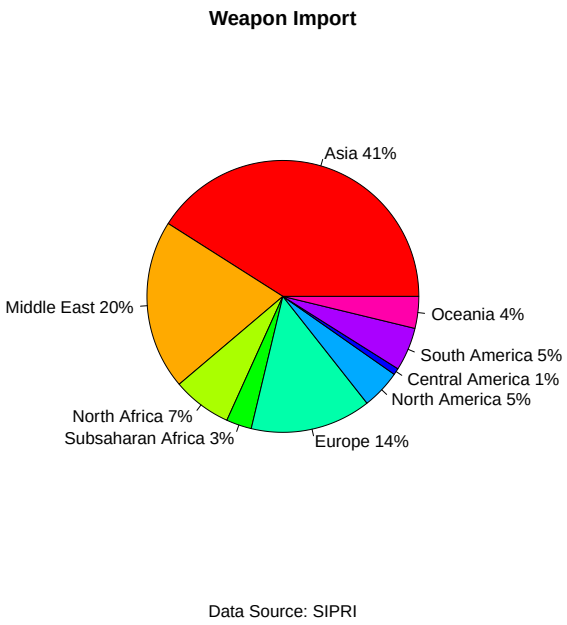


Figure 3: Distribution of worldwide weapon imports in the time period 2009-2013

matrix is a matrix consisting of 1s and 0s where each particular actor in the network is assigned both a row and a column (see definition 3). In our case, a 1 indicates that the nation in the corresponding row is selling weapons to the nation in the corresponding column, while a zero indicates that there is no arms flow from the row nation to the column nation. Since an adjacency matrix only allows binary coding, setting a threshold is neccessary. In this paper we fix the threshold at one million TIV, which means that all trade flows considered more than one million TIV are indicated with a 1 while all others get labeled by a 0. Setting a threshold enables us to distinguish between proper weapon purchases and acquisitions simply made in order to maintain already existing weapons. Experience with the data showed that fixing the threshold at one million TIV is satisfactory for our purposes.

When taking a look at the network graphs in figure 26 and 27 in the appendix one can observe quite easily that the majority of countries, which are actively involved in the arms trade network are only receiving weapons and are not selling their goods to other countries. For the sake of clarity, the countries that did not trade weapons in these years were excluded from the plot. This finding gets even further support by the fact that around 95% of all dyads which are not null, are one-sided. Only the minority of non-null edges are mutual (see Jansen and Schmid [29]). Furthermore,

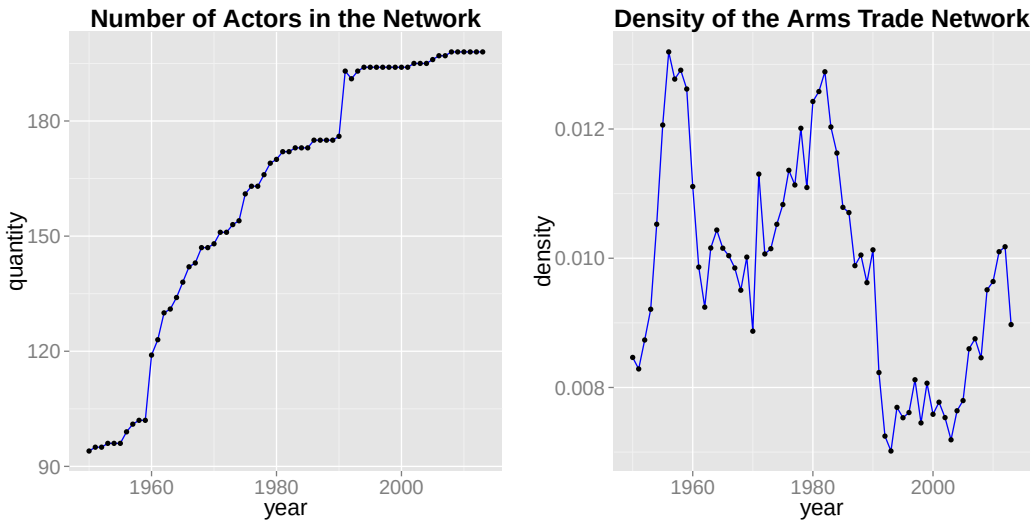


Figure 4: The number of actors included in the arms trade networks (left) and the density of the networks (right) for the time period 1950-2013

by looking at the networks in figure 26 and 27 it catches one's eye that there are a few actors which sell weapons to a vast number of other actors. The total export TIV of the Top 10 weapon selling nations from table 1 corresponds to 88.3% of the total export TIV of all nations in this time period. The United States and Russia alone are each responsible for about one-third of the global arms exports.

Figure 5 visualizes the average in-degree and out-degree distribution for the time period 1950-2013 in percentages. Plotting the distribution on a percentage scale enables a comparison of the distributions for different networks with a different number of actors. In each case, 90% of the corresponding degree value was situated between the black bars. These figures visualize what we already discussed in the paragraph before. The majority of countries are not selling weapons, which can be derived from the fact that over 80% of all actors in the network have an out-degree of zero. On the other hand, the majority of the actors have an in-degree of 0 as well, but this is mainly due to the countries, which are not involved in any weapon trades at all. Besides, there is still a high percentage of actors which are only purchasing weapons from one supplier. These countries usually do not sell weapons and are therefore contingent on a single supplier. We refer to these nodes in the network as *satellites*.

In this paper we will also incorporate other data sets, which are included as ex-

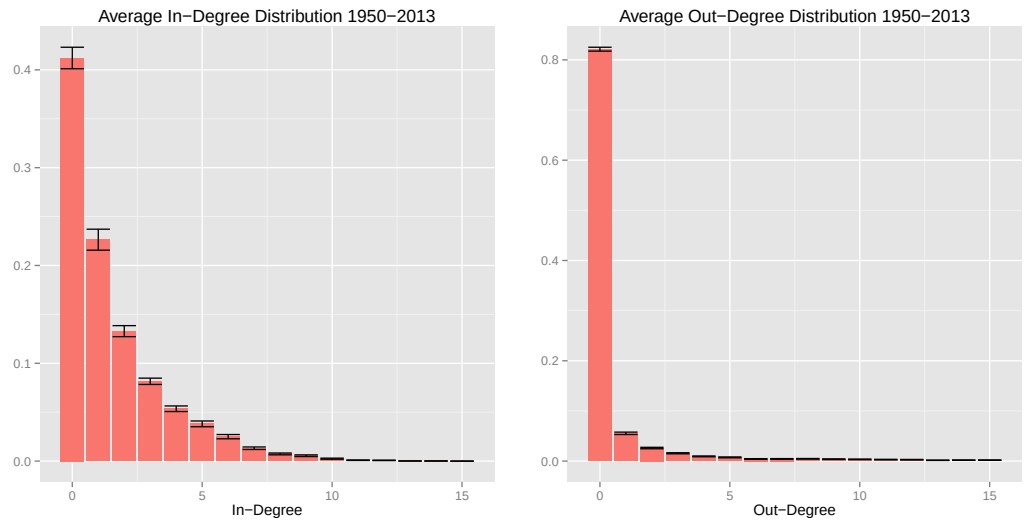


Figure 5: The average in-degree and out-degree distribution for the time period 1950-2013. In each case 90% of the corresponding degree value was situated between the black bars

ogenous covariates. One is the *Formal Interstate Alliance Dataset* provided by Gibler [17], which is part of the Correlates of War (COW) Project. This data set provides insight into the content of military alliance agreements signed by any nation from 1815 on. We incorporate these data as symmetric adjacency matrices for each year, where a 1 indicates that nations i and j signed a *defense agreement*, while a 0 denotes that the corresponding nations have not.

The next data set included into models for this paper contains data from the *Polity IV Project*, which is provided by the *Center for Systemic Peace* (CSP) [34]. This data assigns a *democracy score* between 10 and -10 to each nation on an annual basis, depending on its democratic status. A 10 indicates that nation i has the highest democracy standards while a -10 means the opposite. We created a weighted adjacency matrix with the absolute difference in democracy score between nations i and j as entries.

Furthermore, we include a covariate we are going to refer to as *direct contiguity*. This covariate is a relational covariate (see chapter 3.5) similar to the defense agreement data, and it indicates whether two nations i and j share a common border or not. This data set does not only embrace land borders, but also sea borders. Just as

for the alliance covariate, a 1 indicates that there exists a relation between i and j which in this case means that there is a common border. To the contrary, a 0 indicates that there is no common border between i and j.

Additionally, we use the *GeoDist* Dataset from CEPII [35], the primary French institute for research into international economics. This dataset includes the geographic distance between nations' capitals (measured in kilometers) by using the great circle formula, which uses latitudes and longitudes. Since we assume that these data and the direct contiguity data are dependent on each other, the GeoDist data is only included into the models if it is mentionend particularly. Otherwise, the direct contiguity data is used.

A nodal attribute is the *Composite Index of National Capability* (CINC) from the *National Material Capabilities Dataset* [46]. The CINC is a statistical measure of national power created for the COW project. It uses an average of percentages of world totals in six different components, which represent demographic, economic, and military strength. These components are: total population, urban population, iron and steel production, primary energy consumption, military expenditure, and military personnel. As described by Perkins and Neumayer [40] we include this data curvilinearly (CINC^2+ CINC).

The next covariate in our model gathers information about inter- and intra-state conflicts and includes all episodes of international, civil, ethnic, communal, and genocidal violence and warfare. The data come from the *Major Episodes of Political Violance Project* and are also provided by the CSP [34]. The conflicts are coded on a scale of one to ten according to an assessment of the full impact of their violence on the societies that directly experienced their effects. We distinguish between the inter- and intra-state conflicts by incorporating the inter-state conflicts as relational data and the intra-state conflicts as nodal covariates (see chapter 3.5).

We also include the *Arms Embargoes Dataset*, which is also provided by SIPRI [49]. This database gives information on all multilateral arms embargoes that have been implemented by an international organisation, such as the EU or UN, or by a group of nations. It includes both legally binding embargoes and those that are solely political commitments. However, arms embargoes may be in place for only part of a year, while data on arms transfer is available on a yearly basis. Therefore, in order to prevent legal arms tranfer from biasing the results, only embargos which were imposed for a full calendar year are included. A 1 indicates that country i has an embargo against country j, while a 0 indicates that i does not have an embargo

against j.

Finally, we use the nations' *gross domestic product* (GDP) per capita in US dollars from *The Maddison Project* dataset [50]. In order to make this data more accessible for the networks, we shrink the given numbers by taking the natural logarithm. The Maddison Project is, to our knowledge, the only dataset that also covers socialist and communist countries from before 1990.

3 The Exponential Random Graph Model (ERGM)

The network model we are going to introduce is the *exponential random graph model* (ERGM), which is a probability model for directed or undirected binary models. This means neither the weighting nor the temporal change of ties is considered in the model. In literature, ERGMs are sometimes also referred to as *p-star* or p^* models (see Wassermann and Pattison [53]). Therefore, we consider p-star as a synonym for ERGM. In the following, we will introduce the ERGM for directed networks. This chapter is mainly based on Harris [22], Hunter et al. [26], Hunter [28] and Jansen and Schmid [29].

3.1 The ERGM

In contrast to many other network models the ERGM takes the adjacency matrix of an observed network A^{obs} as the manifestation of a matrix-like random variable Y. According to definition (3) a network of N_V nodes can be defined as adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{N_V \times N_V}$, where $a_{ij} \in \{0, 1\}$ for all $i, j \in \{1, \dots, N_V\}$. $a_{ij} = 1$ means that there is an edge going from actor i to actor j, while $a_{ij} = 0$ indicates that this edge does not exist. Since the model does not involve loops, one has $a_{ii} = 0$ for all $i \in \{1, \dots, N_V\}$. Recall that we simply write i for an actor $v_i \in V$ as long as it is not causing any confusion. Furthermore, define

$$\mathcal{A}(N_V) := \left\{ A \in \mathbb{R}^{(N_V \times N_V)} : a_{ij} \in \{0, 1\},\ a_{ii} = 0 \right\}$$

as the set of all possible networks on N_V nodes without loops. Note that the cardinality of set $\mathcal{A}(N_V)$ is increasing exponentially for every newly included actor, which results in $2^{N_V(N_V-1)}$ elements. Therefore, for an already small number of actors the cardinalty of $\mathcal{A}(N_V)$ turns out to be an astronomically big number. With the definition of $\mathcal{A}(N_V)$ we define

$$Y : \Omega \to \mathcal{A}(N) \quad , \quad \omega \mapsto (Y_{ij}(\omega))_{i,j=1,\dots,N}$$

as a matrix-like random variable. As the probability function from Y to $\mathcal{A}(N_V)$ we define

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\theta^T \cdot \Gamma(A))}{\sum_{A^* \in \mathcal{A}(N)} \exp(\theta^T \cdot \Gamma(A^*))} \tag{1}$$

where

- $\theta \in \mathbb{R}^q$ is a $q-$dimensional vector of parameters
- $\Gamma : \mathcal{A}(N) \to \mathbb{R}^q \ , \ A \mapsto (\Gamma_1(A), \ldots, \Gamma_q(A))^T$ is a q-dimensional function of different network statistics
- $c(\theta) := \sum_{A^* \in \mathcal{A}(N)} \exp(\theta^T \cdot \Gamma(A^*))$ is a normalization constant which ensures that (1) defines a probability function on $\mathcal{A}$

As already mentioned, a specific network A can be considered as a manifestation of a matrix-like random variable, whose probability of occurrence can be modeled with equation (1). A key role when modeling an ERGM is played by the function $\Gamma(\cdot)$. The decision about which network statistics are incorporated into the model affects the model significantly. The selection of endogenous network statistics should be the result of a meticulous analysis of the observed network, since including the wrong statistics can easily cause *degeneracy* problems (see Handcock [19]). We will discuss the meaning of degeneracy at a later point.

Since the adjacency matrix A can be understood as a manifestation of a matrix-like random variable Y, the individual entries a_{ij} of A can be taken as a manifestation of single Bernoulli variables Y_{ij}. This interpretation allows the following calculation regarding the conditional distribution of Y_{ij}:

$$\begin{aligned}
\frac{\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = 0 | Y_{ij}^c = A_{ij}^c)} &= \frac{\mathbb{P}_\theta(Y_{ij} = 1 \ , \ Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = 0 \ , \ Y_{ij}^c = A_{ij}^c)} \\
&= \frac{\mathbb{P}_\theta(Y = A_{ij}^+)}{\mathbb{P}_\theta(Y = A_{ij}^-)} \\
&= \frac{\exp(\theta^T \cdot \Gamma(A_{ij}^+))}{\exp(\theta^T \cdot \Gamma(A_{ij}^-))} \\
&= \exp(\theta^T \cdot (\Gamma(A_{ij}^+) - \Gamma(A_{ij}^-))
\end{aligned}$$

This implies the following equation:

$$\text{logit}(\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)) = \theta^T \cdot (\Gamma(A_{ij}^+) - \Gamma(A_{ij}^-)) \tag{2}$$

In the equation above the following notations were used:

- A_{ij}^{+} emerges from A, while assuming $a_{ij} = 1$
- A_{ij}^{-} emerges from A, while assuming $a_{ij} = 0$
- The condition $Y_{ij}^{c} = A_{ij}^{c}$ is short for: $Y_{pq} = a_{pq}$ for all $(p, q) \in \{1, \ldots, N\}^2$ with $(p, q) \neq (i, j)$
- The expression $(\Delta A)_{ij} := \Gamma(A_{ij}^{+}) - \Gamma(A_{ij}^{-})$ is called the *change statistic*. The kth component of $(\Delta A)_{ij}$ captures the difference between the networks A_{ij}^{+} and A_{ij}^{-} on the kth integrated statistic in the model

As will be illustrated later more precisely, *covariates* can also be included into the model via $\Gamma(\cdot)$. Depending on whether a statistic incorporated into the model uses external covariate information or is based on mere structural network characteristics, one differentiates between *exogenous* and *endogenous* network statistics. Notice that for the sake of simplicity we did and will not condition on exogenous network statistics in this and in the following chapters.

3.2 Parameter Estimation

How can a parameter vector θ be estimated? A first idea could be the following: One can assume that the *dyads* are independent of each other, which means that the random variables Y_{ij} inside the random matrix Y are independent of each other. In this case, the equation (2) reduces to

$$\text{logit}(\mathbb{P}_\theta(Y_{ij} = 1)) = \theta^T \cdot (\Delta A)_{ij}$$

This corresponds with the *logistic regression* approach, where the observations of the dependent variables are simply edge values of the observed adjacency matrix, and the observations of the covariate values are given as the scores of every single change statistic. Therefore, the following structure of the data is given by

$$(a_{ij}, (\Delta(A))_{ij}) \quad \text{for} \quad i, j \in \{1, \ldots, N\}$$

and the estimation of θ can then be obtained as usual using maximum-likelihood estimation. The resulting likelihood function is of the following form:

$$\text{lik}(\theta) = \mathbb{P}_\theta(Y = A) = \prod_{i,j} \frac{\exp\left(\theta^T \Delta(A))_{ij}\right)}{1 + \exp\left(\theta^T \Delta(A))_{ij}\right)} \tag{3}$$

The problem with this simple estimation procedure is that the assumed hypothesis of the independence of the dyads turns out to be erroneous in most cases. This is a systematic problem: The presence of network data is inextricably connected with the presence of *relational data*, which by definition should not be assumed to be independent of each other. If this dependency structure is deliberatley ignored and equation (3) is used to estimate θ, it results in a *pseudo-likelihood estimation*. This technique tends to underestimate the standard error. However, Desmarais and Cranmer [11] show that the pseudo-likelihood provides a consistent approximation of the maximum likelihood.

There are several techniques to circumvent estimators, which underestimate the standard error of θ. In the following, we will introduce a technique based on *Markov Chain Monte Carlo (MCMC)* and maximum-likelihood methods. Later in chapter 6.2 we are going to discuss an approach based on a *bootstrapping* technique, which can also be applied for the ERGM (see Leifeld et al. [33] or Desmarais and Cranmer [10] for further details).

The more rigorous technique is to estimate the parameters directly with the log-likelihood function derived from (1), which has the following form:

$$\text{loglik}(\theta) = \theta^T \cdot \Gamma(A) - \log(c(\theta)) \tag{4}$$

where A is the observed network. For the vector of network statistics, one can assume without loss of generality

$$\Gamma(A) = 0 \tag{5}$$

The reason is the following: If (5) does not apply to the vector of chosen network statistics $\Gamma(\cdot)$, replace $\Gamma(\cdot)$ in (1) with the new network statistic

$$\Gamma^*(\cdot) := \Gamma(\cdot) - \Gamma(A)$$

With this replacement, the probability function of Y remains the same, since after simple recalculation:

$$\frac{\exp(\theta^T \cdot \Gamma(A))}{c(\theta)} = \frac{\exp(\theta^T \cdot \Gamma^*(A))}{c^*(\theta)}$$

where $c^*(\theta) := \sum_{A^* \in \mathcal{A}(N)} \exp(\theta^T \cdot \Gamma^*(A^*))$. This means that *centering* the vector of network statistics does not affect the probability function of the network variable

Y. Therefore, in context of the likelihood function (4) the vector of statistics can always be assumed to be centered around the observed network.
Due to assumption (5), one gets from(4) the simplified log-likelihood function

$$\text{loglik}(\theta) = -\log(c(\theta)) \tag{6}$$

The problem resulting from estimating the parameters with (4) is that the term

$$c(\theta) := \sum_{A^* \in \mathcal{A}(N_V)} \exp(\theta^T \cdot \Gamma(A^*))$$

which sums up the weighted network statistics of all possible networks of N_V nodes, has to be evaluated. Even for networks with small numbers of nodes this presents an enormous computational obstacle, and the necessary calculations for larger networks can not currently be completed in any reasonable timeframe. The arms trade networks we are going to model have an average of 150 actors and therefore, we would have to compute $\exp(\theta^T \cdot \Gamma(A^*))$ for about 2^{22350} networks in order to obtain $c(\theta)$. An astronomically big number! As a result, for sufficiently large networks it is not possible to estimate the parameters directly with the likelihood function.

An expedient for this limitation is based on the following consideration: Fix a vector of parameters $\theta_0 \in \Theta$ from the underlying parameter range Θ and compute for $\theta \in \Theta$ the expected value

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\exp\left((\theta-\theta_0)^T \cdot \Gamma(Y)\right)\right] &= \sum_{A \in \mathcal{A}(N)} \exp\left((\theta-\theta_0)^T \cdot \Gamma(A)\right) \cdot \mathbb{P}_{\theta_0}(Y = A) \\
&= \sum_{A \in \mathcal{A}(N)} \exp\left((\theta-\theta_0)^T \cdot \Gamma(A)\right) \cdot \frac{\exp(\theta_0^T \cdot \Gamma(A))}{c(\theta_0)} \\
&= \frac{1}{c(\theta_0)} \sum_{A \in \mathcal{A}(N)} \exp\left(\theta^T \cdot \Gamma(A)\right) \\
&= \frac{c(\theta)}{c(\theta_0)}
\end{aligned}$$

One gets the equation

$$\frac{c(\theta)}{c(\theta_0)} = \mathbb{E}_{\theta_0}\left[\exp\left((\theta-\theta_0)^T \cdot \Gamma(Y)\right)\right] \tag{7}$$

Equation (7) offers the following possibility: If one draws L random networks $A_1, \ldots, A_L$ out of a distribution $\mathbb{P}_{\theta_0}$ appropriately, one gets with the *law of big*

numbers the following relation:

$$\frac{1}{L} \cdot \sum_{i=1}^{L} \exp\left((\theta - \theta_0)^T \cdot \Gamma(A_i)\right) \quad \xrightarrow{\mathbb{P}_{\theta_0}} \quad \mathbb{E}_{\theta_0}\left[\exp\left((\theta - \theta_0)^T \cdot \Gamma(Y)\right)\right] = \frac{c(\theta)}{c(\theta_0)} \tag{8}$$

For a big enough number, L, of random networks, the following approximation is reasonable:

$$\frac{c(\theta)}{c(\theta_0)} \approx \frac{1}{L} \cdot \sum_{i=1}^{L} \exp\left((\theta - \theta_0)^T \cdot \Gamma(A_i)\right) \tag{9}$$

One can now use equation (9) to determine an approximation of the log-likelihood function (6):

$$\begin{aligned}
\text{loglik}(\theta) - \text{loglik}(\theta_0) &= -\log(c(\theta)) + \log(c(\theta_0)) \\
&= -\log\left(\frac{c(\theta)}{c(\theta_0)}\right) \\
&= -\log\left(\mathbb{E}_{\theta_0}\left[\exp\left((\theta - \theta_0)^T \cdot \Gamma(Y)\right)\right]\right) \\
&\approx -\log\left(\frac{1}{L} \cdot \sum_{i=1}^{L} \exp\left((\theta - \theta_0)^T \cdot \Gamma(A_i)\right)\right)
\end{aligned}$$

By differentiating this equation on both sides with respect to θ one gets an approximate score function:

$$s(\theta) \approx -\frac{\partial}{\partial\theta} \log\left(\frac{1}{L} \cdot \sum_{i=1}^{L} \exp\left((\theta - \theta_0)^T \cdot \Gamma(A_i)\right)\right) \tag{10}$$

This approximate score function now can be used as usual, i.e., it can be iteratively approximately optimized with the *Newton-Raphson algorithm*. As a result, the approximate maximum-likelihood estimator for the parameters can be computed.

As pleasant as this may sound, the immediate question arises: How can a sufficient number of suitable drawings $A_1, ..., A_L$ be taken from the distribution $\mathbb{P}_{\theta_0}$?
For this purpose, the *Markov Chain Monte Carlo (MCMC)* methods can be used. The application of MCMC methods for the simulation of random networks is discussed in the next chapter.

3.3 Simulation of random networks

To be able to compute the approximate likelihood function, which was established in the last paragraph, one needs a sufficiently large number of random networks from the distribution $\mathbb{P}_{\theta_0}$ of the matrix-like random variable Y. Snijders [48] introduces

an approach to select these random networks by using *MCMC methods*.

No matter which kind of MCMC algorithm is used, the basic idea is the following: One constructs a *Markov chain* $(Y_t)_{t\in\mathbb{N}}$ on the set of all possible networks $\mathcal{A}(N)$ of N nodes, whose *stationary distribution* is in conformity with the distribution $\mathbb{P}_{\theta_0}$. One can show that every single realization (or *trajectory*)

$$(A_t)_{t\in\mathbb{N}} := (Y_t(\omega))_{t\in\mathbb{N}}$$

of this stochastic process accomplishes the convergence result (8) (for this version of the *Law of big numbers for Markov chains* we reference Meyn and Tweedie [37]). As a result, sub-sequences of $(A_t)_{t\in\mathbb{N}}$ which are sufficiently large enough can be used for approximation (9).

But how can one construct suitable trajectories of Markov chains from $\mathcal{A}(N)$? To answer this question, two common algorithms are introduced, the *Gibbs Sampling method* and the *Metropolis-Hastings algorithm*.

Gibbs Sampling Method

Begin by choosing a start matrix $A^{(0)} \in \mathcal{A}(N)$ (for instance, the observed matrix could be chosen). Afterwards, the length L of the respective sub-sequence is determined. For $k \in \{0, ..., L-1\}$ execute the following steps recursively (here the network in its kth iteration is denoted as $A^{(k)}$):

1. Randomly choose an edge (i, j) where $i \neq j$ from $A^{(k)}$.
2. Compute with equation (13) the value

$$\pi := \mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = (A_{ij}^{(k)})^c)$$

3. Draw a random number Z from $\text{Bin}(1, \pi)$. If
   - $Z = 0$, define $A^{(k+1)}$ via

$$a_{pq}^{(k+1)} = \begin{cases} 0 & \text{if } (p,q) = (i,j) \\ a_{pq}^{(k)} & \text{if } (p,q) \neq (i,j) \end{cases}$$

- $Z = 1$, define $A^{(k+1)}$ via

$$a_{pq}^{(k+1)} = \begin{cases} 1 & \text{if } (p,q) = (i,j) \\ a_{pq}^{(k)} & \text{if } (p,q) \neq (i,j) \end{cases}$$

4. Start at step 1 with $A^{(k+1)}$.

The depicted algorithm provides a sequence of random networks $A^{(0)}, ..., A^{(L)}$. Since the orignial matrix was chosen randomly and the first simulated networks are very dependent on the chosen matrix (only one edge is changed per iteration!), usually the first B networks, where $N \ll B \ll L$, are discarded as the so called *Burn-In*.

Metropolis-Hastings Algorithm

Again, choose a matrix $A^{(0)} \in \mathcal{A}(N)$ to start with (e.g., the observed network). For $k \in \{0, ..., L-1\}$ recursively proceed as follows:

1. Randomly choose an edge (i,j) where $i \neq j$ from $A^{(k)}$
2. Compute, using the equation (2) the value

$$\pi := \frac{\mathbb{P}_\theta(Y_{ij} \neq a_{ij}^{(k)} | Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = a_{ij}^{(k)} | Y_{ij}^c = A_{ij}^c)}$$

3. Fix $\delta := \min\{1, \pi\}$ and draw a random number Z from $\text{Bin}(1, \delta)$. If
   - $Z = 0$, let $A^{(k+1)} := A^{(k)}$
   - $Z = 1$, define $A^{(k+1)}$ via

$$a_{pq}^{(k+1)} = \begin{cases} 1 - a_{pq}^{(k)} & \text{if } (p,q) = (i,j) \\ a_{pq}^{(k)} & \text{if } (p,q) \neq (i,j) \end{cases}$$

4. Start at step 1 with $A^{(k+1)}$.

Similar to the Gibbs Sampling method, the first B networks are discarded as Burn-In.

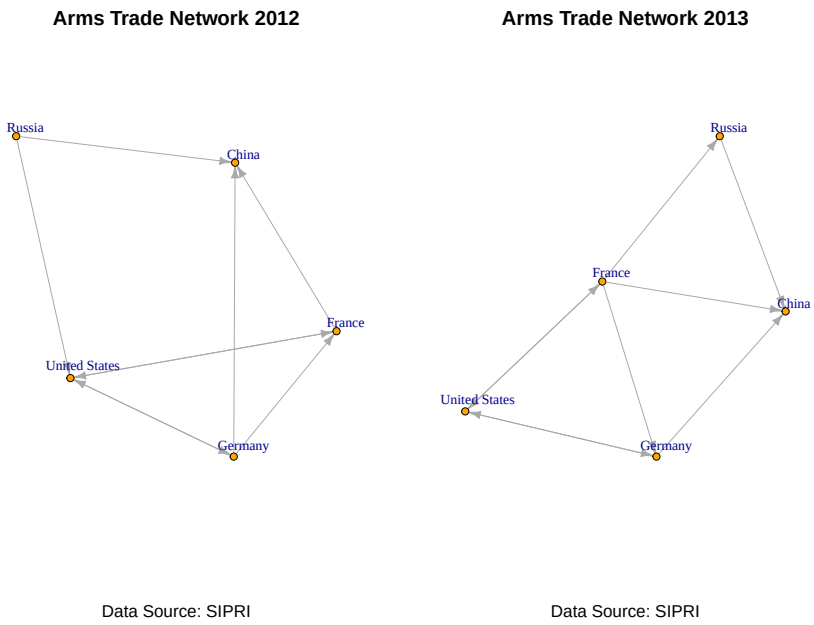


Figure 6: The trade network for 2012 and 2013 for the five main major conventional weapon supplier nations

3.4 Parameter Interpretation

After considering how the vector of parameter θ can be estimated, we now address to the interpretation of these estimates. Interpreting the parameters of an ERGM can be done on two different levels: on the edge level and on the network level. The following discussion is based on Cranmer and Desmarais [8].

We begin by discussing how the parameters of the ERGM can be interpreted on the network level. Therefore, for a network A let A^{k^-} be the network

$$\Gamma_\ell(A^{k^-}) = \begin{cases} \Gamma_\ell(A) & \text{, if } \ell \in \{1, \dots, q\}/k \\ \Gamma_\ell(A) - 1 & \text{, if } \ell = k \end{cases}$$

A^{k^-} is defined as a network where all statistics except the kth one get assigned the same value as in network A and the kth statistic of A^{k^-} is by one smaller than the corresponding statistic of A.

As an example, consider an ERGM with only two statistics: the number of edges and the number of actors with in-degree= 1. A network A^{2^-} is a network, which has the same number of edges as network A, but the number of actors with an in-degree of

1 is one smaller than in network A. The networks in figure 6 are exactly of this kind. They illustrate the network among the five main suppliers of major conventional weapons for the years 2012 and 2013. One can easily assure oneself that the number of edges is 9 in both networks, but the number of actors with in-degree of 1 differ. In the 2012 network only Germany has an in-degree of 1 while in the 2013 network two countries, France and Russia, receive arms from only one of the other four top weapon-selling nations. Therefore, the 2012 network can be written as A^{2^-} of the 2013 network A.

To consider the odds of occurrence of network A compared to A^{k^-}, one realizes through equation (1) the following relationship:

$$\begin{aligned} \frac{\mathbb{P}_\theta(Y = A)}{\mathbb{P}_\theta(Y = A^{k^-})} &= \frac{\sum_{l=1}^{q} \theta_l \cdot \Gamma_l(A)}{\sum_{l=1}^{q} \theta_l \cdot \Gamma_l(A^{k^-})} \\ &= \frac{\exp(\theta_k \cdot \Gamma_k(A))}{\exp(\theta_k \cdot (\Gamma_k(A) - 1))} \\ &= \exp(\theta_k) \end{aligned}$$

Meaning that

$$\frac{\mathbb{P}_\theta(Y = A)}{\mathbb{P}_\theta(Y = A^{k^-})} = \exp(\theta_k) \tag{11}$$

Equation (11) now can be interpretated as follows: The relative plausibility that network A occurs instead of network A^{k^-} is $\exp(\theta_k)$. The higher the value of $exp(\theta_k)$, the more plausible network A is compared to A^{k^-}. This yields the following interpretation for a parameter θ_k:

- if $\theta_k > 0$, then network A is more plausible than network A^{k^-}
- if $\theta_k = 0$, then both networks are equally plausible
- if $\theta_k < 0$, then network A^{k^-} is more plausible than network A

The other interpretation method is the one on the edge level. In order to make a connection between the vector of coefficient θ and the probability $\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)$, observe the following consideration:

Because of (2), one has

$$\text{logit}(\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)) = \theta^T \cdot (\Gamma(A_{ij}^+) - \Gamma(A_{ij}^-))$$

This is equivalent to the equation

$$\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c) = \text{logit}^{-1}(\theta^T \cdot (\Gamma(A_{ij}^+) - \Gamma(A_{ij}^-))) \tag{12}$$

Together with the abbreviation $(\Delta A)_{ij} := \Gamma(A_{ij}^+) - \Gamma(A_{ij}^-)$ and the inverse logit function

$$\text{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

one gets the equation

$$\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c) = \frac{\exp(\theta^T \cdot (\Delta A)_{ij})}{1 + \exp(\theta^T \cdot (\Delta A)_{ij})} \tag{13}$$

With this result, one can compute the odds of occurrence of edge (i, j), conditional on the rest of the network:

$$\begin{aligned}
\frac{\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = 0 | Y_{ij}^c = A_{ij}^c)} &= \frac{\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)}{1 - \mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)} \\
&= \frac{\frac{\exp(\theta^T \cdot (\Delta A)_{ij})}{1 + \exp(\theta^T \cdot (\Delta A)_{ij})}}{1 - \frac{\exp(\theta^T \cdot (\Delta A)_{ij})}{1 + \exp(\theta^T \cdot (\Delta A)_{ij})}} \\
&= \frac{\frac{\exp(\theta^T \cdot (\Delta A)_{ij})}{1 + \exp(\theta^T \cdot (\Delta A)_{ij})}}{\frac{1}{1 + \exp(\theta^T \cdot (\Delta A)_{ij})}} \\
&= \exp(\theta^T \cdot (\Delta A)_{ij})
\end{aligned}$$

With the equation

$$\theta^T \cdot (\Delta A)_{ij} = \sum_{l=1}^{q} \theta_l \cdot (\Delta_l A)_{ij}$$

one gets

$$\begin{aligned}
\frac{\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = 0 | Y_{ij}^c = A_{ij}^c)} &= \exp\left(\sum_{l=1}^{q} \theta_l \cdot (\Delta_l A)_{ij}\right) \\
&= \prod_{l=1}^{q} \exp(\theta_l \cdot (\Delta_l A)_{ij})
\end{aligned}$$

All in all, the calculation can be summarized as

$$\frac{\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij} = 0 | Y_{ij}^c = A_{ij}^c)} = \exp(\theta_1 (\Delta_1 A)_{ij}) \cdot \ldots \cdot \exp(\theta_q (\Delta_q A)_{ij}) \tag{14}$$

Equation (14) now enables a *ceteris-paribus analysis* of the parameters in the model: If the kth change statistic $(\Delta_k A)_{ij}$ increases one unit to $(\Delta_k A)_{ij} + 1$, while all the other change statistics remain unchanged, the odds of occurrence of edge (i, j), conditional on the rest of the network, is multiplied by the factor $\exp(\theta_k)$.
This leads to the following interpretation of the parameter θ_k, $k \in \{1, \ldots, q\}$:

- If $\theta_k > 0$, the conditional odds of occurrence increase.
- If $\theta_k = 0$, the conditional odds stay the same.
- If $\theta_k < 0$, the conditional odds decrease.

Therefore, the interpretation of the parameter happens almost the same way as it is done for logistic regression analysis (compare Fahrmeir et al. [15]). As will be discussed more extensively later on, one has to be very cautious with this kind of interpretation: The increase of the change statistic is not always reasonable. For instance, consider the network statistic

$$\Gamma_1(A) := \text{Number of edges with In-Degree} = 1$$

then the change statistic $(\Delta_1 A)_{ij}$ belonging to dyad (ij) can only attain values $\{-1, 0, 1\}$. As a result, the change statistic can never alter by more than one unit.

3.5 Statistics for the ERGM

Generally speaking, ERGM statistics can be differentiated into three groups: endogenous statistics, nodal covariates, and edge or relational covariates. Endogenous statistics capture the structural form of an observed network, while nodal covariates reflect actors' attributes. For instance, in the case of the international arms trade network this could be a nation's GDP or military expenditure. The third kind of covariate we are considering is the edge or relational covariate. As the name implies this kind of covariate captures other relations between actors in the network. Examples in our case could be covariates which describe whether two nations have a defense agreement or are in conflict with each other. Just as in the observed networks the relation between two actors can be either directed or undirected and therefore, be written as an adjacency matrix. In this paper we will refer to nodal and relational covariates as *exogenous covariates*. In this chapter we will introduce some endogenous network statistics, discuss how relational exogenous covariates can be incorporated into the ERGM, and explain how nodal covariates are included into the model.

The endogenous statistics we are going to discuss in this chapter are called *edges, outstar(2), instar(2), transitive, idegree(k), odegree(k), asymmetric, mutual, dsp(k)* and *esp(k)*, $k \in \mathbb{N}_0$. The statistic *edges* simply counts the number of edges in the network and plays the role of the intercept in the ERGM, since the change statistic in equation (2) is always going to be 1 for the number of edges. This results from the fact that the number of edges in network A_{ij}^{+} is exactly one higher than the number of edges in network A_{ij}^{-}. Therefore, the corresponding parameter θ_{edges} of the change statistic influences every network in the same way.

As already introduced in chapter 3.1, a network statistic $\Gamma_i(A)$ is a mapping from the set of all possible networks on N_V nodes $\mathcal{A}(N_V)$ into $\mathbb{R}$. Formally, this statistc can be written as

$$\Gamma_{edges} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij}$$

The *outstar(2)* statistic is called this way, because the edges radiating from the sender to several receivers form a star shape when drawn. In our case, a positive outstar(2)-parameter indicates that a country which is selling weapons to an other country is more likely to also sell weapons to a third country. The included statistic can be written as

$$\Gamma_{ostar(2)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} \sum_{k=1}^{N_V} a_{ij} a_{ik}$$

Analogously to $\Gamma_{ostar(2)}$ we define the *instar(2)* statistic as

$$\Gamma_{istar(2)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} \sum_{k=1}^{N_V} a_{ji} a_{ki}$$

The next endogenous statistic is called *transitive*. Networks with high values for the transitive statistic are those in which edges are more likely to exist between countries, which obtain weapons from a same third state. Seeing this statistic from a social scientific point of view, transitive incorporates the *a-friend's-friend-is-a-friend*-effect. Mathematically speaking, this statistic can be written as

$$\Gamma_{transitive} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} \sum_{k=1}^{N_V} a_{ij} a_{ik} a_{jk}$$

When looking at the international arms trade network (see figures (26) and (27) in the appendix) one recognizes that a noticeable structure of the network is that in

the case of an existing tie between two actors i and j this tie is usually not mutual. In other words, if nation i is selling weapons to nation j than there is usually no trade flow from j to i. The number of *asymmetric* or *one-sided* dyads in a network A can be defined as

$$\Gamma_{asymmetric} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij}(1 - a_{ji})$$

Analogously, the number of *mutual* dyads is defined as

$$\Gamma_{mutual} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij} a_{ji}$$

When looking at the arms trade network in its entirety, one observes that there is a high number of nodes which only receive weapons from one single country and do not distribute weapons themsevles to other countries. These countries have an in-degree of 1 and an out-degree of 0. However, we want to include statistics into the network, which do count the number of actors with an in- and out-degree of k. These statistics can be incorporated into the model by including the network statistics *idegree(k)* and *odegree(k)*, where $k \in \mathbb{N}_0$. Formally these statistics can be written as

$$\Gamma_{idegree(k)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{m=k}^{N_V - 1} \left\{ \binom{m}{k} \sum_{j=1}^{N_V} \left[\mathbb{1}_{\{m\}} \left(\sum_{i=1}^{N_V} a_{ij} \right) \right] \right\}$$

$$\Gamma_{odegree(k)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{m=k}^{N_V - 1} \left\{ \binom{m}{k} \sum_{j=1}^{N_V} \left[\mathbb{1}_{\{m\}} \left(\sum_{i=1}^{N_V} a_{ji} \right) \right] \right\}$$

Other very useful statistics that are going to play a central role in modeling the arms trade network are the shared-partner statistics *dyad-wise k-shared partners* (dsp(k)) and *edge-wise k-shared partners* (esp(k)). Since dsp(k) is a generalization of esp(k), the following paragraph will primarily discuss the dyad-wise shared partner statistic. This statistic counts the number of vertex pairs (i, j), which share exactly k common neighbors. In a directed graph only vertexes connecting (i, j) over a path of length 2 are counted. To get a better idea of this statistic, take a look at figure (7), where one dsp(3) statistic is visualized. The vertexes A and B share exactly 3 neighbors and are connected over these neighbors by a directed path of length 2. The difference between $esp(k)$ and $dsp(k)$ is that for the $esp(k)$ vertexes A and B would have to be connected by an edge. This is not a necessary requirement for the $dsp(k)$. Therefore,

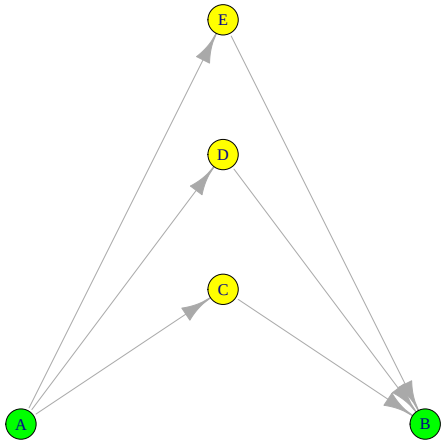


Figure 7: Visualization of dyad-wise shared partners.

figure 7 would demonstrate an esp(3) statistic if A and B were be connected by an edge. Formally, dsp(k) and esp(k) can be written as

$$\Gamma_{dsp(k)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} \left[\mathbb{1}_k \left(\sum_{m=1}^{N_V} a_{im} a_{mj} \right) \right]$$

$$\Gamma_{esp(k)} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} \left[\mathbb{1}_k \left(\sum_{m=1}^{N_V} a_{im} a_{mj} a_{ij} \right) \right]$$

After having introduced all endogenous statistics, which are going to play a role in modeling the arms trade network, we will now turn our focus on the implementation of exogenous data. The way relational covariates are included into the network is quite simple. Since these data can easily be written in the same structure and dimension of the underlying adjacency matrix A they can be included into the network as

$$\Gamma_{rel} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij} c_{ij}$$

while c_{ij} indicates the corresponding entry of the considered relational covariate matrix. For every existing edge, this statistic adds up the associated entries of the covariate matrix.

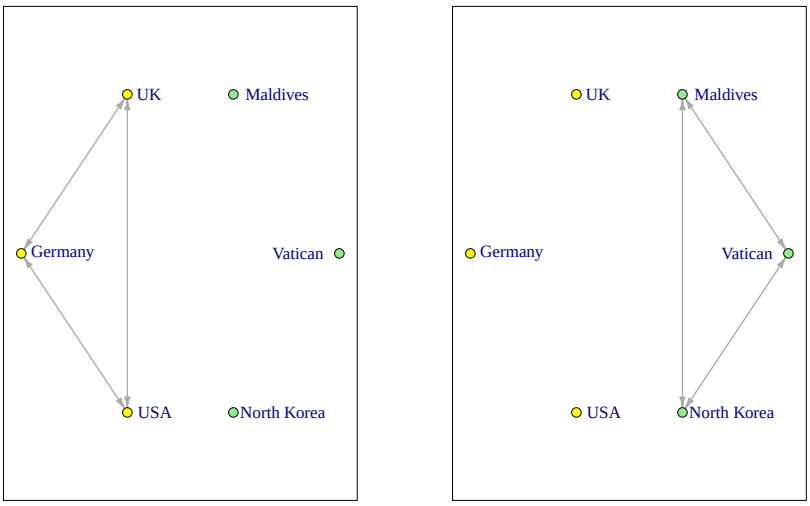


Figure 8: Structural equivalent networks

At this point it should be mentioned that by including covariates one forestalls the following structural problem resulting from only including endogenous statistics: According to (1) the probability distribution $\mathbb{P}_\theta(Y = A)$ only depends via $\Gamma(A)$ on the specific realization of A. This means that two networks A_1 and A_2, which are structurally equivalent on the included endogenous statistics, meaning $\Gamma(A_1) = \Gamma(A_2)$, are equiprobable. As a consequence, by only including endogenous statistics, the model does not distinguish between the nodes, since it only refers to the structure of these networks. In order to visualize this problem one can take a look at figure 8, where two different networks with the same structure are plotted on the same six actors. By only including endogenous statistics both networks are equiprobable, since the structure of both networks is equivalent. Of course, the left network should appear to be more plausible from a contextual point of view than the right one. However, since endogeneous statistics only incorporate structural characteristics of the network the model does not consider node specific attributes. For this reason, it is absolutely essential to include exogenous variables into the network. For example, by introducing the defense agreement covariate into our example, $\Gamma_{rel=defense}$ would count the number of matches between the ties in the observed network and the ties in the defense agreement network. If the model is estimating a positive parameter $\theta_{defense}$, then the model with more accordances with the defense agreement network

turns out to be more likely than the other one. Consequently, a distinction between the two different, but structurally equivalent networks can be made.

Finally, how can nodal covariates be incorporated into the network? When including nodal covariates into an ERGM, the ERGM is expanding the vector of nodal attributes into a matrix. In a directed network we furthermore have to distinguish between sender and receiver effects, i.e., whether the nodal covariate has an effect on the buying or the selling behavior of a country. Take for instance the G8 network from figure 1 in chapter 1. For the nations in this network it could be reasonable to take the number of allied countries in this network into account. When including a nodal covariate as a sender effect, the ERGM is transforming the vector of the number of the actors' allies

$$(5, 5, 5, 5, 1, 0, 5, 6)'$$

into a matrix of the form

$$SM = \begin{pmatrix} & Ca & Fr & Ger & It & Jap & Rus & UK & USA \\ Canada & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ France & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ Germany & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ Italy & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ Japan & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ Russia & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ UK & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ USA & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \end{pmatrix}$$

A nodal covariate is turned into a matrix with the same dimensions as the observed adjacency matrix A and is than included into the ERGM with the statistic

$$\Gamma_{SM} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij} sm_{ij}$$

where $SM = sm_{ij} \in \mathbb{R}^{N_V \times N_V}, i, j, \in \{1, \ldots, N_V\}$ (see Hunter et al [26]). The statistic Γ_{SM} adds for every existing edge in the network the number of the supplier's allies. This statistic is implemented in R as *nodeocov*.

Including a nodal covariate as a receiver effect can be done in a very similar way. Instead of expanding the vector of the number of allies by row into a matrix the size of A, we expand it by column. As a result the expanded receiver matrix RM can

be included into the model with the statistic

$$\Gamma_{RM} : \mathcal{A}(N_V) \to \mathbb{R} \quad , \quad A \mapsto \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} a_{ij} rm_{ij}$$

where $RM = rm_{ij} \in \mathbb{R}^{N_V \times N_V}, i, j \in \{1, \ldots, N_V\}$. After having discussed the most common statistics for the ERGM, we are set to fit our first network model.

3.6 First ERGM for the Arms Trade Network

When fitting an ERGM one usually has to deal with so called *degeneracy* problems, which result in unreliable approximative likelihood estimates for the model's parameters. The reason why degeneracy occurs is that the stochastic process generated by the MCMC-algorithm does not necessarily hold the through the model defined distribution of the random variable Y as stationary distribution (see Handcock [19] for further information). Unfortunately, the models we were fitting with commonly implemented endogenous statistics were either generating degenerated results or producing poor model fits, since the included statistics did not capture the structural form of the networks sufficiently. The best non-degenerated model, with endogenous statistics, which was reasonable according to the structure of the network, was

$$\Gamma(A) = (\Gamma_{edges}, \Gamma_{asymmetric}, \Gamma_{idegree(1)}, \Gamma_{dsp(1)})$$

Almost every ERGM of interest includes the Γ_{edges} statistic for the same reason that nearly every linear regression model contains an intercept term. $\Gamma_{asymmetric}$ makes sense, since the vast majority of the non-null dyads are one-sided. With $\Gamma_{idegree(1)}$ we are trying to capture the fact that the arms trade network includes a lot of satellite countries, i.e., countries which only purchase their weapons from a single supplier. Extending the model with $\Gamma_{odegree(k)}$ as well as $\Gamma_{ostar(k)}$ and $\Gamma_{istar(k)}$ caused degeneracy. With the statistic *dsp(1)* we intended to capture the structural characteristic that the arms trade network has a few central weapon distributors (see Jansen and Schmid [29] for a more precise discussion).

Besides the endogenous statistics introduced in the previous paragraph we are going to include a range of exogenous covariates. A general explanation about the included covariates can be found in chapter 2. The covariates *defense agreement, direct contiguity* and *polity score* are captured as edge covariates, as well as a variable we are going to refer to as *path dependency*. This covariate sums up the total TIV sold from country i to country j the five years before the year of consideration. Further

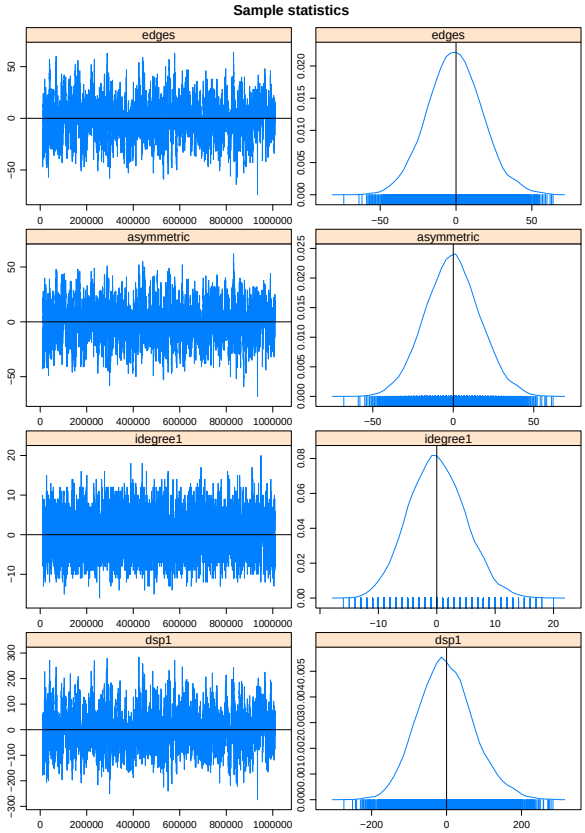


Figure 9: MCMC diagnostics for the ERGM of 2013 with endogenous statistics Γ_{edges}, $\Gamma_{asymmetric}$, $\Gamma_{idegree(1)}$ and $\Gamma_{dsp(1)}$

on, we include the nodal covariates *GDP, CINC* and *intra-state conflict* into the network. The covariates GDP and CINC are incorporated for the supplier as well as for the recipient, while the intra-state conflict data are only added for the recipient. Moreover, we decided to exclude the inter-state conflict and the embargo data from our models since these networks turned out to be extremely sparse, and as a consequence, generated highly oscillating parameter estimates with enormous variance values. For certain years the included covariate networks were empty networks, and therefore caused degenerated model fits.

Perkins and Neumayer [40] argue that there is a time delay between the order date of arms and the delivery date, which, according to our calculations, turns out to be an average of two years. Therefore, all exogenous covariates are included with a two year lag, i.e., for the network of year t we use the exogenous covariates of year $t-2$.

In order to verify whether a model is degenerated or not, one can take a look at the MCMC diagnostics as plotted in figure 9. For simplicity's sake, only the MCMC diagnostics of the endogenous statistics are shown. The plots on the left side visualize the attained values via MCMC simulated networks for every single statistic included into the model. Doing so centers the attained values around the values of the observed network. We refer to this kind of visualization as a *trace plot*. The plots on the right side visualize the empirical density function of the respective statistic, based on the simulated networks (see Hunter and Handcock [27]).

After having understood the meaning of MCMC diagnostics plots, the next logical question concerns what good MCMC diagnostics look like. The empirical density function should be symmetrical around zero for every included centered statistic Γ_*, since the expected value of the centered statistic

$$\Gamma_*(\cdot) - \Gamma(A^{obs})$$

should be zero. Otherwise, the values in the simulated networks systematically differ from the corresponding statistics in the observed network, making it unreasonable to assume that the simulated networks originate from the same distribution as the observed network. Furthermore, the trajectories in the trace plot should neither indicate a dependence structure nor remain on a constant level. This would be a signal that the constructed stochastic process violates the Markov properties. When looking at the MCMC diagnostics in figure 9 we observe that a dependence structure is not identifiable and that the empirical density functions are symmetrical around zero. Thus, the model is not degenerated.

The MCMC diagnostics of a clearly degenerated model can be seen in figure 10. This model was fitted with the endogenous statistics Γ_{edges} and $\Gamma_{odegree(0)}$. In this case we can observe obvious dependency structures in the trace plots as well as an empirical density function which is not symmetrical around zero. Both are strong indicators for the degeneracy of the model.

The first model did not degenerate, but does it also provide a good model fit? In order to answer this question, we are going to compare the fitted models using four hyper-statistics: The in-degree distribution, the out-degree distribution, the geodesic distance distribution between two actors, and the edgewise-shared partner distribution. To be able to judge whether a model fit is good, one can take a look at the *goodness-of-fit* plots as seen in figure 11. After having estimated the vector of parameters $\hat{\theta}$, one is interested in how similar the distribution of $\mathbb{P}_{\hat{\theta}}$ is to

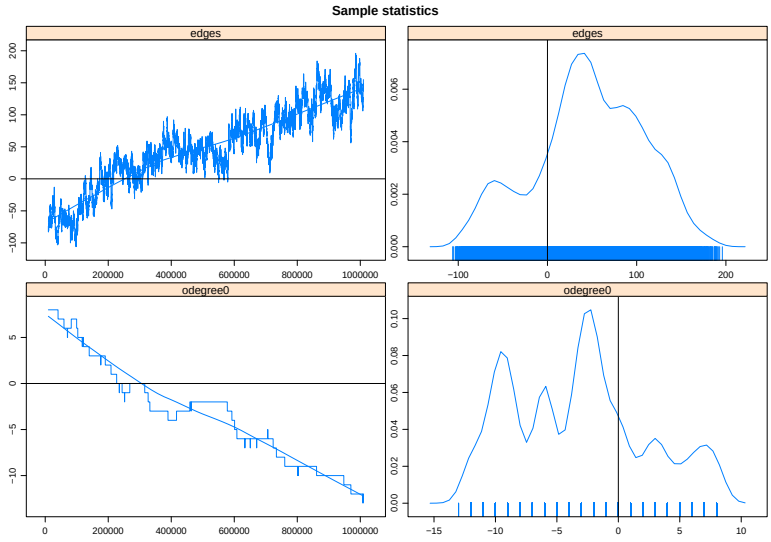


Figure 10: MCMC diagnostics for the ERGM of 2013 with endogenous statistics edges and odegree(0)

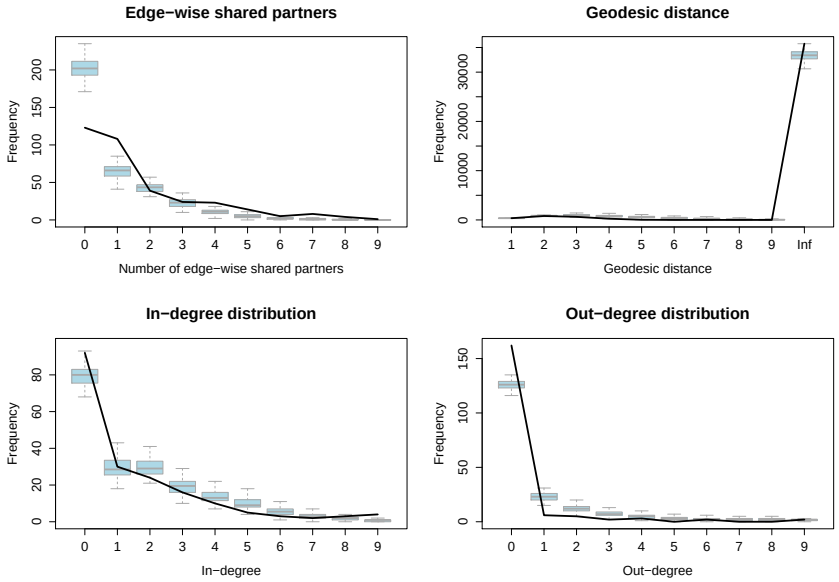


Figure 11: Goodness-of-fit plots for the ERGM for 2013

the distribution of $\mathbb{P}_\theta$. In order to answer this question, we are simulating a large number of networks out of the distribution $\mathbb{P}_{\hat{\theta}}$ via MCMC as described in chapter 3.3 and comparing the simulated networks based on the distributions of the hyper statistics with the originally observed network. The bold black line illustrates the hyper statistic distribution of the observed network, while the range bounded by the boxplots displays the range where 95% of the simulated networks' hyper statistics can be found. According to this, a model provides a good fit if the bold black line passes through every single boxplot, and even better if it hits the median of each boxplot (see Hunter et al. [26]). Therefore, one can say that, with the exception of the in-degree distribution, none of the three remaining hyper statistic distributions are well captured in this model. As a consequence, the distribution of $\mathbb{P}_{\hat{\theta}}$ is not similar to the distribution of $\mathbb{P}_\theta$.

A reason for the poor model fit could be that the change statistic $(\Delta A)_{ij}$ increases linearly, a fact that can cause instability and hence, result in degenerated models. Kauermann[1] suggests circumventing this instability problem by replacing the parameter vector θ with smooth functions. In doing so, we ccan rewrite equation (2) as

$$\text{logit}(\mathbb{P}_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)) = s((\Delta A)_{ij})$$

where $s((\Delta A)_{ij}) := s_1((\Delta_1 A)_{ij}) + \cdots + s_p((\Delta_p A)_{ij})$ and $s_k(\cdot), k \in \{1, \ldots, p\}$ are smooth functions, which have to be estimated from the data. We are going to discuss the estimation of smooth functions in the next chapter. This model generalizes the ERGM similar to how the GAM generalizes the GLM. Consequently, just as we will discuss in chapter 5.3 for the GAM, we need some additional identifiability constraints for the smooth functions $s_k(\cdot)$. However, since this model is not yet implemented in R we are going to introduce a different approach to improving our model.

Something that catches one's eye in figure 11 is that the number of actors with an in-degree of 1 is perfectly captured in the model. Recall that this model was fit with $\Gamma_{idegree(1)}$. It seems like this single statistic is enough to capture the entire in-degree distribution in a satisfying way. By including statistics into the network which adequately reflect the in- and out-degree distribution there is hope that the distribution of the hyper statistics geodesic distance and edge-wise 1-shared partners of $\mathbb{P}_{\hat{\theta}}$ might improve. Unfortunately, including statistics with an in-degree of k, where $k \in \mathbb{N}\backslash\{1\}$, or any statistic which counts the number of actors with an out-degree of ℓ, $\ell \in \mathbb{N}$,

[1] This paper has not been published at the time of this study (05/2015).

caused degenerated models. However, in order to fit a reasonable network model one has to include endogenous statistics, which capture the structure of the model in a suitable way.

Furthermore, in order to find a statistic that is able to incorporate a network's entire degree distribution, one should consider that the effects of an actor's in- and out-degrees might not be linear in nature. The difference between the chance that an actor with an in-degree of 1 is forming a new tie instead of an actor with an in-degree of 0 might be higher than the difference between the chances of two actors with an in-degree of 10 and 11. This assumption would result in an effect that flattens more the higher the in- and out-degrees of an actor are.

But what does this non-linear effect look like and how can one detect this relation? To explore this question, we are going to establish an approach to model the network with *generalized additive models* (GAM), which can incorporate non-parametric effects of covariates with so-called *smooth functions*. Even though, the results will be biased, since these models ignore a network's dependency structure, we will obtain an approximate impression of a node's in- and out-degree effects. The idea is to detect the degree distribution's functional effects and to adjust the ERGM fit by adapting *geometrically weighted statistics* to this relation. These statistics intend to use degree counts with geometrically decreasing weights. We will introduce these statistics in chapter 7. However, first we are interested in the degree distribution's non-parametric effect. Therefore, we are going to fit a GAM, which presupposes smoothing techniques. As a consequence, we are going to discuss some basic smoothing theory in the next chapter.

4 Scatterplot Smoothing

In this chapter we will introduce some techniques for editing nonparametric functions. As was already assumed in the previous chapter, the relation between response and covariates does not seem to be linear in every case. Therefore, we introduce smoothing splines, which create approximate functions to capture important patterns in the data. The most important property of smooth functions are their nonparametric nature, and as a consequence, we do not assume a rigid form of dependence between the response Y_{ij} and the influential variables $X_{ij1}, \ldots, X_{ijp}$. However, the name *non parametric* is not always well chosen. Even though there are several smoothing techniques, like kernel smoothers, where the term nonparametric applies, many other techniques such as spline smoothers, which will be discussed in

this paper and later used for computation, are described by parameters. However, these parameters only regulate the adjustment of splines to data and therefore cannot be interpreted in a scientific way.

But how can one detect the most appropriate smooth function for a covariate's effect? As a first step, we focus on how the effect of a single metric covariate on an approximately Gaussian distributed target value can be estimated. These results then serve as the basis for smoothing methods for several non-linear metric covariates. This chapter is mostly based on Hastie and Tibshirani [23], Wood [55], Fahrmeir et al [15] and Eilers and Marx [14].

4.1 Polynomial Splines

For the following, we assume given data in the form (y_{ij}, x_{ij}) , $i, j = \{1, \dots, N_V\}, i \neq j$, where y_{ij} are observations of the dependent response variable and x_{ij} are the corresponding metric covariates. Given that assumption, if we take y_{ij} as dyads in a network on N_V nodes we obtain $N = N_V^2 - N_V$ observations. We assume that the response variable can be described by a function $s(\cdot)$ and a measuring error ε_{ij}

$$y_{ij} = s(x_{ij}) + \varepsilon_{ij} \tag{15}$$

The first approach that probably comes to mind is to approximate the relation between the target value and the covariate with a polynomial function

$$s(x_{ij}) = \alpha_0 + \alpha_1 x_{ij} + \cdots + \alpha_b x_{ij}^b$$

where $b \in \mathbb{N}$ and $\alpha_k \in \mathbb{R}$, $k \in \{0, \dots, b\}$. This idea could be, for instance, realized by the least square method. However, in most cases a pure polynomial approach does not provide satisfying results. In order to understand this, take a look at figure 12. Here data was simulated with a nonpolynomial function

$$f : [-4, 4] \to \mathbb{R}$$

$$f(x) = 2(-0.4exp(-0.5(x+1)^2) - 0.6exp(-0.5(x-2)^2)) + 0.9 \tag{16}$$

and $y = f(x) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.2)$. A similar example is used in Fahrmeir [15]. The first plot shows the simulated data together with function $f(\cdot)$. When looking at the second picture one can see that assuming a linear relation between x and y is not the best choice. The linear function does not only disregard the local minima and maximum, it also neglects the slope at the domain boundary. But the linear

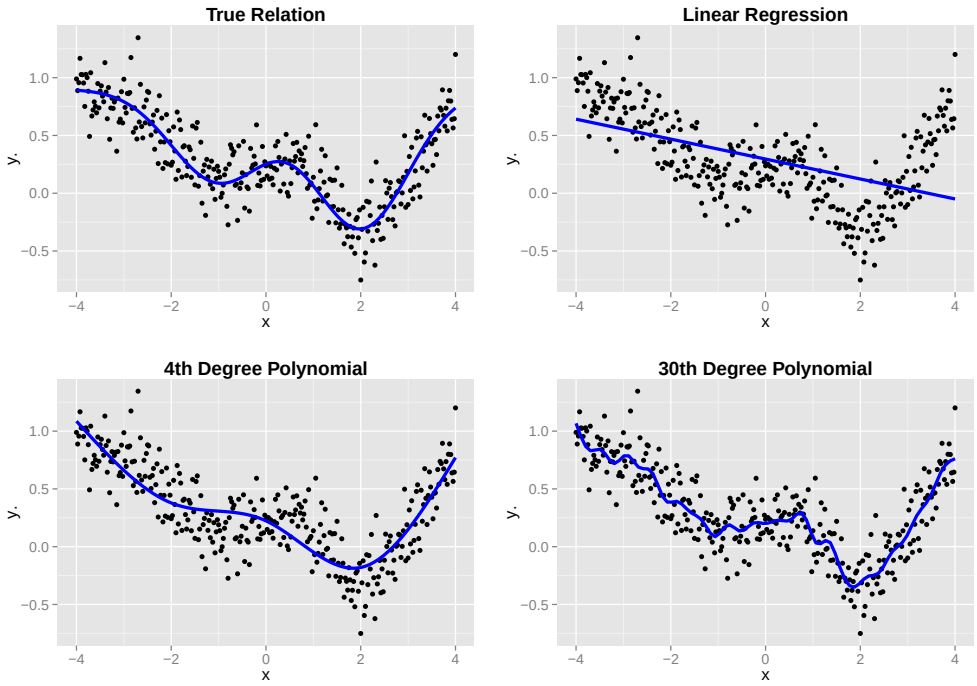


Figure 12: Polynomial regression for the simulated data

regression is not the only one that results in a bad fit. Even a polynomial approach, applied in the third and forth pictures, visualizes the problems with pure polynomial methods. While polynomials with low degrees do not capture the true relation of the data sufficiently (for instance, the local maximum at $x \approx 0$ and the local minima at $x \approx -1$ and $x \approx 2$ are not captured adequately), polynomials with high degrees provide *wiggly* fits of the data. In this case we talk about *overfitting*, which is not ideal either.

In order to find a way out of this quandary one could divide the codomain into m parts $c = \kappa_0 < \cdots < \kappa_m = d$ and capture the relation between x and y on each interval $[\kappa_l, \kappa_{l+1})$, $l \in \{0, \ldots, m-1\}$ with a b-th degree polynomial. The problem with this approach is that, since the estimates are done independently for each interval, the piecewise estimated functions are not necessarily connected. A method for how one can gain functions which are estimated on intervals $[\kappa_l, \kappa_{l+1})$ but still provide continuous transitions will be given in the next chapter.

4.2 B-Splines

There are several ways to adopt measures to forestall non-connected estimates. The two most common ones are polynomial spines with truncated powers and so-called *B-splines*. Since B-splines are usually chosen over polynomial splines with truncated powers for numerical reasons, we content ourself discussing the B-spline method. We mainly refer to Gu [18] or Ruppert et al. [44].

The problem resulting from the previous paragraph is that piecewise estimated polynomials usually provide smooth functions, which are neither continuous nor differentiable on the entire codomain. The main idea of B-splines is a construction to guarantee that piecewise estimated functions on knots $\kappa_1, \ldots, \kappa_{m-1}$ are composed in a sufficient, $(b-1)$-times differentiable way. In order to estimate $s(\cdot)$ with B-splines, one has to represent the smooth function in such a way that $y = s(x) + \varepsilon$ becomes a linear model. This is done by choosing specific *basis functions* $B_1(x), \ldots, B_t(x) : [c, d] \to \mathbb{R}^+$, $t = m + b - 1$. Then, one can write

$$s(x) = \sum_{i=1}^{t} \alpha_i B_i(x) \tag{17}$$

A huge advantage the B-spline method has over other spline approaches is that B-splines are defined as non-zero functions on only a few intervals $[\kappa_l, \kappa_p]$, $l, p \in \{0, \ldots, m\}$, $l \neq p$. This results in numerical benefits, as we will discover later. Let

$$B_i(x) = \begin{cases} f(x) & \text{, if } x \in [\kappa_i, \kappa_{i+b+1}) \\ 0 & \text{, else} \end{cases}$$

where $f : \mathbb{R} \to \mathbb{R}^+$ is constructed from polynomial pieces and

$$\sum_{i=1}^{t} B_i(x) = 1$$

More exactly, the function $f(\cdot)$ is composed of $b+1$ polynomial pieces of degree b, which are put together in a $b-1$-times continuously differentiable way. On the left side of figure 13 we can see a single B-spline basis function of degree 1. B-splines are defined by covering each interval $[\kappa_l, \kappa_{l+b+1})$ by $b+1$ basis functions of degree b. This results in $(b-1)$-times differentiable functions. For better understanding, the simple example of linear B-splines over equidistant knots is given in figure 13. When looking at the interval $[\kappa_6, \kappa_9]$, with $m = 3$, one can see that each interval is covered by $b + 1 = 2$ linear basis splines and that a total of $t = m + b - 1 = 4$ basis

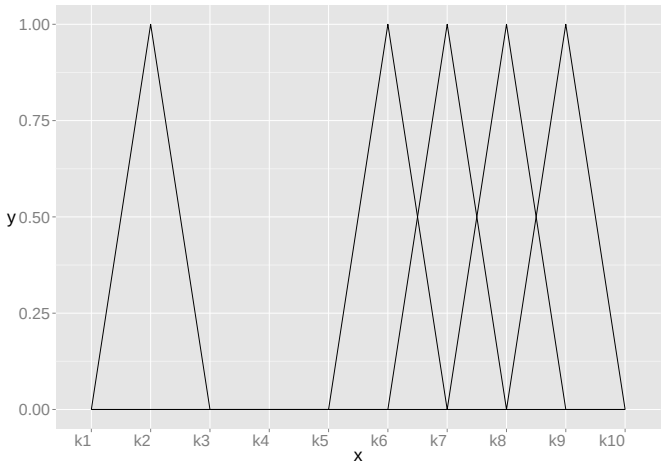


Figure 13: Illustrations of linear B-splines on equidistant knots

functions are needed to cover interval $[\kappa_6, \ldots, \kappa_9]$.

By looking at the basis functions in figure 13 we can easily verify the actual definition of linear B-spline basis functions

$$B_i^1(x) = \frac{x - \kappa_i}{\kappa_{i+1} - \kappa_j} \mathbb{1}_{[\kappa_i, \kappa_{i+1}]}(x) + \frac{\kappa_{i+2} - x}{\kappa_{i+2} - \kappa_{i+1}} \mathbb{1}_{[\kappa_{i+1}, \kappa_{i+2}]}(x)$$

where the 1 in $B_i^1(x)$ points out the linear form of the piecewise defined polynomials. This definition obviously alludes to the fact that $B_i^1(x)$ consists of two linear pieces. In general, B-spline basis functions for higher degrees can be defined recursively

$$B_i^b(x) = \frac{x - \kappa_i}{\kappa_{i+1} - \kappa_j} B_i^{b-1}(x) + \frac{\kappa_{i+b+2} - x}{\kappa_{i+b+2} - \kappa_{i+1}} B_{i+1}^{b-1}(x)$$

Due to the linear form of (17) and by defining X and α as

$$X = \begin{pmatrix} B_1(x_{12}) & \ldots & B_t(x_{12}) \\ \vdots & & \vdots \\ B_1(x_{(N_V-1)N_V}) & \ldots & B_t(x_{(N_V-1)N_V}) \end{pmatrix} \quad , \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_t \end{pmatrix} \tag{18}$$

one can write (15) in linear form

$$y = X\alpha + \varepsilon \tag{19}$$

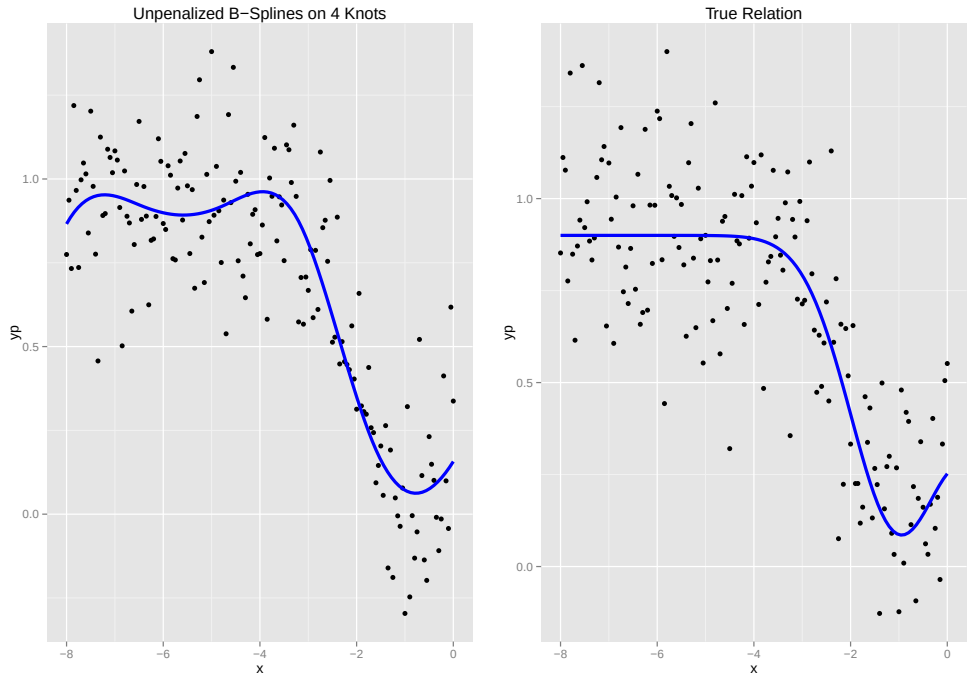


Figure 14: B-Spline regression for the simulated data

where $y = (y_{12}, \dots y_{(N_V-1)N_V})'$ and $\varepsilon = (\varepsilon_{12}, \dots, \varepsilon_{(N_V-1)N_V})'$. As a consequence, the parameter vector α can be estimated by the ordinary least square method

$$\hat{\alpha} = (X'X)^{-1}X'y \tag{20}$$

As already mentioned above, the design matrix X holds some beneficial characteristics. The most important characteristic stems from the local definition of the basis functions, which mainly yield matrix entries of 0. The only non-zero entries occur in a tube along the diagonal of the matrix. These kinds of band matrices are desirable, since solving (20) with these matrices is numerically more efficient.

However, the parameter vector α cannot be interpreted in a reasonable way. Instead, one is interested in the form of the estimated function $\hat{s}(\cdot)$, which is a result of $\hat{\alpha}$:

$$\hat{s}(x) = B\hat{\alpha}$$

where $B = (B_1(x), \dots, B_t(x))$.

As we can see on the left side of figure 14, spline estimators with basis functions can

be wiggly. Both plots were generated with the same function from example (16), but on a different codomain. The left side shows a B_i^3 basis spline estimate on 4 equidistant knots, while the plot on the right side presents the actual relation in the data. There are several reasons why basis splines turn out wiggly, such as the selection of the basis dimension or the selection of the knots. It is reasonable that a smooth, but not too wiggly function would be preferred over a spline estimator as seen on the left side of figure 14. But how can one control the 'wiggliness' of a smoother? A common strategy is by controlling the degree of smoothing by *penalized* B-splines.

4.3 Penalized B-Splines (P-Splines)

Penalized B-splines differ from the methods discussed in the previous chapter, since instead of minimizing

$$||y - B\alpha||^2$$

we are going to minimize

$$||y - B\alpha||^2 + \lambda \int_C s''(x)^2 dx \tag{21}$$

with regard to α, where C is the codomain of x and $s''(x)$ is the second derivative of function $s(x)$. Forming a penalty function by the second derivative of a fitted curve was first introduced by O'Sullivan [39]. The second derivative of a function yields information about a functions curvature, and therefore by minimizing (21) we penalize models that are too wiggly. With the *smoothing parameter* λ one can control the trade-off between the model's fit and smoothness. While $\lambda = 0$ results in spline estimates without penalization and hence in wiggly models, $\lambda \to \infty$ leads to the linear regression of the data. In the next chapter, we discuss a method for finding a fitting smoothing parameter λ, but for now we treat λ as given.

As a first step, we are going to show that we can write the penalty in (21) as

$$\int_C s''(x)^2 dx = \alpha^T S \alpha \tag{22}$$

where $S \in \mathbb{R}^{t \times t}$ is a matrix that can be expressed by the basis functions $B_i(x)$. The proof is fairly straightforward: Recall that we define function $s(x)$ as

$$s(x) = \sum_{i=1}^{t} \alpha_i B_i(x)$$

which yields

$$s''(x) = \alpha^T B''(x)$$

for the second derivative. The second derivative exists thanks to the polynomial nature of the piecewise composed basis functions. Since $s''(x)$ is a scalar and scalars are their own transpose we can write

$$\begin{aligned} \int_C s''(x)^2 dx &= \int \alpha^T B_i''(x) B_i''(x)^T \alpha \; dx \\ &= \alpha^T \underbrace{\int_C B_i''(x) B_i''(x)^T dx}_{:=S} \; \alpha \end{aligned}$$

This already finishes the proof. As a consequence, instead of minimizing (21) one can minimize

$$||y - B\alpha||^2 + \lambda \alpha' S \alpha$$

with regard to α. Minimizing this equation with the least square method yields

$$\begin{aligned} LS(\alpha) &= (y - B\alpha)'(y - B\alpha) + \lambda \alpha' S \alpha \\ &= y'y - 2y'B\alpha + \alpha' B' B \alpha + \lambda \alpha' S \alpha \end{aligned}$$

Here we have used $\alpha' B' y$ and $y' B \alpha$ as scalars and therefore $(\alpha' B' y)' = y' B \alpha$. Together with the rules of derivation

$$\frac{\partial w'v}{\partial v} = w \quad \text{and} \quad \frac{\partial v' A v}{\partial v} = 2Av$$

where $v, w \in \mathbb{R}^t$ are vectors and $A \in \mathbb{R}^{t \times t}$ is a symmetrical matrix, one gets for the first and second derivation

$$\frac{\partial LS(\alpha)}{\partial \alpha} = -2B'y + 2B'B\alpha + 2\lambda S\alpha \tag{23}$$

$$\frac{\partial^2 LS(\alpha)}{\partial \alpha \partial \alpha'} = 2B'B + 2\lambda S \tag{24}$$

$B'B + \lambda S$ is positive definite and therefore invertible (see Fahrmeir [15] for further details). As a result, we get a solution to our minimization problem by zeroing (23). Solving this equation for α finally yields the least square estimator for α

$$\hat{\alpha} = (B'B + \lambda S)^{-1} B'y \tag{25}$$

Even though this approach is straightforward, one has to compute the second derivative

$$s''(x) = \alpha^T B''(x)$$

Let $B_i^b(x)$ be the value of x of the ith B-spline of degree b. de Boor [3] introduces a simple formula for the derivatives of B-splines and shows that

$$\begin{aligned} s'(x) &= -\sum_i^t \Delta\alpha_{i+1} B_i^{b-1}(x) \\ s''(x) &= \sum_i^t \Delta^2\alpha_i B_i^{b-2}(x) \end{aligned}$$

where $\Delta\alpha_i = \alpha_i - \alpha_{i-1}$ and $\Delta^2\alpha_i = \Delta\Delta\alpha_i = \alpha_i - 2\alpha_{i-1} + \alpha_{i-2}$. However, these derivatives lead to rather complex systems of equations. Therefore, Eilers and Marx [14] suggest a simple approximation of the derivatives, which can be used for the construction of the penalty terms. Instead of (21) we are going to minimize

$$||y - B\alpha||^2 + \lambda \sum_{i=3}^{t} (\Delta^2\alpha_i)^2 \tag{26}$$

Besides easy computation, this approach has the advantage of being able to penalize linear B-splines in a reasonable way, since the second derivative is not constantly zero. The spline functions estimated in chapter 6.3 are going to apply this approximation.
After having discussed how to estimate α it remains to be seen how one can adequately establish an appropriate smoothing parameter λ.

4.4 Cross Validation

Selecting an appropriate smoothing parameter λ is crucial for a good model fit. If λ is too small $\hat{s}_\lambda(\cdot)$ will be too wiggly and if λ is too large, the data will be oversmoothed. In either case, the spline estimate $\hat{s}_\lambda(\cdot)$ is not close to the true function $s(\cdot)$ and, as a consequence, is a bad fit. In the ideal case one would select λ in a way that $\hat{s}_\lambda(\cdot)$ is as close as possible to $s(\cdot)$. Hence, an appropriate criterion could be to choose λ in order to minimize

$$W := \frac{1}{N} \sum_{i=1}^{N_V} \sum_{\substack{j=1 \\ i \neq j}}^{N_V} \Big(\hat{s}_\lambda(x_{ij}) - s(x_{ij}) \Big)^2 \tag{27}$$

However, minimizing (27) in a direct way is not possible, since $s(\cdot)$ is unknown, but one can derive an estimate for the expected squared error $\mathbb{E}(W) + \sigma^2$ by using *cross validation*.

Cross validation is a statistical method that involves the partitioning of a sample of data into two subsets: a training set for model fitting and a validation set for the evaluation of the model. The main idea of cross validation is to reuse the data by switching the roles of the training and validation samples. However, this method is not sufficient for small data samples.

The cross validation method we are going to introduce is the *leave-one-out* method. According to the name of this method one can easily imagine that it works by leaving the points $(y_{ij}, x_{ij}), ij = \{12, \ldots, N_V(N_V - 1)\}$ out one at a time as the validation set and estimating the smooth function with the remaining $N - 1$ points. By doing so, the omitted data becomes independent of the model fit. Therefore, one can construct the (ordinary) cross validation sum of squares

$$CV(\lambda) = \frac{1}{N} \sum_{i=1}^{N_V} \sum_{\substack{j=1 \\ i \neq j}}^{N_V} \Big(y_{ij} - \hat{s}_\lambda^{-ij}(x_{ij}) \Big)^2 \tag{28}$$

where $\hat{s}_\lambda^{-ij}(x_{ij})$ stands for the model fitted to all data except the observation (x_{ij}, y_{ij}). A cross validation estimate of λ is the minimizer of (28). $CV(\lambda)$ is computed by leaving out each observation one at a time, estimating the model on the remaining data, computing the squared difference of $\hat{s}_\lambda^{-ij}(x_{ij})$ and y_{ij} and by averaging them over all the data.

As a next step recall that $y_{ij} = s(x_{ij}) + \varepsilon_{ij}$. Then, equation (28) can be written as

$$\begin{aligned} CV(\lambda) &= \frac{1}{N} \sum_{i=1}^{N_V} \sum_{\substack{j=1 \\ i \neq j}}^{N_V} \Big(s(x_{ij}) - \hat{s}_\lambda^{-ij}(x_{ij}) + \varepsilon_{ij} \Big)^2 \\ &= \frac{1}{N} \sum_{i=1}^{N_V} \sum_{\substack{j=1 \\ i \neq j}}^{N_V} \left(\Big(s(x_{ij}) - \hat{s}_\lambda^{-ij}(x_{ij}) \Big)^2 - 2 \Big(s(x_{ij}) - \hat{s}_\lambda^{-ij}(x_{ij}) \Big) \varepsilon_{ij} + \varepsilon_{ij}^2 \right) \end{aligned}$$

Let us assume that $\hat{s}_\lambda^{-ij}(x_{ij})$ and ε_{ij} are independent. Therefore, together with $\mathbb{E}(\varepsilon_{ij}) = 0$ one can vanish the second term by taking expectations

$$\mathbb{E}\big(CV(\lambda)\big)= \frac{1}{N}\,\mathbb{E}\left(\sum_{i=1}^{N_V}\sum_{\substack{j=1\\ i\neq j}}^{N_V}\Big(s(x_{ij}) - \hat{s}_\lambda^{-ij}(x_{ij})\Big)^2\right)+\sigma^2 \tag{29}$$

By using the leave-one-out method for large sample sets one gets $\hat{s}_\lambda(x_{ij}) \approx \hat{s}_\lambda^{-ij}(x_{ij})$ and, as a consequence, $\mathbb{E}\big(CV(\lambda)\big)\approx \mathbb{E}(W) + \sigma^2$. This means we have found a way to approximate (27) and consequently, in order to minimize W, we can choose λ in order to minimize $CV(\lambda)$.

We have found a reasonable approach for estimating a fitting smoothing parameter λ, but as one can easily imagine it is quite inefficient to compute CV by fitting the model to each of the N resulting data sets where one observation is left out one at a time. Fortunately, there exists a short cut, since one can show that

$$CV(\lambda) = \frac{1}{N}\sum_{i=1}^{N_V}\sum_{\substack{j=1\\ i\neq j}}^{N_V}\left(\frac{y_{ij} - \hat{s}_\lambda(x_{ij})}{1 - a_{ijij}}\right)^2 \tag{30}$$

where a_{ijij} are the diagonal elements of the influence matrix A.
For the proof, let, for reasons of clarity, $p := ij$ and $q := kl \in \{12, \ldots, N_V(N_V - 1)\}$. Then, we define

$$\tilde{y}_{pq} = \begin{cases} y_q & \text{if } p \neq q \\ \hat{s}_\lambda^{-p}(x_p) & \text{if } p = q \end{cases}$$

which yields $\tilde{y}_p = (\tilde{y}_{p1}, \ldots, \tilde{y}_{pN})'$. Therefore, $\tilde{y}_p$ is the vector one gets by replacing the ith element in y, which is the observation left out, by the estimate $\hat{s}_\lambda^{-p}(x_p)$. Furthermore, we define $\tilde{s}_\lambda^{-p}(\cdot)$ as the estimate of $s(\cdot)$ given data $\tilde{y}_p$. Then, one can prove (see Wahba [51]) that

$$\tilde{s}_\lambda^{-p}(x_p) = \hat{s}_\lambda^{-p}(x_p)\ ,\ p = \{12, \ldots, N_V(N_V - 1)\}$$

As a next step, recall that $\hat{s}_\lambda(\cdot) = A(\lambda)y$, where $A(\lambda)$ is the influence matrix for the model fitted to all the data, which can be computed by

$$A(\lambda) := B(B'B + \lambda S)^{-1}B'$$

where B and S are defined as in the previous chapter. Thus, when replacing y by $\tilde{y}_p$, we get

$$\left(\tilde{s}_\lambda^{-p}(x_{12}), \ldots, \tilde{s}_\lambda^{-p}(x_{N_V(N_V-1)})\right) = A(\lambda)\tilde{y}_p$$

which finally yields

$$\hat{s}_\lambda^{-p}(x_p) = \tilde{s}_\lambda^{-p}(x_p) = \sum_{q=1}^{N} a_{pq}\tilde{y}_q = \sum_{\substack{q=1 \\ q\neq p}}^{N} a_{pq}y_q + a_{pp}\hat{s}_\lambda^{-p}(x_p) \tag{31}$$

Given

$$\hat{s}_\lambda(x_p) = \sum_{q=1}^{N} a_{pq}y_q \tag{32}$$

and subtracting (31) from (32), we get

$$\hat{s}_\lambda(x_p) - \hat{s}_\lambda^{-p}(x_p) = a_{pp}(y_p - \hat{s}_\lambda^{-p}(x_p))$$

Rearranging the equation finishes the proof:

$$y_p - \hat{s}_\lambda^{-p}(x_p) = \frac{y_p - \hat{s}_\lambda(x_p)}{1 - a_{pp}}$$

The cross validation we are going to use when fitting GAMs is going to be the *generalized cross validation* GCV(λ), which results from (30) by replacing a_{pp} by the average of the trace of $A(\lambda)$. This yields

$$GCV(\lambda) = \frac{1}{N}\sum_{i=1}^{N_V}\sum_{\substack{j=1 \\ i\neq j}}^{N_V}\left(\frac{y_{ij} - \hat{s}_\lambda(x_{ij})}{1 - \frac{trA(\lambda)}{n}}\right)^2 \tag{33}$$

The $GCV(\lambda)$ can be computed faster and, as a consequence, minimized more easily than with the ordinary cross validation approach (28), since one only has to fit the full data once and compute the average of the trace of the influence matrix $A(\lambda)$ instead of using the CPU-intensive leave-one-out method.
Now that we are able to estimate smoothers, we are going to discuss in the next chapter how one can estimate relational data by using generalized linear and generalized additive models.

5 Statistical Regression Models

In this chapter we are going to discuss the *logit model*, the *additive model* (AM) and the *generalized additive model* (GAM). The Logit Model is a *generalized linear model* (GLM) that generalizes linear regression by allowing the *linear model* (LM) to be related to the response variable via a specific link function. The AM, first introduced by Friedman and Stuetze [16], is a linear model including a sum of smooth functions for the influence variables. The GAM, first introduced by Hastie and Tibshirani [23], is a Generalized Linear Model (GLM) with a linear predictor including a sum of smooth functions of covariates. Just as the GLM generalizes the LM, the GAM generalizes the AM. The GAM can even be seen as a generalization of the GLM and therefore of ordinary linear regression.

5.1 Regression Review: The Logit Model

Since we are interested in a binary target variable Y_{ij}, for which an edge either exists or does not exist between two actors, the codomain of the model has to be restricted to [0,1]. In the following, we denote random variables with capital letters, while a specific realization is denoted by lower-case characters. The aim of binary regression is to model and estimate the effects of given covariates $x_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ for the conditional probability

$$\begin{aligned} \pi_{ij} &= \mathbb{E}(Y_{ij} \mid X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}) \\ &= \mathbb{P}(Y_{ij} = 1 \mid X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}) \end{aligned}$$

where $ij \in \{12, 13, \ldots, N_V(N_V - 2), N_V(N_V - 1)\}$, $p \in \mathbb{N}$, for the occurence of y_{ij}. Note that $N = |\{12, 13, \ldots, N_V(N_V - 2), N_V(N_V - 1)\}|$ is the number of possible edges. Modeling the probability of the occurrence of an edge with the linear model

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, has serious disadvantages, such as a restriction for the parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)'$, which are difficult to manage. For all possible values β and x_{ij} the linear predictor

$$\eta_{ij} = \beta_0 + \beta_1 x_{ip1} + \cdots + \beta_p x_{ijp} = x_{ij}^{'} \beta \tag{34}$$

has to attain a value in the interval [0,1]. Furthermore, the error value ε_{ij}'s error variance $Var(\varepsilon_{ij}) = Var(y_{ij}|x_{ij})$ is not homoscedastic, i.e., is equal to a constant

σ^2. This is based on the premise that y_{ij} is Bernoulli distributed and therefore one gets $Var(y_{ij}) = \pi_{ij}(1 - \pi_{ij})$, which results in different variance values for each dyad (ij).

For that reason, a common way to fit models with binary response values is to link the probability π_{ij} to the linear predictor η_{ij} through a relation of the form

$$\begin{aligned} \pi_{ij} &= \mathbb{P}(Y_{ij} = 1 | X_{ij1} = x_{ij1}, \dots, X_{ijp} = x_{ijp}) \\ &= h(\eta_{ij}) = h(\beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp}) \end{aligned} \tag{35}$$

The function $h(\cdot)$, which is called a *response function*, has to be a distribution function from the exponential family, such that for any β and any x_{ij} one gets $h(\eta) \in [0, 1]$. The covariate's effects are expected to be linear within the parameters, but the distribution of the response function, and therefore also of the link function, can be general. Since $h(\cdot)$ is strictly monotonically increasing there exists an inverse function $g(\cdot) = h^{-1}(\cdot)$, called a *link function*, and thus relation (35) can also be written as

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} = g(\pi_{ij})$$

One approach to define a suitable function $h(\cdot)$ is the following:

$$\pi_{ij} = h(\eta_{ij}) = \frac{exp(\eta_{ij})}{1 + exp(\eta_{ij})}$$

This yields, for the link function

$$g(\pi_{ij}) = log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \eta_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp}$$

A model with such a link function is called a *logit model*, since one gets a linear model for logarithmized odds. Furthermore, note that multiplying $g(\cdot)$ with the exponential function yields

$$\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = exp(\beta_0) exp(\beta_1 x_{ij1}) \cdot \dots \cdot exp(\beta_p x_{ijp}) \tag{36}$$

This relation shows that the covariates take a multiplicative exponential effect on the chance $\dfrac{\pi_{ij}}{1 - \pi_{ij}}$. For a random variable

$$Y_{ij} : \Omega \to \{0, 1\}, \quad \omega \to (Y_{ij}(\omega))_{i,j \in \{1,\dots,N\}}$$

we can now define the probability function of Y_{ij} as

$$\mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij}) = h(\eta_{ij}) = \frac{exp(\eta_{ij})}{1 + exp(\eta_{ij})} \tag{37}$$

where $X_{ij} = x_{ij}$ is short for $X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}$.
Then, one can compute the odds of occurence of edge (i, j), conditional on the covariates:

$$\begin{aligned}
\frac{\mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij})}{\mathbb{P}(Y_{ij} = 0|X_{ij} = x_{ij})} &= \frac{\mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij})}{1 - \mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij})} \\
&= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \Big/ \Big(1 - \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}\Big) \\
&= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \Big/ \Big(\frac{1}{1 + \exp(\eta_{ij})}\Big) \\
&= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \cdot \Big(1 + \exp(\eta_{ij})\Big) \\
&= \exp(\eta_{ij})
\end{aligned}$$

This implies the equation:

$$\text{logit}(\mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij})) = \eta_{ij} \tag{38}$$

This yields the following *ceteris paribus* interpretation for the parameters $\beta_k, k \in \{1, \ldots, p\}$: If, for instance, the value of x_{ijk} increases by 1, while all other values remain the same, then quotient (36) gets multiplied by $exp(\beta_k)$, since

$$\begin{aligned}
&\frac{\mathbb{P}(Y_{ij} = 1|X_{ij1} = x_{ij1}, \ldots, X_{ijk} = x_{ijk} + 1, \ldots, X_{ijp} = x_{ijp})}{\mathbb{P}(Y_{ij} = 0|X_{ij1} = x_{ij1}, \ldots, X_{ijk} = x_{ijk} + 1, \ldots, X_{ijp} = x_{ijp})} \quad = \\
&\quad \frac{\mathbb{P}(Y_{ij} = 1|X_{ij1} = x_{ij1}, \ldots, X_{ijk} = x_{ijk}, \ldots, X_{ijp} = x_{ijp})}{\mathbb{P}(Y_{ij} = 0|X_{ij1} = x_{ij1}, \ldots, X_{ijk} = x_{ijk}, \ldots, X_{ijp} = x_{ijp})} \quad \cdot \quad exp(\beta_k)
\end{aligned}$$

As a result, the chance for $\dfrac{\pi_{ij}}{1 - \pi_{ij}}$

- increases, if $\beta_k > 0$
- stays the same, if $\beta_k = 0$
- decreases, if $\beta_k < 0$

5.2 The Additive Model

The class of *additive models* (AM) is a useful extension of the class of linear models, since it mitigates the strong linear assumption of each covariate $X_1, \ldots, X_p$, $p \in \mathbb{N}$ towards the response Y. Linear regression can be seen as an approach to estimate $\mathbb{E}(Y|X_1, \ldots, X_p)$, by assuming the model structure to be

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij1} + \cdots + \alpha_q x_{ijq} + \varepsilon_{ij}$$

with parameters $\alpha_0, \ldots, \alpha_p$ and i.i.d. $\varepsilon_{ij} \sim N(0, \sigma^2)$. For additive models, we generalize the linear predictor with smooth functions $s(\cdot)$

$$y_{ij} = \alpha_0 + s_1(x_{ij1}) + \cdots + s_q(x_{ijq}) + \varepsilon_{ij} \tag{39}$$

where $\mathbb{E}\Big(s_k(X_k)\Big) = 0$ for all $k = \{1, \ldots, p\}$. Note that including more than one function into the model causes an identifiability problem. Two functions are only estimable within an additive constant. The resulting problem is that a constant could be synchronously added to the first and subracted from the second function, without changing the model prediction. We are going to represent the additive model using penalized B-splines just as we discussed in chapter 4.

Since the arms trade model will include binary covariates as well, which do not have to be smoothed, we are going to rewrite (39) as

$$y_{ij} = s_1(x_{ij1}) + \cdots + s_p(x_{ijp}) + Z_{ij}\beta + \varepsilon_{ij} \tag{40}$$

where $\beta' = (\beta_{p+1}, \ldots \beta_q)$ is a vector of parameters and $Z_{ij} = (x_{ij(p+1)}, \ldots, x_{ijq})$ is the vector of covariates we assume to have a linear effect.

An advantage of the linear model towards other models is that it is additive in the predictors' effects. This yields the following opportunity: If a linear model is fitted, it is possible to investigate the predictors' effects separately, since we assume the covariates to be independent of each other. If one holds all but one predictor fixed and takes a look at the variation of the fitted response, then it does not depend on the values of the other predictors. When taking a look at additive models we can observe that they retain this important feature of linear models. Their predictors' effects are additive as well (see 40), which yields the conclusion that once the additive model is fitted, we are able to examine the functions of the covariates separately. Therefore, we can analyze the roles of the predictors in modeling the response variable individually.

But how can one estimate $s_1(\cdot), \ldots, s_p(\cdot), \beta$ simultaneously? In chapter 4 we have only considered univariate smoothing. There are several methods to get estimators $\hat{s}_1(\cdot), \ldots, \hat{s}_p(\cdot), \hat{\beta}$ for (40). A simple, and therefore commonly used method is the *backfitting* approach, first introduced by Breiman and Friedman [5]. The main idea of backfitting is to estimate each smooth component $\hat{s}_1(\cdot), \ldots, \hat{s}_p(\cdot), \hat{\beta}$ by iteratively smoothing partial residuals, with respect to the covariates the smooth relates to. The partial residuals, which correspond to the jth smooth term, are the residuals we gain by subtracting all but the jth smooth from the response variable. A reason for the popularity of this method is undoubtably that the estimates for $s_1(\cdot), \ldots, s_p(\cdot)$ can be realized for any simple smoothing methods. It even allows the combination of different smoothing methods.

If one is neglecting the error ε in the additive model one approximately gets for all $k \in \{1, \ldots, p\}$ that

$$s_k \approx y - s_1 - \cdots - s_{k-1} - s_{k+1} - \cdots - s_p - Z\beta$$

Therefore, for given estimators $\hat{s}_1, \ldots, \hat{s}_{k-1}, \hat{s}_{k+1}, \ldots, \hat{s}_p, \hat{\beta}$, the expression

$$y - \sum_{\substack{i=1 \\ i \neq k}}^{p} \hat{s}_i - Z\hat{\beta} \tag{41}$$

can be seen as a partial vector of residuals without $\hat{s}_k$. As a next step, we are going to refer to R_k, $k \in \{1, \ldots, p\}$ as the design matrix of the kth covariate as defined in (18). Then, we define the to s_k corresponding spline smoother as

$$K_k := (R_k' R_k + \lambda_k S_k)^{-1} R_k' \tag{42}$$

where S_k refers to the penalty matrix (22) introduced in chapter 4.3. As a consequence, one can estimate $\hat{s}_k$ by applying the spline smoother (42) on the vector of residuals (41). One gets

$$\hat{s}_k = K_k(y - \sum_{\substack{i=1 \\ i \neq k}}^{p} \hat{s}_i - Z\hat{\beta})$$

Based on simple starting assumptions, one can now iteratively estimate $\hat{s}_1, \ldots, \hat{s}_p$, $\hat{\beta}$:

1. Fix $\hat{s}_1, \ldots, \hat{s}_p, \hat{\beta}$. For instance: $\hat{s}_1 \equiv 0, \ldots, \hat{s}_p \equiv 0, \hat{\beta} \equiv 0$

2. For k in $1 : p$ improve estimator $\hat{s}_k$ by

$$\hat{s}_k = K_k(y - \sum_{\substack{i=1 \\ i \neq k}}^{p} \hat{s}_i - Z\hat{\beta})$$

3. Improve the estimator $\hat{\beta}$ by

$$\hat{\beta} = (Z'Z)^{-1}Z'(y - \sum_{i=1}^{p} \hat{s}_i)$$

4. Repeat steps 2 and 3 until the estimated functions stop changing less than a given error

Note that by adjusting K_k one can apply almost any other smoothing method, such as polynomial splines, kriging etc. For more in-depth discussions we refer to Härdle et al. [21]. The backfitting algorithm discussed in this chapter is only the simplest version of backfitting. The GAM introduced in the next chapter will be fit by *penalized iteratively re-weighted least squares* (P-IRLS), a weighted version of the backfitting algorithm.

5.3 The Generalized Additve Model

The Generalized Linear Logit Model introduced earlier boasts the ability to model binary response variables. As a next step, one can generalize the strong assumption of the covariates' $X_1, \ldots, X_p$ linear relation in the model by assuming a nonparametric effect, and thereby an additive extension of the family of GLMs. A GAM is a GLM with a linear predictor including a sum of smooth functions of covariates. Hence, they extend GLMs the same way that additive models extend linear models. As a consequence, the linear predictor now expresses the outcome of some known monotonic function of the expected value of the response, while the response follows any exponential family distribution. As already seen in chapter 3.6 it seems to make sense to assume a nonlinear effect for the degree distribution. Therefore we extend the linear predictor (34) with smooth functions $s_1(\cdot), \ldots, s_p(\cdot)$ to

$$\eta_{ij} = s_1(x_{ij1}) + \cdots + s_p(x_{ijp}) + Z\beta + \varepsilon_{ij} \tag{43}$$

where the errors ε_{ij} are independent of the x_{ij}, with $\mathbb{E}(\varepsilon_{ij}) = 0$ for all $ij = \{12, \ldots, N_V(N_V - 1)\}$ and $\mathbb{E}(s_k(X_{ijk})) = 0$, $k \in \{1, \ldots, p\}$, since otherwise there would be free constants in each of the functions. $\beta = (\beta_{p+1}, \ldots, \beta_q)$ and $Z = (x_{ij(p+1)}, \ldots, x_{ijq})$ are defined as in the previous chapter. The non-parametric functions $s_1(\cdot), \ldots, s_p(\cdot)$ are smooth functions, one for each covariate $x_{ij1}, \ldots, x_{ijp}$. It should be mentioned at this point that for linear functions $s_1(\cdot), \ldots, s_p(\cdot)$ one gets

the linear predictor (34). Similar to the logit model introduced in chapter 5.1 we define the response function as

$$\pi_{ij} = h(\eta_{ij}) = \frac{exp(\eta_{ij})}{1 + exp(\eta_{ij})}$$

with the corresponding link function

$$g(\pi_{ij}) = log\Big(\frac{\pi_{ij}}{1 - \pi_{ij}}\Big) = \eta_{ij}$$

Note that an additive model's estimated functions are the analogues of the coefficients in linear models. For now, we treat each of the functions $s_1(\cdot), \ldots, s_p(\cdot)$ as a smooth function which can individualy be estimated by a scatterplot smoother. GAMs do not incorporate terms of interaction between two covariates. Models which extend the GAM by terms of interaction are called *generalized additive mixed models* (GAMM) and will not be considered in this paper. Due to this, the GAM can be seen as an extension of the GLM and even of the linear model. These models are suitable for exploring the data set and visualizing the relationship between the response variable Y and the independent covariates $X_1, \ldots, X_p$. We are going to estimate the nonparametric functions $s_1(\cdot), \ldots, s_p(\cdot)$ by using penalized B-splines and an iterative method called penalized iteratively re-weighted least square (P-IRLS), a weighted version of the backfitting algorithm. The question of how one can appropriately estimate a smooth function $s_k(\cdot)$ was already discussed in chapter 4.

Recall from chapter 4.2 that each smooth funtion $s_k(\cdot)$ can be estimated by

$$\hat{s}_k(x) = R_k \hat{\alpha}_k$$

where

$$R_k = \begin{pmatrix} B_1(x_{12k}) & \ldots & B_t(x_{12k}) \\ \vdots & & \vdots \\ B_1(x_{(N_V-1)N_V k}) & \ldots & B_t(x_{(N_V-1)N_V k}) \end{pmatrix} \quad , \quad \hat{\alpha}_k = \begin{pmatrix} \hat{\alpha}_{1k} \\ \vdots \\ \hat{\alpha}_{tk} \end{pmatrix}$$

for $t \in \mathbb{N}$.

By defining the rows of a matrix X by

$$X_{ij} := \quad \Big(B_1(x_{ij1}), B_2(x_{ij1}) \ldots, B_t(x_{ij1}), \ldots,$$

$$B_1(x_{ijp}), B_2(x_{ijp}) \dots, B_t(x_{ijp}),$$
$$x_{ij(p+1)}, \dots, x_{ijq}\Big) \tag{44}$$

and the paramter vector γ as

$$\gamma := (\alpha_{11}, \alpha_{21}, \dots, \alpha_{t1}, \dots, \alpha_{1p}, \alpha_{2p}, \dots, \alpha_{tp}, \beta_{p+1}, \dots, \beta_q)' \tag{45}$$

we can write the predictor (43) in linear form

$$y = X\gamma + \varepsilon$$

where $y := (y_{12}, \dots, y_{N_V(N_V-1)})$, $\varepsilon := (\varepsilon_{ij}, \dots, \varepsilon_{N_V(N_V-1)})$. For similar reasons we can also write $\eta = X\gamma$.

When using GAMs we assume that the observations y_{ij} are coming from a distribution in the exponential family with probability denisty function

$$f(y|\theta) = exp\left(\frac{y\theta - b(\theta)}{\phi} \cdot \omega + c(y, \phi, \omega)\right) \tag{46}$$

where $b(\theta)$ is an at least twice differentiable function, ϕ is called the *dispersion paramter* and ω is a known *prior weight*. The Bernoulli distribution can be shown to be an exponential family distribution. Therefore, let

$$\theta = log\Big(\frac{\pi}{1-\pi}\Big)$$

Note that θ is a function of π. However, for the sake of simplicity we will write θ instead of $\theta(\pi)$. Then, the Bernoulli probability density function

$$f(y|\pi) = P(Y = y) = \pi^y(1-\pi)^{1-y}$$

can be written in the exponential family form

$$f(y|\theta) = exp(y\theta - log(1 + exp(\theta))) \tag{47}$$

where $b(\theta) = log(1 + exp(\theta))$, $\phi = \omega = 1$ and $c(y, \phi, \omega) = 0$. Furthermore, one can easily show that

$$\mathbb{E}(y) = \pi = b'(\theta) = \frac{exp(\theta)}{1 + exp(\theta)} \quad , \quad Var(y) = \pi(1-\pi) = b''(\theta) = \frac{exp(\theta)}{(1 + exp(\theta))^2}$$

When conducting maximum-likelihood estimation with this model, one can take advantage of the pratical feature that GAMs can be estimated by the P-IRLS method. For this method, one has to guesstimate the parameter vector $\hat{\gamma}^{[0]}$ and calculate the linear predictor $\hat{\eta}_{ij}^{[0]} = X_{ij}\hat{\gamma}^{[0]}$ in order to obtain the fitted values $\hat{\pi}_{ij}^{[0]} = h(\hat{\eta}_{ij}^{[0]})$. Continue by iterating k. We calculate the working dependent variable

$$z_{ij}^{[k]} := \eta_{ij}^{[k]} + (y_{ij} - \pi_{ij}^{[k]}) \cdot g'(\pi_{ij}^{[k]})$$

where $g'(\pi_{ij}^{[k]})$ is the derivative of the link function evaluated by $\gamma^{[k]}$. Furthermore, we need to calculate the iterative weights

$$w_{ij}^{[k]} \propto \frac{1}{b''(\theta)g'(\pi_{ij}^{[k]})^2}$$

where $b''(\theta)$ was evaluated by $\hat{\gamma}_{ij}^{[k]}$. Note that $w_{ij}^{[k]}$ is inversely proportional to the variance of $z_{ij}^{[k]}$ (see Rodriguez [43]). We finally get an improved estimate $\hat{\gamma}^{[k+1]}$ by minimizing the penalized weighted least square estimate

$$||\sqrt{W}(z - X\gamma)||^2 + \lambda_1\gamma' S_1\gamma + \cdots + \lambda_p\gamma' S_p\gamma$$

where X is the model matrix defined in (44), W is a diagonal matrix with weights $w_{ij}^{[k]}$ as entries, S_k , $k \in \{1, \dots p\}$ is a matrix of known coefficients as defined in (22) and $z = (z_{12}^{[k]}, \dots, z_{N_V(N_V-1)}^{[k]})$ is a response vector. This algorithm can be repeated until the estimates change less than a specified constant. McCullagh and Nelder [36] successfully proved that the P-IRLS algorithm is equivalent to Fisher scoring and results in maximum-likelihood estimation. For a more detailed discussion we refer to Wood [55] and Rodriguez [43].

In summary, it can be said that in order to estimate a GAM, one hast to turn the GAM into a GLM with coefficients γ and a smoothing parameter λ. Hence, one has to choose fitting basis functions $B_i(x)$. The smoothing parameter λ acts as a as trade-off parameter in order to control the relative weight given to the two conflicting goals: matching the data and estimating a smooth funtion. A common way to estimate λ is by using cross-validation. Finally, the parameter vector γ can be estimated using the penalized iteratively re-weighted least square method.

6 Modeling Networks with GLMs and GAMs

In this chapter we are going to discuss some approaches to modelling networks with GLMs and GAMs. For GLMs we will present a pseudo-likelihood approach which uses a bootstrapping technique to adjust the biased coefficient estimates. Besides the strategy for modeling networks with GLMs, we will discuss two approaches for modeling networks with GAMs. The first approach, however, will turn out to be unsuitable for our purposes and the second approach, which does not consider a network's dependency structure will provide biased results. Nevertheless, the estimated smooth functions will visualize a general impression of the covariates' effects, and therefore, will justify the generalization of the ERGM, which will then be discussed in chapter 7.

6.1 First Approach to Modeling Networks with GLMs and GAMs

After having introduced the generalized additive model in the previous chapter, the question arises around how one can fit networks with these models. According to definition 3 a directed network on N_V nodes in year x can be written as an adjacency matrix $A = (a_{ij})$, where $a_{ij} \in \{0, 1\}$ for all $ij, \in \{12, \ldots, N_V(N_V - 1)\}$. Here $a_{ij} = 1$ means that an edge exists between actor i and actor j, i.e., country i exports weapons to country j, and $a_{ij} = 0$ indicates that there is no arms flow from country i to country j. Since the model does not take loops into account, i.e., the arms trade inside countries, we define $a_{ii} = 0$ for all $i \in \{1, \ldots, N_V\}$. This notation accents once more that we are only considering binary networks. This means the weighting of ties is not incorporated into the model and the only interest is whether two countries trade weapons or not. We take the particular entry a_{ij} of A as a manifestation of the Bernoulli variable Y_{ij}. With the additive predictor

$$\eta_{ij} = s_1(x_{ij1}) + \cdots + s_p(x_{ijp}) + Z_{ij}\beta$$

we define the probability function of Y_{ij} as

$$\mathbb{P}(Y_{ij} = 1 | X_{ij}^A = x_{ij}) = h(\eta_{ij}) = \frac{exp(\eta_{ij})}{1 + exp(\eta_{ij})} \tag{48}$$

where $X_{ij}^A = x_{ij}$ is short for $X_{ij1}^A = x_{ij1}, \ldots, X_{ijq}^A = x_{ijq}$, with $q \geq p, \quad q, p \in \mathbb{N}$. Just as defined in chapter 5.2 the vector of covariates we assume to have a linear effect is described by Z_{ij}. The A in $X_{ij}^A = x_{ij}$ simply indicates that the dependent variables do not necessarily only exist as exogenous variables, but can

also include dyad-specific characteristics from network A such as the existence of the reciprocal tie Y_{ji} or the sender's or receiver's in- or out-degree. We will amplify this in a later paragraph. Note that the additive predictor is just a generalization of a linear predictor. Consequently, the following considerations also apply for GLMs. Furthermore, we are going to define the term *dyad* in this chapter slightly differently from how we defined it in chapter 1. In the following, let a dyad be the *directed* relation from i to j, i.e., an edge e_{ij} from i to j either does or does not exist. In chapter 1 we have defined a dyad in general as the relation between two actors, which could either be mutual, unidirectional or null.

This first and simple model treats all dyads as pairwise independent, which means that one assumes the occurrence of Y_{ij} as independent from the occurrence of other ties and, consequently, independent from the structure of the network. In this simple case, the estimation of the parameter vector γ can then be computed using normal pseudo-likelihood estimation

$$\text{plik}(\gamma) = \prod_{\substack{i,j=1 \\ i \neq j}}^{N_V} \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}$$

The pseudo-likelihood approach is simple and fast, but contains the substantial disadvantage that the assumed hypothesis of the independence of dyads turns out to be erroneous in most cases. The presence of network data is inextricably connected with the presence of relational data. In the case of the arms trade network it is, for instance, reasonable to assume that the occurrence of a tie between countries i and j has an effect on the occurrence of a tie between countries i and k. These dependency relations are disregarded with the pseudo-likelihood approach.

In order to incorporate the dependency structure of a dyad while avoiding intensive MCMC methods as discussed in chapter 3.2 for the ERGM, consider the following approach, which was proposed by Kauermann[1]:
Assume for the sake of simplicity that the number of actors N_V in the observed network A is even. As a first approach, one can posit that the occurrence of a tie Y_{ij} only depends on the dyads either directly connected to actor i or j or some exogenous covariates. This idea allows the assumption that the occurrence of ties Y_{ij} and Y_{st}, with $i, j, s, t \in \{1, \ldots, N_V\}, i \neq j \neq s \neq t$ are independent of each other, given the rest of the network. Therefore, given a network of N_V nodes, one can arrange the actors into pairs of two, e.g., $\mathbb{D}(N_V) = \{(12), (34), \ldots, ((N_V - 1)N_V)\}$, and take

[1]This paper has not been published at the time of the study (05/2015).

the occurrence of Y_{ij} as independent from Y_{st}, with $(ij), (st) \in \mathbb{D}(N_V), (ij) \neq (st)$, conditioning on all other dyads $Y^c_{\mathbb{D}(N_V)} := \{Y_{kl} \mid (kl) \notin \mathbb{D}(N_V)\}$ in the network. Formally,

$$Y_{12} \perp\!\!\!\perp Y_{34} \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{(N_V-1)N_V} \quad | \quad Y^c_{\mathbb{D}(N_V)}$$

Since we are investigating directed networks, this also implies

$$Y_{21} \perp\!\!\!\perp Y_{43} \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{N_V(N_V-1)} \quad | \quad Y^c_{\mathbb{D}(N_V)}$$

and any other combination of mutually independent dyads with pairs in $\mathbb{D}(N_V)$. By including proper dyad-specific characteristics

$$\Phi : \mathcal{A}(N_V)^c_{\mathbb{D}(N_V)} \to \mathbb{R}^\ell \ , \ A^c_{\mathbb{D}(N_V)} \to (\Phi_1(A^c_{\mathbb{D}(N_V)}), \ldots, \Phi_\ell(A^c_{\mathbb{D}(N_V)}))'$$

one can model

$$logit(\mathbb{P}(Y_{ij} = 1 \mid Y^c_{\mathbb{D}(N_V)} = A^c_{\mathbb{D}(N_V)}, X^{ex}_{ij} = x^{ex}_{ij})) = \beta_0 + s_{en}(\Phi(A^c_{\mathbb{D}(N_V)})) + s_x(x^{ex}_{ij}) \quad (49)$$

where

- $A^c_{\mathbb{D}(N_V)}$ is the network A without dyads in $\mathbb{D}(N_V)$
- $\mathcal{A}(N_V)^c_{\mathbb{D}(N_V)}$ is the set of all possible $A^c_{\mathbb{D}(N_V)}$
- $s_{en}(\Phi(A^c_{\mathbb{D}(N_V)})) := s_1(\Phi_1(A^c_{\mathbb{D}(N_V)})) + \cdots + s_\ell(\Phi_\ell(A^c_{\mathbb{D}(N_V)}))$ are dyad-specific characteristics
- $s_x(x^{ex}_{ij}) := s_{\ell+1}(x_{ij(\ell+1)}) + \cdots + s_q(x_{ijq})$ are conventional covariates
- $(X^{ex}_{ij} = x^{ex}_{ij}) := (X^{ex}_{ij(\ell+1)} = x^{ex}_{ij(\ell+1)}, \ldots, X^{ex}_{ijq} = x^{ex}_{ijq})$

(49) can be modeled for any dyad $(ij) \in \mathbb{D}(N_V)$. Note that we treat $\Phi(A^c_{\mathbb{D}(N_V)})$ as regular covariates and not as endogenous statistics as in the ERGM. By proper characteristics we define statistics as the in- or out-degree of actor i or j or network statistics built from k-stars or triangles, which do not violate the independence assumption made above. More complex statistics, such as loops of size 4 or higher, can not be incorporated into this model without violating the independence assumptions. Otherwise the occurrence of Y_{ij} might depend on Y_{st}, even though $(ij), (st) \in \mathbb{D}(N_V)$. This approach has the crucial advantage that, conditional on $\mathbb{A}^c_{\mathbb{D}(N_V)}$, the results are not biased, i.e., we can compute proper parameter estimates and standard deviations.

	1	2	3	4	5	6
1	0	1	2	3	4	5
2	6	0	4	5	3	2
3	7	9	0	1	5	3
4	8	10	6	0	2	4
5	9	8	10	7	0	1
6	10	7	8	9	6	0

Table 2: Latin square with a unique diagonal for N=6

Furthermore, this approach has another huge advantage: Networks can be simulated faster by improving computationally intensive MCMC-methods such as Gibbs sampling or Metropolis-Hasting. Therefore, let $\mathcal{D}$ be a sequence of sets $\mathbb{D}_n(N_V), n \in \{1, \ldots, 2(N_V - 1)\}$, such that each index pair (ij), where $i, j \in \{1, \ldots, N_V\}, i \neq j$, is an element of exactly one set $\mathbb{D}_n(N_V)$. Then, a network can be simulated by using so-called *latin squares* with a unique diagonal (see Andersen and Hilton [2]).

Take, for instance, a network on $N_V = 6$ nodes. The numbers in the latin square shown in table 2 can be seen as simulation steps of the parallelized Gibbs sampling. According to table 2 ties Y_{12}, Y_{34} and Y_{56} can be simulated in parallel in the first step, due to their independence, followed by Y_{13}, Y_{26} and Y_{45} etc. Finally, with $N_V/2$ computing cores one can simulate an entire network in just $2(N_V - 1)$ steps. This means that the simulation steps only increase linearly for an even number of nodes. Parallel simulation via Gibbs sampling is also possible for an odd number N_V of actors, but takes a few more steps.

A first approach to modelling the probability of occurrence of a dyad Y_{ij} by a logit model could be to assume that Y_{ij} depends on the existence of the reciprocal dyad Y_{ji} and on the in- and out-degree of actors i and j. As a consequence, $\Phi(A^c_{\mathbb{D}(N_V)})$ from equation (49) is defined as

$$\Phi(A^c_{\mathbb{D}(N_V)}) := \left(a_{ji}, \sum_{\substack{k=1 \\ k \neq j}}^{N_V} a_{ik}, \sum_{\substack{k=1 \\ k \neq i}}^{N_V} a_{jk}, \sum_{\substack{k=1 \\ k \neq j}}^{N} a_{ki}, \sum_{\substack{k=1 \\ k \neq i}}^{N_V} a_{kj} \right) \tag{50}$$

Together with the exogenous covariates X_{ij}^{ex}, one can model

$$logit(\mathbb{P}(Y_{ij} = 1 \mid Y^c_{\mathbb{D}(N_V)} = A^c_{\mathbb{D}(N_V)}, X^{ex}_{ij} = x^{ex}_{ij}) =$$

$$\gamma_0 + \gamma_1 a_{ji} + \gamma_2 \sum_{\substack{k=1 \\ k\neq j}}^{N_V} a_{ik} + \gamma_3 \sum_{\substack{k=1 \\ k\neq i}}^{N_V} a_{jk} + \gamma_4 \sum_{\substack{k=1 \\ k\neq j}}^{N} a_{ki} + \gamma_5 \sum_{\substack{k=1 \\ k\neq i}}^{N_V} a_{kj} + \gamma_{ex} x^{ex}_{ij} \tag{51}$$

where $\gamma_{ex} x^{ex}_{ij}$ is short for $\gamma_6 x_{ij6} + \cdots + \gamma_p x_{ijp}$. Interpreting the parameters can be done just as in a regular GLM, since we assume ties Y_{ij}, $(ij) \in \mathbb{D}(N_V)$ to be independent of each other. If $\gamma_2 > 0$, then the higher the out-degree of the sender i, the more likely the occurrence of an edge from i to j is. However, as already discussed in chapter 3.6 for the ERGM it is reasonable to assume a nonparametric effect for the non-binary covariates. Then, model (51) changes to

$$logit(\mathbb{P}(Y_{ij} = 1 \mid Y^c_{\mathbb{D}(N_V)} = A^c_{\mathbb{D}(N_V)}, X^{ex}_{ij} = x^{ex}_{ij}) =$$

$$\gamma_0 + \gamma_1 a_{ji} + s_2(\sum_{\substack{k=1 \\ k\neq j}}^{N_V} a_{ik}) + s_3(\sum_{\substack{k=1 \\ k\neq i}}^{N_V} a_{jk}) + s_4(\sum_{\substack{k=1 \\ k\neq j}}^{N_V} a_{ki}) + s_5(\sum_{\substack{k=1 \\ k\neq i}}^{N_V} a_{kj}) + s_x(x^{ex}_{ij}) \tag{52}$$

where $s_x(x^{ex}_{ij})$ is short for $s_6(x_{ij6}) + \cdots + s_p(x_{ijp})$. Note that for $s_k(x_{ijk}) = \gamma_k x_{ijk}$, $k \in \{6, \ldots, p\}$ one gets the linear relation assumed in model (51).

Unfortunately, this simple and fast method turns out to be unsuitable for the arms trade network or any other sparse network. To understand the reason, consider the network for the year 2012. For this year the directed network contains $N_V = 192$ actors, which implies $N_V^2 - N_V = 36672$ possible edges. Compared to this large number, the actual number of observed edges $N_E = 376$ is extremely low. When randomly drawing independent pairs (ij) out of a network with 192 actors, we get a sample of the size $\frac{N_V}{2} = 96$, but the chance of drawing a pair with an edge is just a little higher than one percent. As a result, the sampled data is not suitable for inference, since the overwhelming number of the attribute of concern is set zero. In some extreme, but not too improbable cases, this can even mean that a drawn sample does not hold a single observation with an edge.

6.2 The Bootstrap Logit Model

After the model introduced in the previous section turned out to be unsuitable for our case, we have to consider a different approach: Again, a simple and fast strategy is to take the observations of the dependent variable y_{ij} as independent and calculate

the log pseudo-likelihood

$$Pliklog(\gamma) = \sum_{\substack{i,j=1 \\ i \neq j}}^{N_V} log\mathbb{P}(Y_{ij} = 1 \mid Y_{ij}^c = A_{ij}^c, X_{ij}^{ex} = x_{ij}^{ex}) \tag{53}$$

However, by calculating the pseudo-likelihood we face the same problem as discussed earlier in this chapter. Even though the calculation is simple and fast, the dependency structures in the network are ignored, since we treat the observations of the response as independent of each other. Therefore, the parameter estimates are biased and the variance estimates appear unreasonable. For this reason, we suggest a bootstrapping strategy to adjust the parameter and variance estimates. However, the following bootstrapping approach is only reasonable for models of type (49), i.e., for models without smooth functions $s(\cdot)$. For a general introduction to bootstrapping we refer to Efron and Tibshirani [13], Shao and Tu [45] and Davidson and Hinkley [9].

In the following, let the vector $\hat{\gamma}$ be the pseudo-likelihood estimate of γ for network A. Via the MCMC algorithm we simulate a new network A^* by using the pseudo-likelihood estimated $\hat{\gamma}$ as the parameter. An approach for simulating networks using MCMC was introduced by Snijders [48] and already discussed in chapter 3.3.

Once a new network A^* has been simulated, one can estimate the pseudo-likelihood

$$Pliklog(\gamma) = \sum_{\substack{i,j=1 \\ i \neq j}}^{N_V} log\mathbb{P}(Y_{ij} = 1 \mid Y_{ij}^c = (A_{ij}^*)^c, X_{ij}^{ex} = x_{ij}^{ex})$$

of the simulated network A^* and refer to the newly obtained estimator by $\hat{\hat{\gamma}}^*$. The principal idea of bootstrapping is the assumption that we can draw inference from the simulated distribution of $\hat{\hat{\gamma}}^* - \hat{\gamma}$ about the difference of interest $\hat{\gamma} - \gamma$. As a consequence, one can rectify the biased pseudo-likelihood estimate $\hat{\gamma}$. The idea is to consider the bootstrap bias $b(\hat{\gamma}^*) = \mathbb{E}^*_{\hat{\gamma}}(\hat{\hat{\gamma}}^*) - \hat{\gamma}$ as an estimate for the unknown bias $b(\hat{\gamma}) = \mathbb{E}_{\hat{\gamma}}(\hat{\hat{\gamma}}) - \hat{\gamma}$. Here, we denote the pseudo-likelihood estimate computed from simulated networks with parameter $\hat{\gamma}$ with $\hat{\hat{\gamma}}$. The bootstrap bias $b(\hat{\gamma}^*)$ can be approximated by simulating B networks $A^{*1}, \ldots, A^{*B}$ and by computing the pseudo-likelihood estimates $\hat{\hat{\gamma}}^{*1}, \ldots, \hat{\hat{\gamma}}^{*B}$ for each of them. Since calculating the ideal bootstrap sample would result in high computational cost, we proceeded by

drawing B new networks. We then estimate $\mathbb{E}^*_{\hat{\gamma}}(\hat{\hat{\gamma}}^*)$ by

$$\hat{\hat{\gamma}}^*(\cdot) := \frac{1}{B}\sum_{r=1}^{B} \hat{\hat{\gamma}}^{*r}$$

We can then write

$$\hat{b}^*(\hat{\gamma}) = \hat{\hat{\gamma}}^*(\cdot) - \hat{\gamma}$$

If we now assume $b(\hat{\gamma}) \approx b(\gamma)$, where $b(\gamma) = \mathbb{E}_\gamma(\hat{\gamma}) - \gamma$, this yields the bias-adjusted estimator $\bar{\gamma}$

$$\begin{aligned} \bar{\gamma} &= \hat{\gamma} - \hat{b}^*(\hat{\gamma}) \\ &= \hat{\gamma} - (\hat{\hat{\gamma}}^*(\cdot) - \hat{\gamma}) \\ &= 2\hat{\gamma} - \hat{\hat{\gamma}}^*(\cdot) \end{aligned}$$

If one is interested in a suitable estimate for the unknown variance $Var(\hat{\gamma})$, one can draw on the boostrapping technique once again and compute the bootstrap variance $Var^*(\hat{\hat{\gamma}}^*)$. With the assumption $Var(\hat{\gamma}) \approx Var(\gamma)$ one can draw conclusions about the actual variance of interest.

Let $\hat{\hat{\gamma}}^{*1}, \ldots, \hat{\hat{\gamma}}^{*B}$ be the pseudo-likelihood estimates of the simulated networks $A^{*1}, \ldots, A^{*B}$. Then, one can compute the bootstrap variance via

$$Var^*(\hat{\hat{\gamma}}^*) = \frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\hat{\gamma}}^{*r} - \hat{\hat{\gamma}}^*(\cdot)\right)^2$$

where $\hat{\hat{\gamma}}^*(\cdot)$ is the arithmetic mean of $\hat{\hat{\gamma}}^{*r}$, $r \in \{1, \ldots, B\}$ as defined above. This result yields the bootstrap-estimated standard deviation

$$se^*(\hat{\hat{\gamma}}^*) = \sqrt{Var^*(\hat{\hat{\gamma}}^*)}$$

After having computed the bootstrap standard deviation one can also estimate the bootstrap t-intervals by computing

$$Z^*(r) = \frac{\hat{\hat{\gamma}}^{*r} - \bar{\gamma}}{se^*(r)}$$

where $se^*(r)$ is an estimation of the standard error of $\hat{\hat{\gamma}}^{*r}$. After having computed all $Z^*(r)$, $r \in \{1, \ldots, B\}$ one has to arrange them according to the size and estimate

the quantiles $\hat{t}^{(\alpha)}$ and $\hat{t}^{(1-\alpha)}$ for a $(1-2\alpha)$ confidence interval by calculating

$$\alpha = \frac{\#\{Z^*(r) \leq \hat{t}^{(\alpha)}\}}{B}$$

where $\#\{\cdot\}$ is simply the count of $Z^*(r)$ less than or equal to $\hat{t}^{(\alpha)}$. This finally yields the bootstrap-t-interval for a $1-2\alpha$ confidence level

$$[\ \bar{\gamma} - \hat{t}^{(1-\alpha)} \cdot se^*(\hat{\bar{\gamma}}^*)\ ,\ \bar{\gamma} - \hat{t}^{(\alpha)} \cdot se^*(\hat{\bar{\gamma}}^*)\ ]$$

However, the approach introduced above is only reasonable for GLMs, i.e., for models with a linear predictor. We will refer to this model as the *bootstrap logit model* (BLM).

In the following section we will present the results obtained by fitting a BLM. We treat the occurrence of each tie y_{ij} as an independent observation. The occurrence of a tie depends on the covariates, which can be divided into endogenous covariates and conventional covariates. The endogenous covariates are going to be the same as in (50). The conventional covariates include the supplier's and receiver's GDP and CINC, the receiver's intra-state conflict score as well as the relational covariates: defense agreement, direct contiguity, polity score, and path dependency (see chapter 2 and 3.6). Just as for the ERGM we are going to include the conventional covariates with a $t-2$ time lag. The resulting parameter estimates can be found in appendix 9.2. These plots show the time series for each estimated parameter and the corresponding 95% confidence interval for the period $1952-2013$. A green node indicates that the estimate is statistically significant at the 5% level, orange indicates that the estimate is statistically significant at the 10% level, but not at the 5% level, and red indicates that the estimate is not significant at the 10% level.

The time series for the estimated intercept has a clearly negative, statistically significant effect during the entire time period, which indicates that the network is rather sparse. The time series for the reciprocal tie reveals the interesting fact that the estimates are mostly positive, meaning that the chance of a tie occurrence between actors i and j increases when there is a tie going from j to i, until the turn of the millenium. It is worthwile to mention that the existence of a tie between j and i becomes insignificant from 1999 on. This result seems to relate to the result we obtained from the defense agreement time series, where the parameter estimates also become insignificant around the turn of the millennium. As Brzoska [6] is covering in his paper, during the time of the Cold War, allied nations were trading weapons

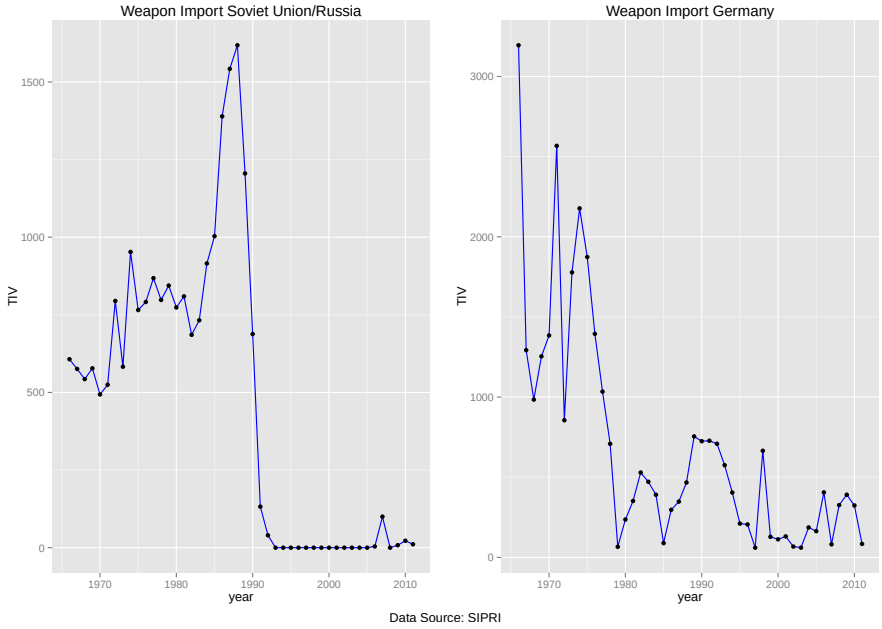


Figure 15: Weapon import trends for the Soviet Union/Russia and Germany

with each other, especially the countries involved in NATO and the Warsaw Treaty. But in the past 15 years, arms have been traded for more economic reasons. This theory is further supported by the time series of the receiver's GDP. While a recipient's GDP did not play a central role in the sale of weapons until the 1970s, the estimates turn into positive statistically significant values from the 1980s on. This supports the conclusion that, today, countries with higher GDPs per capita are more likely to purchase major conventional weapons.

The time series for the seller's in-degree provides interesting insights as well, since over time it changes from having a clearly positive effect to having a negative effect. This can be explained by the argument that the world's main weapon suppliers are not currently importing weapons at the same level as during the Cold War, but are instead focusing on distributing their own products globally. A prime example is given by figure 15 where the trends in arms imports are visualized for two of the world's main weapon suppliers: the Soviet Union/Russia and Germany. Both nations' weapons imports clearly decrease over our timeframe of examination. In Russia's case we can even observe that the country has become self-sufficient in terms of weapons supply. The seller's out-degree and the buyer's in-degree have the expected positive effects, indicating that sellers and buyers that already have high

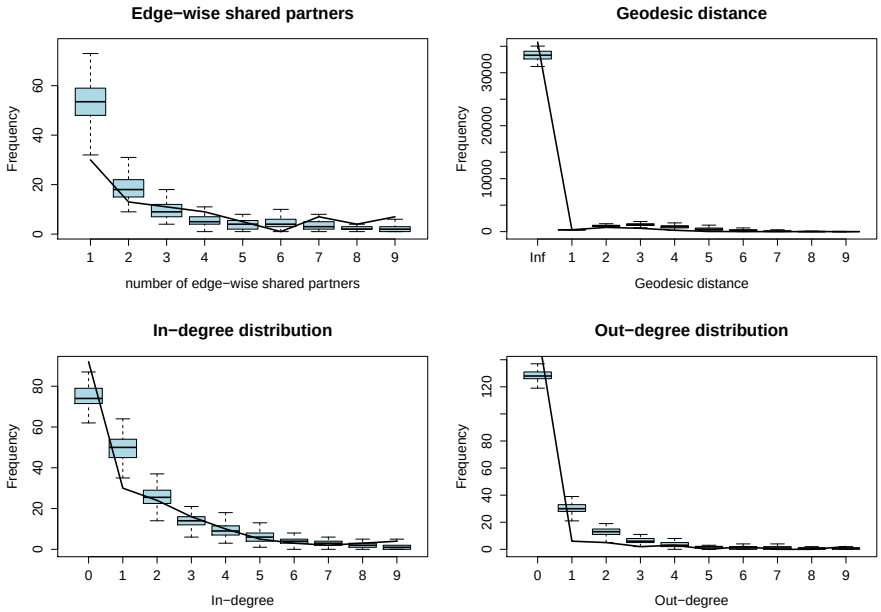


Figure 16: Goodness-of-fit of the bootstrap logit model for 2013

out/in-degrees are more likely to form ties. An interesting insight comes from the result for the buyer's out-degree, which is negative for the whole observation period. This outcome leads to the conclusion that most countries that purchase weapons are not arms suppliers themselves.

In order to be able to compare the model fit of the BLM with the ERGM, we apply the same method that was used for the ERGM to evaluate the model's fit. Figure 16 visualizes the goodness-of-fit for the bootstrap logit model for the year 2013. These plots can be interpreted just as the goodness-of-fit plots discussed in chapter 3.6. The bootstrap bias $b(\hat{\hat{\gamma}}^*)$ was approximated by $B = 100$ simulated networks and the boxplots were generated using another 100 simulated networks. When comparing figure 16 with figure 11 from chapter 3.6 we do not observe a clear improvement in any of the four hyper statistics. The networks simulated by the distribution of $\mathbb{P}_{\hat{\gamma}}$ do not describe the observed network in a better way than the ERGM fit did earlier. For models fitted for different years, we get similar results. The goodness-of-fit results are still not desirable, especially since the edge-wise shared partners and the out-degree distribution are not captured in a satisfying way. For this reason, we are going to determine the effect an actor's in- and out-degree have on formation of ties by using smooth functions. This will be done in the next chapter. The visualized

smooth functions are then going to justify the extension of the ERGM into a *curved* ERGM as we will present in chapter 7.

6.3 The Generalized Additive Model for Networks

By introducing the BLM we deviated from the specified goal set in chapter 3.6 to estimate the nonparametric effect of an actor's in- and out-degree on the occurrence of a new tie. In this chapter, we are going to fit a generalized additive model as in (48), which disregards a network's dependency structure and thus, only provides pseudo-likelihood estimates. Although the results are going to be biased, we are going to get a first impression of the estimated effects.

We assume the occurrence of a tie $Y_{ij} = 1$ to be dependent on the sender's and receiver's in- and out-degrees, as well as on the existence of the reciprocal tie Y_{ji}. This yields the model

$$\begin{aligned} &logit(\mathbb{P}(Y_{ij} = 1) = \\ &\gamma_0 + \gamma_1 Y_{ji} + s_1(\sum_{\substack{k=1 \\ k\neq j}}^{N_V} Y_{ik}) + s_2(\sum_{\substack{k=1 \\ k\neq i}}^{N_V} Y_{jk}) + s_3(\sum_{\substack{k=1 \\ k\neq j}}^{N} Y_{ki}) + s_4(\sum_{\substack{k=1 \\ k\neq i}}^{N_V} Y_{kj}) + s_x(x_{ij}^{ex}) \end{aligned} \quad (54)$$

where $s_x(x_{ij}^{ex})$ is short for the exogenous covariates $s_6(x_{ij6}) + \cdots + s_p(x_{ijp})$ and $i, j \in \{1, \ldots, N_V\}$, $i \neq j$. This model treats an actor's in- and out-degrees, as well as Y_{ji}, as covariates of a regular GAM. This implies that we are taking these covariates to be independent of each other and furthermore disregard a network's dependency structure.

Figure 17 shows the estimated smooth function for the supplier's out-degree and the receiver's in-degree for the time period 2004-2013. We used linear B-splines as discussed in chapter 4.2, penalized them with the method introduced by Eilers and Marx [14] (see chapter 4.3) and optimized the smoothing parameter λ by generalized cross validation as illustrated in chapter 4.4. The effect we observe for both, the seller's out-degree and the receiver's in-degree, are clearly non-linear, but decline for higher in- and out-degrees. For that reason, a black graph was added into these plots to visualize the log function and demonstrate that the effect is non-linear in nature. The log function for the supplier's out-degree was shifted by the constant $c = 1.5$. A steep slope is noticeable at the beginning of both visualizations, which then starts to decline incrementally. We did not visualize the effects of the supplier's in-degree and the recipient's out-degree, since a clear trend was not identifiable and

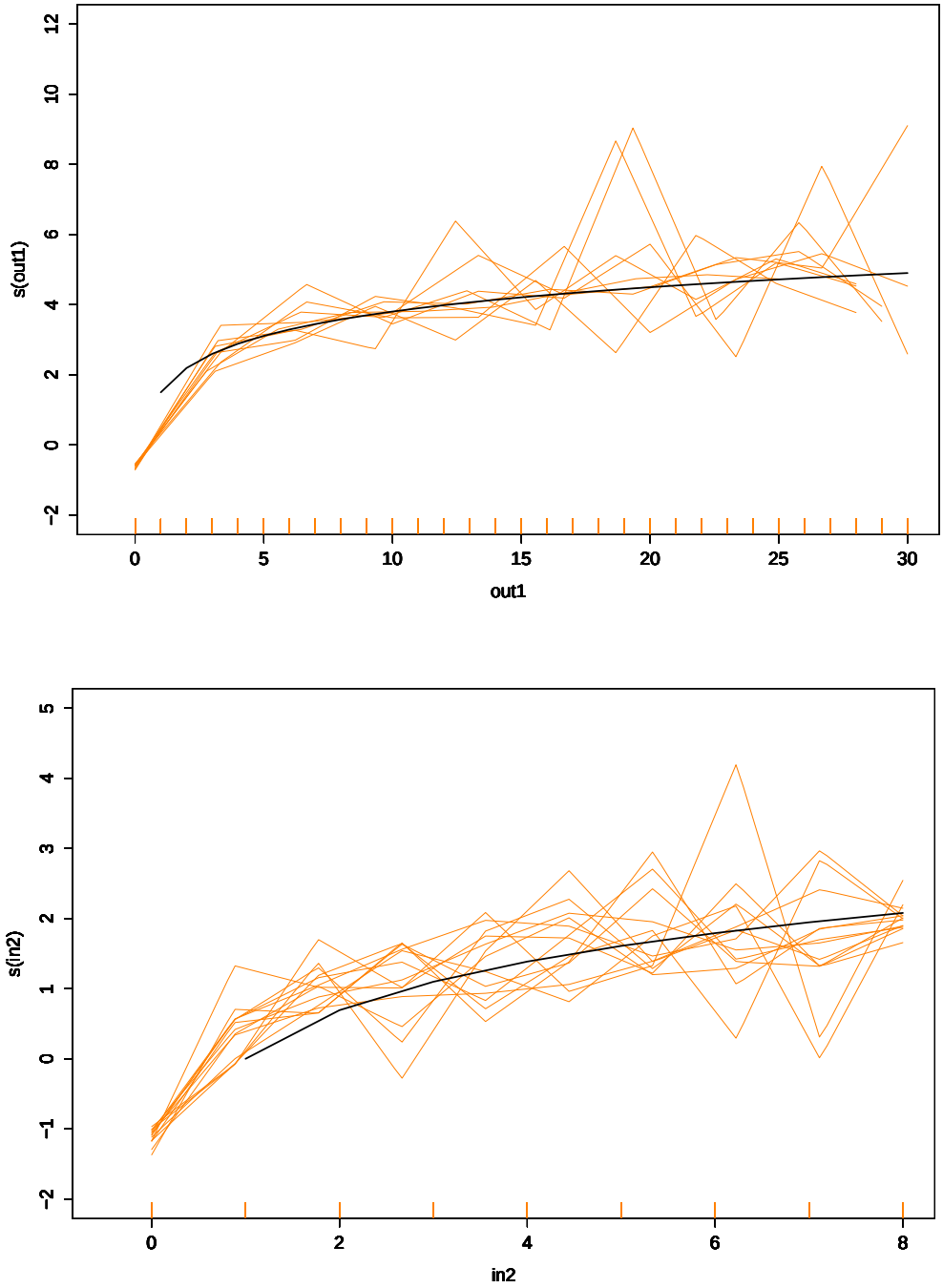


Figure 17: Penalized B-spline fit for the in- and out-degree for the years 2004-2013. The black line illustrates the logarithm (in-degree) and the logarithm + 1.5 (out-degree)

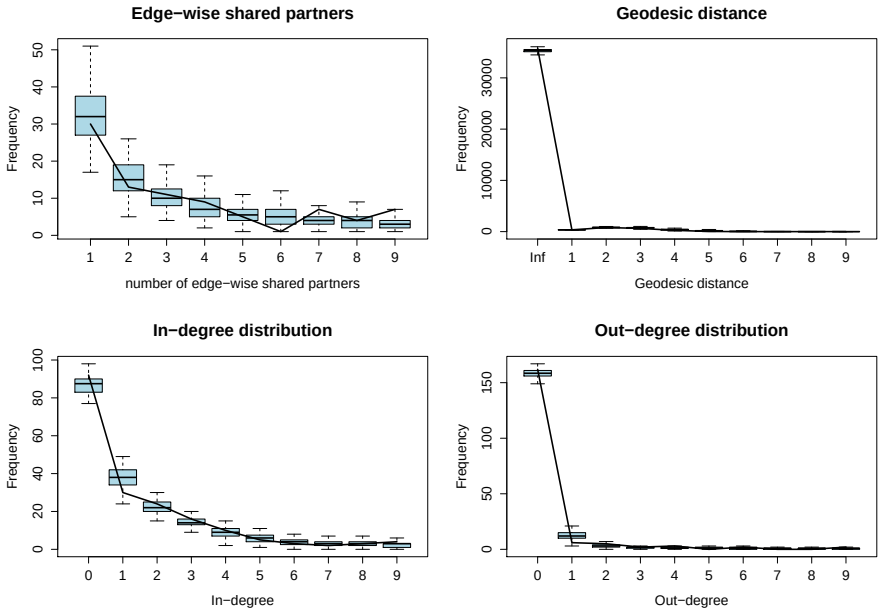


Figure 18: Goodness-of-fit for the generalized additive model of the year 2013

the results were not significant for most of the years.

Even though the coefficients were estimated by a pseudo-likelihood approach, the model fit turned out to be acceptable. In figure 18 the goodness-of-fit as described in chapter 3.6 is visualized for the year 2013. Again, we obtained similar goodness-of-fit results for other years. The results show a clear improvement over the goodness-of-fit of the ERGM simulated in chapter 3.6 and over the results of the BLM in chapter 6.2. Besides the in-degree distribution, which was already well captured in the two other models, the distributions of the three other hyper statistics improved as well. The black bold line, which represents the distribution in the observed network, passes through every single boxplot.

Certainly, we have to be very careful about drawing conclusions from the estimated smooth functions of the supplier's out-degree and the receiver's in-degree about the ERGM. The ERGM always considers the impact that the change in the occurrence of a tie has on the entire network, and by introducing the geometrically weighted degree statistics in the next chapter we will recognize that the ERGM can consider the change that the occurrence of a tie has on the entire degree distribution of a network. The fitted GAM from this chapter misses this universal view of an entire network's degree distribution since it is confined to considering the in- and out-degree of a

particular actor without taking the change in the network into account. In the GAM, an actor's in- and out-degrees are merely taken as *dyad-specific* characteristics and therefore, lack the global consideration of the ERGM. Nevertheless, the results of the smooth functions in this chapter justify the consideration to extend the conventional ERGM into a more generalized model. This generalized ERGM enables us to down-weight the contributions of high-degree nodes in a geometrically decreasing way.

From figures 17 we learn that the effect decays bit by bit for higher degrees. Hence, it would be desirable to extend the ERGMs from chapter 3.6 in a way that allows these models to capture this effect. This can be accomplished by generalizing the ERGM by so-called *curved exponential random graph models (CERGM)*, since the CERGM enables us to add specific endogenous statistics that capture the decay of a degree's effect. The decay is controlled by a *decay parameter* and can be estimated along with the regular parameters. The CERGM will be introduced in the following chapter.

7 The Curved Exponential Random Graph Models (CERGM)

In this chapter we are going to introduce the *curved exponential random graph model (CERGM)*, a generalization of the ERGM. In particular, these models involve *geometrically weighted degree*, *geometrically weighted edge-wise shared partner*, and *geometrically weighted dyad-wise shared partner* statistics. These statistics enable the inclusion of a network's degree, edge-wise shared partner, and dyad-wise shared partner distributions into the model. This chapter is mainly based on Hunter [25], Hunter and Handcock [27], and Robins et al. [42].

7.1 The CERGM

When modeling the ERGM with the endogenous statistics

$$\Gamma(A) = (\Gamma_{edges}, \Gamma_{asymmetric}, \Gamma_{idegree(1)}, \Gamma_{dsp(1)})$$

as is done in chapter 3.6, and examining the quality of the model by simulating the goodness-of-fit plots, one observes that the simulated networks reproduce the number of nodes with in-degree 1 perfectly (see figure 11). However, even though the number of nodes with an in-degree of 1 was captured sufficiently, one might also want to include the statistics $\Gamma_{idegree(0)}, \Gamma_{idegree(2)}, \Gamma_{idegree(3)}, \ldots$ as well. One even

might want to go a step further and fit the model with $\Gamma_{odegree(0)}, \Gamma_{odegree(1)}$, $\Gamma_{odegree(2)}, \ldots$ in the hope that $\mathbb{P}_{\hat{\theta}}$ becomes more similar to $\mathbb{P}_\theta$. Unfortunately, these models degenerate when these steps are attempted. Even if these models did not degenerate, however, we would still obtain a high number of estimated coefficients. It would be desirable to include a statistic into the ERGM that could refer to a network's degree distribution without causing a high number of estimated coefficients. Therefore, consider the vector $\theta \in \mathbb{R}^q$ as a function of a vector $\varrho \in \mathbb{R}^p$, where $p < q$.

This implies that we can write (1) as

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\theta(\varrho)^T \cdot \Gamma(A))}{c(\theta(\varrho))} \tag{55}$$

We are going to refer to (55) as a *curved exponential random graph model*. As was already discussed in chapter 3.2, the maximum-likelihood estimator (MLE) of the parameter vector θ is the vector which maximizes $\mathbb{P}_\theta(Y = A^{obs})$. With A^{obs}, we refer to the observed network. For this reason we might also refer to the MLE $\hat{\theta}$ as

$$\hat{\theta} = \arg\max_{\theta \in \mathbb{R}^q} \frac{\exp(\theta^T \cdot \Gamma(A))}{c(\theta)} \tag{56}$$

By assuming that θ is a function of ϱ the MLE of (55) turns out to be

$$\hat{\varrho} = \arg\max_{\varrho \in \mathbb{R}^p} \frac{\exp(\theta(\varrho)^T \cdot \Gamma(A))}{c(\theta(\varrho))} \tag{57}$$

A detailed approach for computing (57) through Fisher scoring is given by Hunter and Handcock [27]. At this point it should be mentioned that in the case of $\theta(\varrho)$ being a linear function, we could write $\theta(\varrho)$ as $M\varrho$, where $M \in \mathbb{R}^{q \times p}$. Consequently, (55) turns into

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\varrho^T M^T \cdot \Gamma(A))}{c(M\varrho))}$$

which is basically a standard ERGM. This can easily be verified by setting $\Gamma^*(A) := M^T \cdot \Gamma(A)$ and $c^*(\varrho) := c(M\varrho)$. This proves that distinguishing between CERGMs and ERGMs only makes sense when $\theta(\varrho)$ is a non-linear function of ϱ.

But how can one include a network's degree distribution as a network statistic $\Gamma_{dist}(A)$? Snijders et al. [47] introduced an approach involving k-star (or star(k)) statistics $S_1(A), \ldots, S_{N_V-1}(A)$, where $S_k(A)$ denotes the number of k-stars in the network, $k \in \{1, \ldots, N_V - 1\}$. In a directed network we have to distinguish between outstar(k) and instar(k), which can be defined with the star definition given in

chapter 3.5 as

$$S_k^o(A) := \Gamma_{ostar(k)}(A)$$

$$S_k^i(A) := \Gamma_{istar(k)}(A)$$

Note that in every network $S_1^o(A) = S_1^i(A) = \Gamma_{edges}(A)$, i.e., $S_1^o(A)$ and $S_1^i(A)$ are equal to the number of edges in the network. However, if each $\Gamma_{ostar(k)}$ has its own coefficient in the network, the resulting ERGM would look something like

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\sum_{k=1}^{n-1} \theta_k S_k^o(A))}{c(\theta)} \tag{58}$$

On this basis, Snijders introduces the *alternating k-star statistics*

$$\mathfrak{S}(A, \lambda_{out}) := \sum_{k=2}^{N_V-1} (-\frac{1}{\lambda_{out}})^{k-2} S_k^o(A) = S_2^o(A) - \frac{S_3^o(A)}{\lambda_{out}} + \cdots + (-1)^{N_V-3} \frac{S_{N_V-1}^o(A)}{\lambda_{out}^{N_V-3}}$$

and

$$\mathfrak{S}(A, \lambda_{in}) := \sum_{k=2}^{N_V-1} (-\frac{1}{\lambda_{in}})^{k-2} S_k^i(A) = S_2^i(A) - \frac{S_3^i(A)}{\lambda_{in}} + \cdots + (-1)^{N_V-3} \frac{S_{N_V-1}^i(A)}{\lambda_{in}^{N_V-3}}$$

where we refer to $\lambda_{out}, \lambda_{in} \in \mathbb{R}^+$ as the *decay parameter*. Models with these statistics and a fixed decay parameter turn out to be standard ERGMs:

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\xi \cdot \mathfrak{S}(A, \lambda_{out})}{c(\xi, \lambda_{out})} \tag{59}$$

But the question arises around how one should chose the decay parameter. If one wants to automatically estimate the λ_{out} then the model turns out to not be a standard ERGM anymore, but, rather, a CERGM. In order to clarify this, verify that model (59) is just like model (55) with

$$\theta_1 = 0 \text{ and } \theta_k \equiv \theta_k(\xi, \lambda_{out}) = \frac{(-1)^k \xi}{\lambda_{out}^{k-2}}\ ,\ 2 \leq k \leq N_V - 1$$

Hunter and Handcock [27] succeeded in proving that one can also rewrite alternating k-stars as a function of a network's degree distribution

$$\mathfrak{S}(A, \lambda_{out}) = \lambda_{out} \left(\lambda_{out} \sum_{j=1}^{N_V-1} \left(1 - \left(1 - \frac{1}{\lambda_{out}} \right)^j \right) D_j^o(A) + 2S_1^o(A) \right) \tag{60}$$

$$\mathfrak{S}(A, \lambda_{in}) = \lambda_{in}\left(\lambda_{in}\sum_{j=1}^{N_V-1}\left(1-\left(1-\frac{1}{\lambda_{in}}\right)^j\right)D_j^i(A) + 2S_1^i(A)\right) \tag{61}$$

where $D_j^o(A) := \Gamma_{odeg(j)}(A)$ and $D_j^i(A) := \Gamma_{ideg(j)}(A)$ are the number of nodes with and in- and out-degree of j, respectively. In the next step, we define the *geometrically weighted out-degree* (gwod) and the *geometrically weighted in-degree* (gwid) statistics as the first summand of (60) and (61)

$$\Gamma_{gwod}(A, \lambda_{out}) := \lambda_{out}\sum_{j=1}^{N_V-1}\left(1-\left(1-\frac{1}{\lambda_{out}}\right)^j\right)D_j^o(A) \tag{62}$$

$$\Gamma_{gwid}(A, \lambda_{in}) := \lambda_{in}\sum_{j=1}^{N_V-1}\left(1-\left(1-\frac{1}{\lambda_{in}}\right)^j\right)D_j^i(A) \tag{63}$$

At this point it also becomes obvious where the *geometrically* in the name of gwod and gwid comes from. It simply refers to the geometric sequence $(1-\frac{1}{\lambda})^j$ which appears in these statistics.

Equations (60) and (61) also demonstrate that the alternating k-star statistic is a linear combination of the geometrically weighted degree statistic and the number of edges. Therefore, it is possible to invert this equation to express the geometrically weighted degree statistic as a linear combination of the alternating k-star and the number of edges. Combined with the fact that the number of *edges* is essential in every ERGM, since it is playing the role of the intercept, this yields the result that the geometrically weighted degree statistic and the alternating k-star statistic are interchangeable when fitting a model.

But how can one interpret the parameters? For the sake of simplicity, consider the model

$$\mathbb{P}_\theta(Y = A) = \frac{\exp(\xi \cdot \Gamma_{gwod}(A, \lambda_{out})}{c(\theta(\xi, \lambda_{out})))} \tag{64}$$

Adding one single edge to the network changes the out-degree distribution of the network so that one actor with an out-degree of k turns into an actor with an out-degree of $k+1$. With our notation this means that D_k^o and D_{k+1}^o get replaced by $D_k^o - 1$ and $D_{k+1}^o + 1$, while no changes for D_ℓ^o , $\ell \in \{0, \ldots, N_V - 1\} \setminus \{k, k+1\}$ are made. This changes the probability of a graph in the following way (for reasons of

clarity the fraction has already been reduced)

$$\begin{aligned}\frac{\mathbb{P}_\theta(Y_{ij}=1|Y_{ij}^c=A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij}=0|Y_{ij}^c=A_{ij}^c)} &= \frac{\exp(\xi\lambda_{out}((D_k^o-1)(1-\tau^k)+(D_{k+1}^o+1)(1-\tau^{k+1})))}{\exp(\xi\lambda_{out}((D_k^o)(1-\tau^k)+(D_{k+1}^o)(1-\tau^{k+1})))} \\ &= \exp(\xi\lambda_{out}((1-\tau^{k+1})-(1-\tau^k))) \\ &= \exp(\xi\lambda_{out}(\tau^k-\tau^{k+1})) \\ &= \exp(\xi\lambda_{out}((\tau-1)\tau^k)) \\ &= \exp(\xi\tau^k)\end{aligned}$$

where $\tau = (1-\frac{1}{\lambda_{out}})$. This implies

$$\frac{\mathbb{P}_\theta(Y_{ij}=1|Y_{ij}^c=A_{ij}^c)}{\mathbb{P}_\theta(Y_{ij}=0|Y_{ij}^c=A_{ij}^c)} = \exp(\xi(1-\frac{1}{\lambda_{out}})^k) \tag{65}$$

What does this mean for the interpretation? Recall that $\lambda_{out} \in \mathbb{R}^+$. When taking a look at $(1-\frac{1}{\lambda_{out}})^k$ with $\lambda_{out} > 1$, one can observe that by adding a new edge into the network the benefit decreases geometrically by the degree of the involved nodes. This seems to be the perfect statistic for including the decreasing effect we observed for the GAM's out-degree in figure 17. Furthermore, we can perceive that the higher the decay parameter, the slower the decay. Therefore, the geometric decay is controlled by λ_{out} and this is the reason why we refer to λ_{out} as the *decay parameter*. The interpretation of parameter ξ is very similar to the interpretation of parameter θ in a standard ERGM (1). As a consequence,

- $\xi > 0$ implies a preference for adding an edge
- $\xi = 0$ implies no preference
- $\xi < 0$ implies a preference for deleting an edge

This implies that with $\xi > 0$ the model prefers networks containing nodes with high out-degrees, while $\xi < 0$ indicates the opposite. Putting both parameters' interpretation together, we can observe that for a very large parameter λ_{out} the preference of adding an edge to the network does not decrease much, since $(1-\frac{1}{\lambda_{out}})^k \approx 1$ and consequently remains almost constant at ξ. The other extreme case $\lambda_{out} = 1$ yields $\exp(\xi(1-\frac{1}{\lambda_{out}})^k) = 1$, which implies no preference, regardless of the value of ξ. Interpreting the parameters for $\lambda_{out} \in (0,1)$ turns out to be difficult, since the value of $(1-\frac{1}{\lambda_{out}})^k$ starts alternating.

Aside from the geometrically weighted degree distributions, we are going to present

two more CERGM statistics, which were also introduced by Hunter [25]: the *geometrically weighted dyad-wise shared partners* (gwdsp) and the *geometrically weighted edge-wise shared partners* (gwesp). For this reason, we denote $DP_k(A) := \Gamma_{dsp(k)}(A)$ and $EP_k(A) := \Gamma_{esp(k)}(A)$ in an analogously to how we defined the degree distributions. Given this, we can define Γ_{gwdsp} and Γ_{gwesp} as

$$\Gamma_{gwdsp}(A, \lambda_{dsp}) := \lambda_{dsp} \sum_{j=1}^{N_V-1} \left(1-\left(1-\frac{1}{\lambda_{dsp}}\right)^j\right) DP_j(A) \tag{66}$$

$$\Gamma_{gwesp}(A, \lambda_{esp}) := \lambda_{esp} \sum_{j=1}^{N_V-1} \left(1-\left(1-\frac{1}{\lambda_{esp}}\right)^j\right) EP_j(A) \tag{67}$$

Similar to the geometrically weighted degree distributions, these statistics include a geometric sequence in their definition. Since the functional forms of $\Gamma_{gwdsp}(A, \lambda_{dsp})$ and $\Gamma_{gwesp}(A, \lambda_{esp})$ are the same as those for $\Gamma_{gwod}(A, \lambda_{out})$ and $\Gamma_{gwid}(A, \lambda_{in})$, they can be interpreted similarily. Incidentally, just as gwod and gwid interrelate with the alternating k-star statistics, one can show that gwdsp and gwesp interrelate with the *alternating k-triangle* and the *alternating k-twopath* statistics, which were also introduced by Snijders et al. [47].

7.2 Results for the CERGM

After having justified the generalization from a conventional ERGM to a curved ERGM by visualizing the non-linear effect of the in- and out-degree in chapter 6.3, and after having introduced the CERGM together with the related geometrically weighted statistics in chapter 7.1, we are finally set to fit a CERGM to the arms trade data and present the results.

Including the geometrically weighted statistics in the model causes a stepwise down-weighted effect. At first glance, it seems to be reasonable to include all four statistics, Γ_{gwod}, Γ_{gwid}, Γ_{gwesp} and Γ_{gwdsp}, into the network in order to guarantee a decent model fit. However, it turned out that including all four or any combination of three or two of these statistics into our arms trade model caused degeneracy. Luckily, including just one geometrically weighted statistic resulted in non-degenerated models for the majority of cases. We therefore decided to include the geometrically weighted out-degree statistic, since we anticipated being able to capture the networks' out-degree distribution in a more reasonable way. According to the goodness-of-fit plots in figure 11, the in-degree distribution was appropriately captured by including $\Gamma_{idegree(1)}$, but adding any statistic $\Gamma_{odegree(k)}, k \in \mathbb{N}$ caused degeneracy. By

adequately including a model's in- and out-degree distribution, it is furthermore reasonable to expect an improvement of the geodesic distance distribution and even of the edge-wise shared partners distribution.

By extending the ERGM into a CERGM by including Γ_{gwod}, the model appeared to be more robust against degeneracy. Statistics which initially could not be added into the ERGM, could now be put into the model without causing degenerated models. We therefore decided to incorporate the statistics $\Gamma_{dsp(0)}$ and $\Gamma_{esp(0)}$. With the statistic $\Gamma_{dsp(0)}$ we intended to capture the fact that the majority of actors in the network are not connected by a directed two-path through a third actor. We justify this on the basis of the geodesic distance distribution of the networks (see figures 11, 16 and 18), which indicate that there is no directed path between most actors. The reason for this is that most actors do not sell weapons and therefore have an out-degree of zero (see figure 5). Consequentely, a directed two-path cannot originate from these actors. With the statistic $\Gamma_{esp(0)}$ we emphazise the direct trades between two actors, since this statistic counts the number of pairs (i, j) which are directely connected and do not close deals via a third party. Γ_{edges} is the final endogenous statistics included in our model. The included covariates in our model are the same as those in the fitted ERGM in chapter 3.6. Just as in the ERGM, we include these covariates with a $t - 2$ time lag.

Figure 19 shows the MCMC diagnostics of the fitted model for the year 2013. As already discussed in chapter 3.6, the plots on the left side visualize the values for every included statistic obtained via MCMC-simulated networks. These values are centered around the statistic's value for the observed network. The right side shows the empirical density function for each case.
The MCMC diagnostics show good results, since every single empirical density function is centered around the value of the observed statistic and the trace plots on the left side do not display questionable dependency structures. This also holds true for the MCMC diagnostics of years other than 2013.

So the model does not degenerate, but does it also provide a good fit? In order to answer this question and in order to compare the fitted CERGM with the fitted ERGM of chapter 3.6 and fitted BLM of chapter 6.2, we take a look at the goodness-of-fit diagnosis in figure 20. Compared to the goodness-of-fit of the ERGM (figure 11) and the goodness-of-fit of the BLM (figure 16), we observe a clear improvement. Besides the in-degree distribution, which was already well-captured in figures 11 and 16, the new model also covers the distribution of the remaining three hyper-

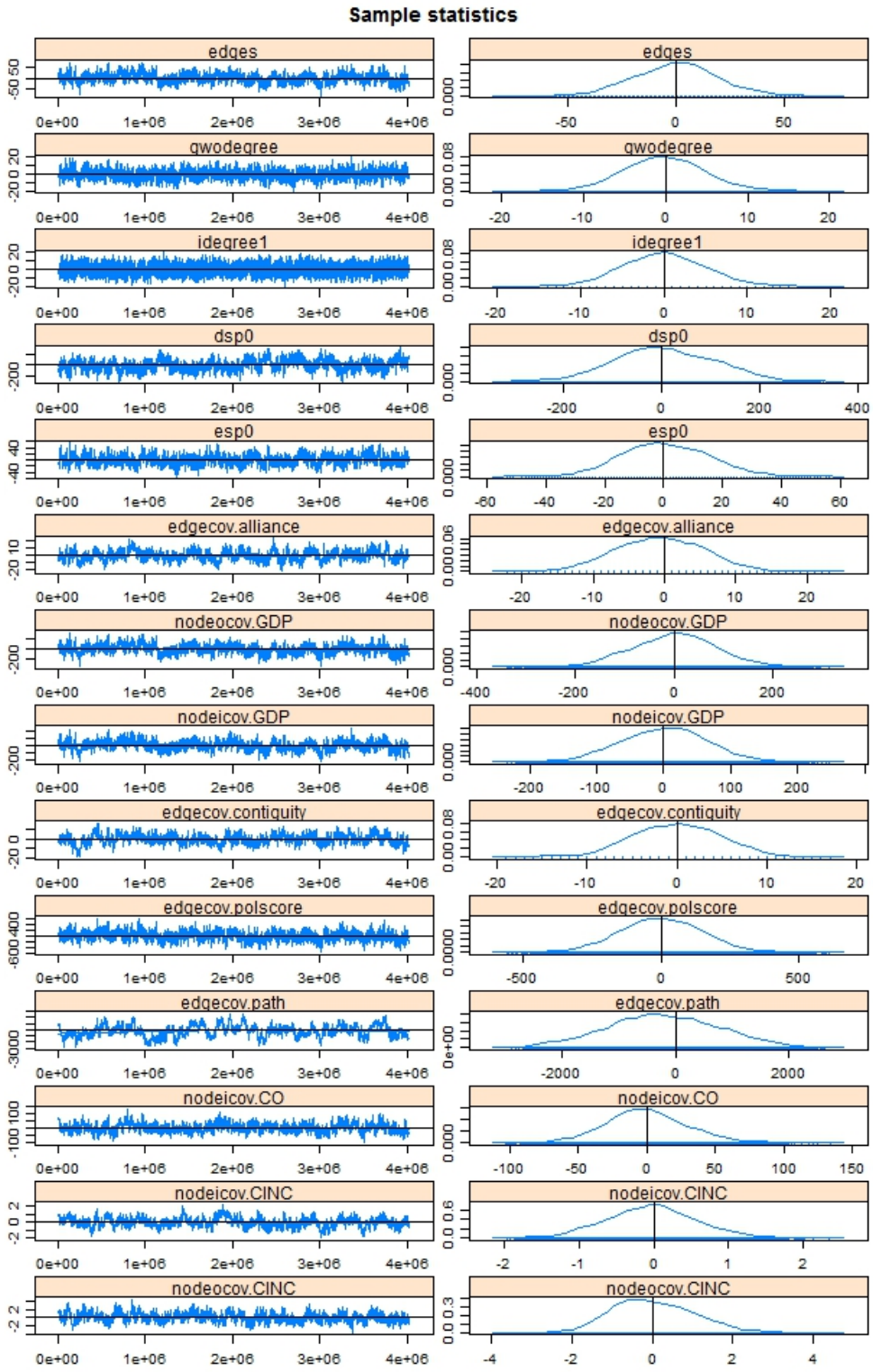


Figure 19: MCMC diagnostics for the CERGM of 2013

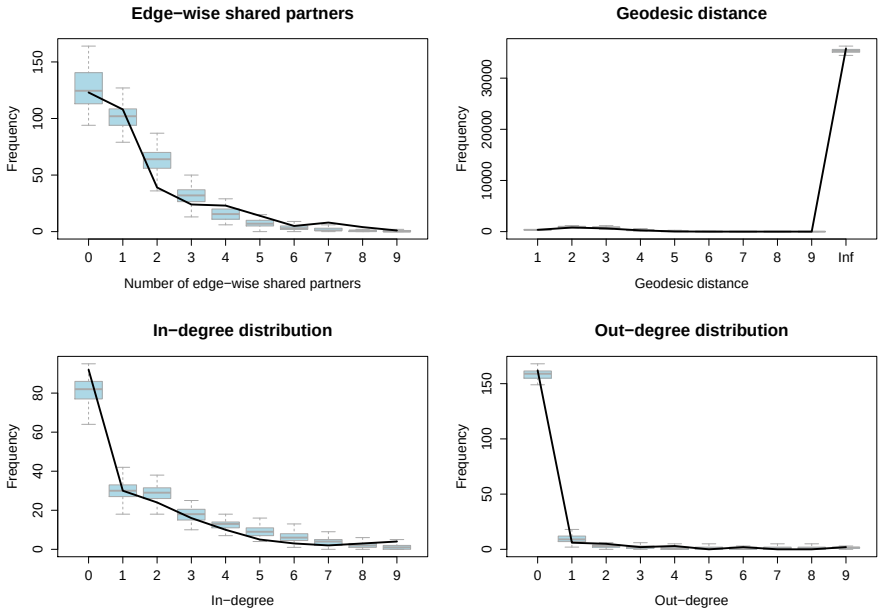


Figure 20: Goodness-of-fit diagnosis for the fitted CERGM for 2013

statistics in a satisfying way. It is remarkable that with the integration of Γ_{gwod} the out-degree's goodness-of-fit improved markedly. Since the same holds true for the geodesic distribution, it looks like our assumption proves to be true: By fitting a model that adequately captures a network's in- and out-degree distribution, the geodesic distance distribution improves automatically.
Finally, the edge-wise shared partners distribution of the simulated networks also improved. Certainly, this is thanks to the fact that we were able to include $\Gamma_{esp(0)}$ into the CERGM.

The results for the parameter estimates for every fitted ERGM from 1952 until 2013 can be seen in figures 21 and 22. The first ERGM was fitted for 1952 and not for 1950, since we include the exogenous variables with a two year time lag. The parameter estimates are visualized with 95% confidence intervals and a color index indicates the significance level of the variables. A green node indicates that the included endogenous statistic or exogenous covariate is statistically significant at the 5% level, orange indicates that the variable is statistically significant at the 10% level, but not at the 5% level, and a red node indicates that a statistically significant relation could not be determined at a 10% level.

The time series for Γ_{edges} has the expected highly negative effect in every year's net-

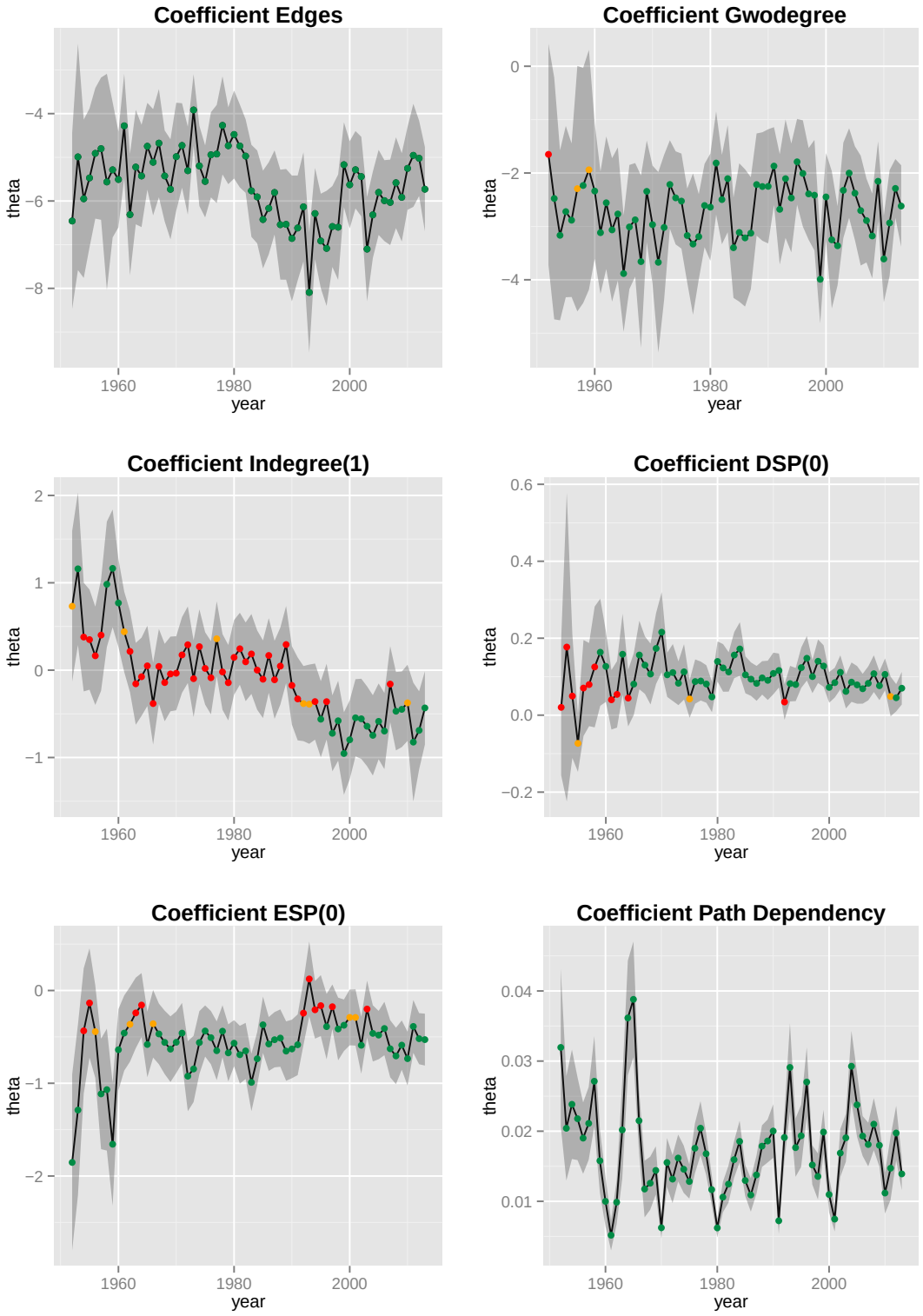


Figure 21: Time series of the estimated parameters for the time period 1952-2013

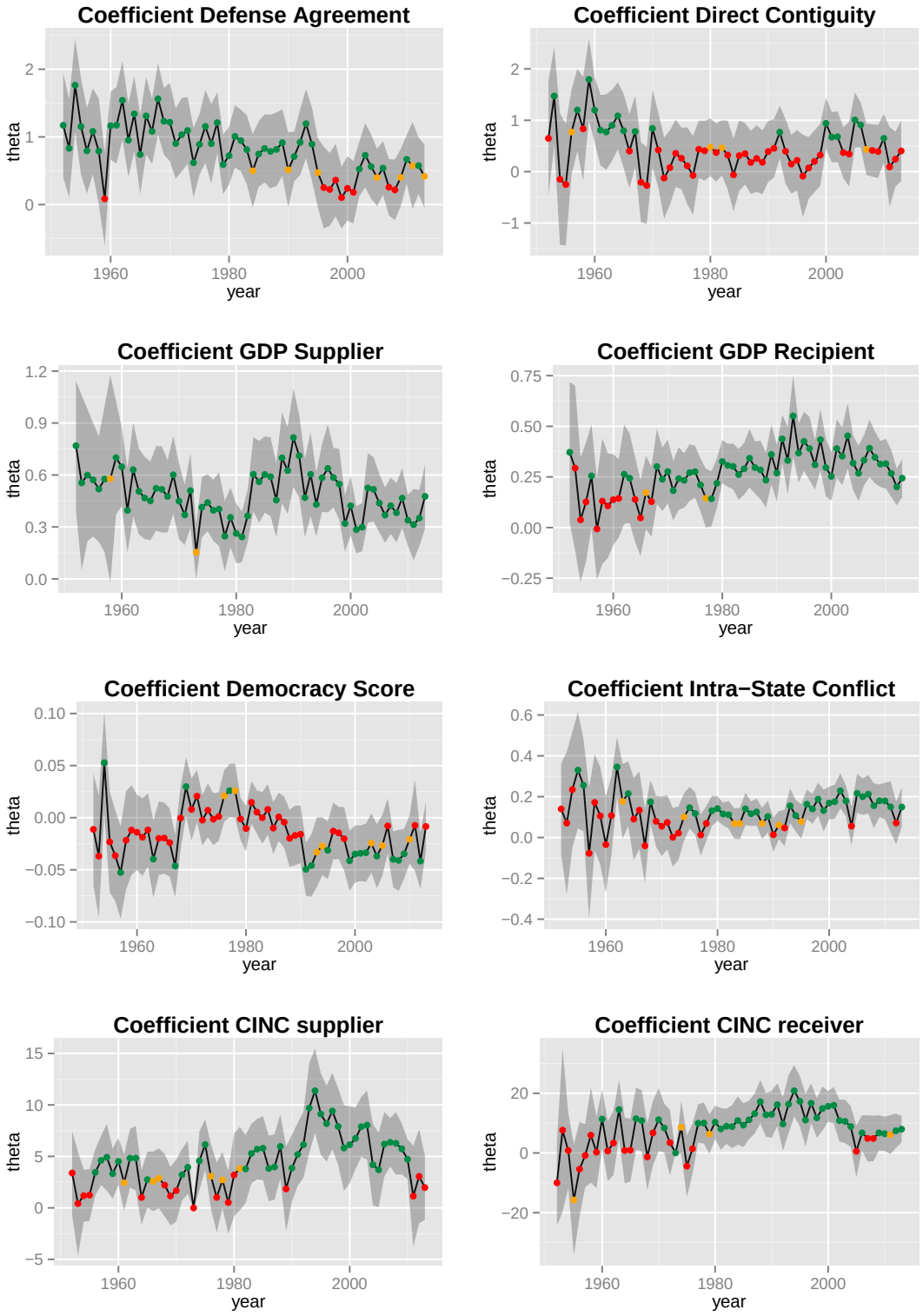


Figure 22: Time series of the estimated parameters for the time period 1952-2013

work, indicating that the observed networks are all rather sparse. An interpretation of Γ_{edges} on the edge level is not possible, since networks where the change-statistic for the number of edges differs by one do not exist. The change-statistic for the number of edges is equal to 1 in every network. Therefore, the only way left is to interpret Γ_{edges} on the network level. For two networks A and A^{edges-}, network A^{edges-}, which has one edge less than network A while all the other statistics are equal, is more plausible. Recall that A^{k^-} is defined as a network where all statistics except the kth have the same value as in network A, but the kth statistic of A^{k^-} is one smaller than that in A. Roughly speaking, our model is tending to sparse networks, i.e., models with less ties.

With the exception of the early years, we observe a mostly statistically significant positive effect for the dyad-wise 0-shared partner statistic. Unlike Γ_{edges}, we can interpret $\Gamma_{dsp(0)}$ on the edge level. Therefore, consider two networks A and B, which are both completely known, except for edge e_{ij}, and where the change-statistic

$$(\Delta_{dsp(0)}A)_{ij} := \Gamma_{dsp(0)}(A^+_{ij}) - \Gamma_{dsp(0)}(A^-_{ij})$$

of A is one higher than the change-statistic

$$(\Delta_{dsp(0)}B)_{ij} := \Gamma_{dsp(0)}(B^+_{ij}) - \Gamma_{dsp(0)}(B^-_{ij})$$

of B, while all other change-statistics of A and B are identical. Recall that A^+_{ij} emerges from A, while assuming $a_{ij} = 1$, and that A^-_{ij} emerges from A, while assuming $a_{ij} = 0$. It thus follows that edge e_{ij} is more likely to occur in network A, i.e., in the network with more dyad-wise 0-shared partners. Just as for Γ_{edges}, $\Gamma_{dsp(0)}$ can also be interpreted on the network level. Given two networks A and $A^{dsp(0)-}$, network $A^{dsp(0)-}$ is less plausible than network A. The statistics $\Gamma_{idegree(1)}$ and $\Gamma_{esp(0)}$ can be interpreted similarly. However, notice for the $\Gamma_{idegree(1)}$ results that the estimated parameters change from having a positive effect in the 1950s, to having no effect in the 1960s through the 1980s, to having a negative effect in the more recent years. This indicates that for the earlier years, networks with more in-degree-1-actors are more plausible, but for more recent years the models with less actors with an in-degree of 1 are more plausible.

While the decay parameter estimates for the geometrically weighted out-degree statistic can be found in figure 23, the regular parameter estimates for this statistic are plotted next to all other results in figure 21. The decay paramter estimates are greater than 1 througout the entire time period with an average value of 3.05. This

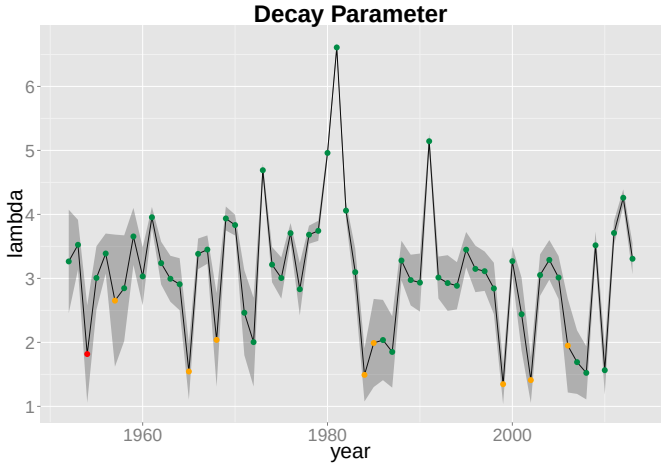


Figure 23: Time series of the decay parameter for the time period 1952-2013

implies a fairly fast geometrical decrease, which aligns with the results in figure 17 from chapter 6.3. The regular parameter estimates are the multipliers for this effect and turn out to be negative for the whole time period, which emphazises once more that our models tend towards sparse networks.

As we have already pointed out in chapter 3.5, a network that merely includes endogenous statistics can not distinguish between structurally equivalent networks. However, through the positive results of the exogenous variable *defense agreement*, the model assigns more probability mass to a network, which has more conformity with the defense agreement network. Consequently, the results of the parameter time series for the defense agreement estimates reveal very interesting insights. Just as in the parameter time series of the BLM (see appendix 9.2) we observe a positive, but clearly decreasing effect. Even though we have to be careful with the interpretation of trends in the time series, since the number of edges increases over time, we can see that the estimates start becoming statistically insignificant around the turn of the millennium. This bolsters the theory that the existence of alliances has played an increasingly minor role in countries' decisions to engage in arms trading. Brzoska [6] discusses that back in the 1960s and 1970s weapons were sold mainly to allies in order to bolster desired power dynamics and further personal political interests, while nowadays economic factors play a much more decisive role. The fitted CERGM and the BLM further support this theory.

The results for the direct contiguity data turn out to be statistically insignificant

at the 10% level for most of our examined timeframe. However, for some scattered years we obtain positive statistically significant results, which might be counterintuitive at first glance since it means that countries which share a common border are more likely to trade weapons with each other. Likely reasoning for this is that the NATO countries in Europe supply each other with military goods.

We included the democracy score in our analysis by generating a matrix with the absolute difference of the corresponding countries' democracy scores as entries. In this analysis, a negative parameter would indicate that countries with similar democracy scores are more likely to trade weapons than countries with highly dissimilar scores. On the other hand, a positive parameter would indcate that countries with dissimilar scores are more likely to trade. Our results do not allow either of these conclusions, since they oscillate around zero with occasional statistically significant results on both sides. After 1990, however, the model tends to prefer ties between countries with similar democracy scores.

The time series of the intra-state conflict estimates goes from being highly variable in the early portion of our examined timeframe to demonstrating more consistent estimates from the 1980s on. From this decade on, we obtained, with some exceptions, statistically significant estimates that were positive, which implies that countries characterized by political disturbances are indeed more likely to purchase weapons.

The supplier's GDP time series shows the expected highly positive effects, which are statistically significant througout the entire time period. Similar results are achieved for the BLM, as one can verify in appendix 9.2. This result indicates that countries with a high GDP per capita are more likely to be the tail of a tie, i.e., the supplier of arms. The results for the receiver's GDP time series reveal more interesting insights, since it seems that a country's GDP was not a major factor driving arms purchases in the 1950s and 1960s, but starts playing a crucial role from the 1970s on.

The results for a nation's CINC reveals outcomes similar to those for GDP. Just as for the receiver's GDP time series, the receiver's CINC time series goes from being largely statistically insignificant in the early portion of our examined timeframe to demonstrating statistically significant and positive estimates from 1978 on. Both the receiver's GDP and CINC estimates bolster Brzoska's [6] theory that economic factors are increasingly influential drivers in the global armament market.

8 Summary and Outlook

In this paper, we investigate the arms trade data of major conventional weapons that was provided by SIPRI. After a short introduction of network analysis in chapter 1, we introduced the arms trade data in chapter 2 along with data about the included covariates. Furthermore, we present descriptive results of the data, which provide the basis for the included endogenous statistics in chapter 3.6. However, before being able to fit the networks, the corresponding model, the exponential random graph model, and some crucial network statistics had to be introduced. By looking at the goodness-of-fit of the first model, we recognize that the underlying model provides an insufficient fit. Since the out-degree distribution and the edge-wise shared partners distribution particularly were captured in an inadequate way, we generalize the ERGM in chapter 7 and introduce the CERGM. The new results presented in chapter 7.2 are more satisfying and reveal some interesting insights.

In order to justify the generalization step from ERGM to CERGM, we model the networks with GAMs. Even though this model yields biased results, since it ignores the dependency structures inside the networks, we get approximate insight into a degree's impact in the model. The estimated smooth functions for a supplier's out-degree and a recipient's in-degree visualize the steadily decreasing effect, which can be adequately incorporate into the CERGM by geometrically weighted degree statistics. In the course of discussing different approaches to modelling networks with GAMs, we also introduce a bootstrapping approach in chapter 6.2 for fitting networks with GLMs.

In this thesis we only consider binary and stationary models. In all probability, the model could be improved by ERGMs which do consider temporal dependencies (TERGM). Hanneke et al. [20] propose a dynamic model, which allows network structures to change over time. This network has been extended by Desmarais and Cranmer [11]. Another limitation of the ERGM is that it has been applied to binary relations only. This might be acceptable for networks where a relation between two actors is either present or absent, but for valued networks such as the arms trade network, this is a serious limitation since we had to dichotomize the data. The model does not distinguish between significant arms transactions, such as the shipment of sixteen F-16 aircrafts from the United States to Italy, and comparatively negligible trade agreements such as the supply of a single armoured vehicle from Indonesia to Pakistan. Therefore, the ERGM loses some important information and provides biased results.

9 Appendix

9.1 Comments on the Electronic Appendix

In this chapter, we explain the most important functions and codes written for this paper. All calculations and visualizations presented in this paper were produced with R [41]. The primary packages that were used are *igraph, statnet, xergm, network, mgcv, ggplot2, reshape* and *gridExtra*.

- **natnum** This code loads the list of all actors (see chapter 9.3) in alphabetical order. Furthermore, it creates a column for a country's ID assigned by the Correlates of War project (COW) and a column for the IDs assigned by CEPII. With this list the data of each covariate can be assigned to the corresponding nation.

- **matrix_of_existing_countries** This code creates a matrix EX, where the columns refer to the years from 1950 to 2013 and where each row refers to one specific country. $e_{ij} = 1$ indicates that country i did exsist in year j, while $e_{ij} = 0$ indicate that country i did not exist in year j.

- **amk** This code generates a list of weighted adjacency matrices, one for each year from 1950 until 2013 and based on the alphabetical order of natnum.

The covariate data sets are rearranged into the same order as the adjacency matrices in amk. Relational covariates are put into a list of matrices, just as the observed networks in amk, while nodal attributes are rearranged into vectors of the amk order. The R-files which transform the covariates into the requested forms are *alliance cow, conflict, distance, GDP per capita, CINC, path dependency* and *polity iv*.

- **amallr(year, mod, tiv)** With the use of the matrix EX this function cuts out all actors that did not exist in the corresponding year, meaning that each year's adjacency matrix only displays the countries that actually existed at this point. In list 9.3 we indicate the time range within which each country is incorporated into the models. No entry was made for countries that existed for the entire time period 1950-2013. With the *mod*-parameter one controls the return of the function. $mod = 1$ returns the adjacency matrix of the observed network. $mod = 2$: defense agreement, $mod = 3$: direct contiguity, $mod = 4$: embargo, $mod = 5$: GDP, $mod = 9$: polity iv, $mod = 10$: CINC, $mod = 11$: path dependency, $mod = 12$: distance, $mod = 13$: intra-state conflict, $mod = 14$: inter-state conflict. The *tiv* paramter serves as a threshold for the adjacency matrix.

- **ergm out** This code computes CERGMs as described and visualized in chapter 7.2. For a given time period between $1950 - 2013$, CERGMs are calculated and the resulting parameter estimates, as well as the corresponding standard errors and p-values, are saved into a matrix. The plots were generated with the help of these matrices. The plotting codes are located at the very end of this R-file.

- **bootstrap logit out** Similar to the *ergm out*-file, this file contains the code for the bootstrap logit model as described in chapter 6.2 and visualized in chapter 9.2. For a given time period between $1950 - 2013$, BLMs are calculated and the resulting parameter estimates, as well as the corresponding standard errors and p-values, are saved into a matrix. The plots were generated with the help of these matrices. The plotting codes are located at the very end of this R-file.

- **gam splines** This code computes the GAMs as described in chapter 6.3 and plots, for a given time period between $1950 - 2013$, the estimated penalized B-splines for a supplier's out-degree and a recipient's in-degree as visualized in figure 17.

Besides the codes presented in this chapter, the electronic appendix includes a number of further codes, which were written for the remaining figures in this paper. We labeled these codes with informative names so that each figure could be matched easily to its corresponding code.

9.2 Results for the BLM

In the following section, the results of the BLM as described in chapter 6.2 are visualized. Each estimated parameter is plotted with the corresponding 95% confidence interval. The color of the nodes indicates the significance level of the corresponding estimate. A green node indicates that the estimate is statistically significant at the 5% level, orange indicates that the estimate is statistically significant at the 10% level, but not at the 5% level and finally red indicates that the estimate is not significant at the 10% level.

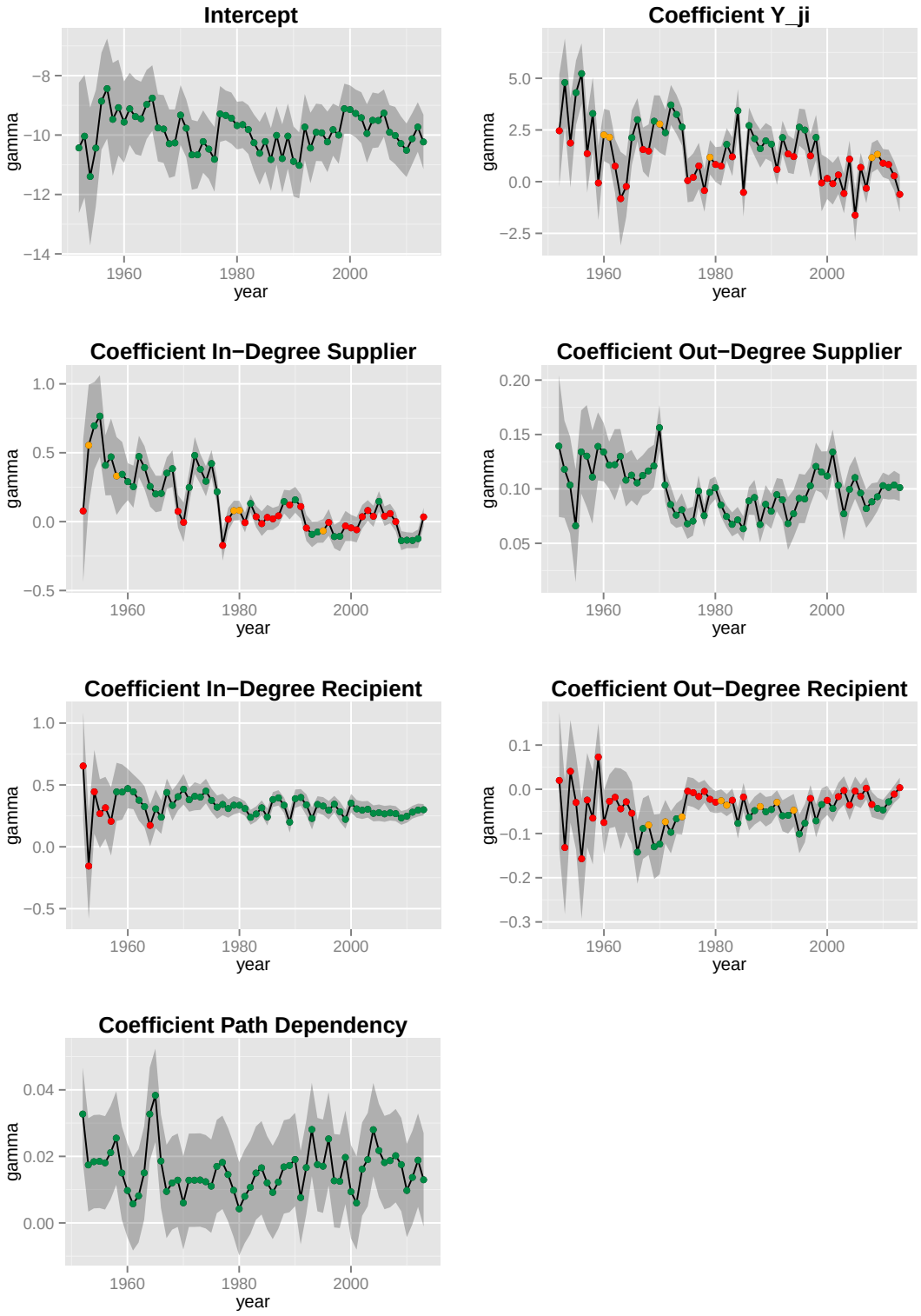


Figure 24: Time series of the estimated parameters for the time period 1952-2013

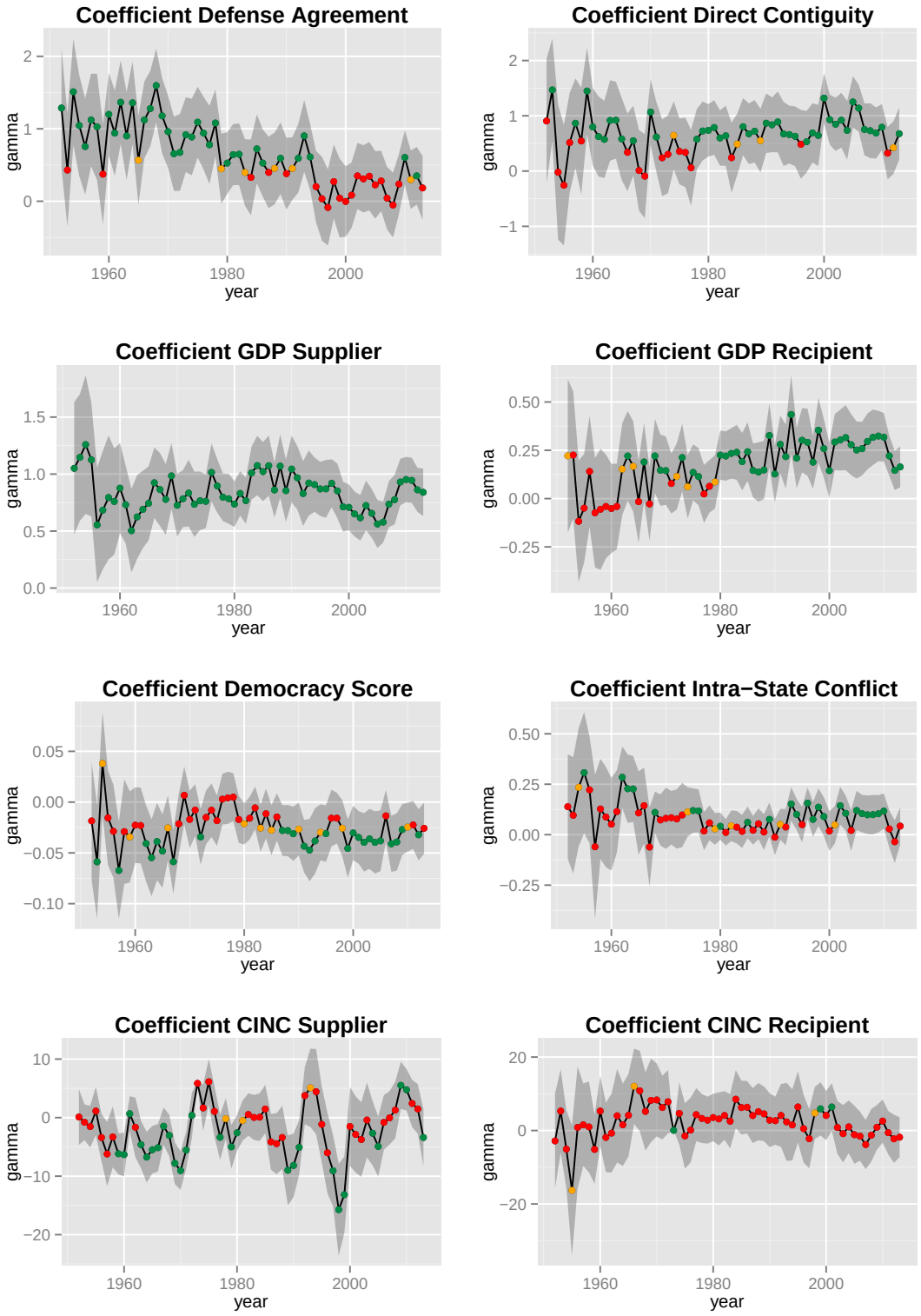


Figure 25: Time series of the estimated parameters for the time period 1952-2013

9.3 List of all Actors

In the following table, all countries and areas for which the MCW-data was gathered by SIPRI are listed. The IDs correspond with the IDs used in the R-codes. The entry in the 'Years' column indicates the time period within which the corresponding country is included into the networks. A blank entry in this column denotes that this country existed during the whole time period of interest ($1950-2013$) and hence, is included in every network.

ID	Country	Years	ID	Country	Years
1	Abkhazia	since 1992	31	Burundi	since 1962
2	Afghanistan		32	Cambodia	since 1953
3	Albania		33	Cameroon	since 1960
4	Algeria	since 1962	34	Canada	
5	Andorra		35	Cape Verde	since 1975
6	Angola	since 1975	36	Central African Republic	since 1960
7	Antigua and Barbuda	since 1981	37	Chad	since 1960
8	Argentina		38	Chile	
9	Armenia	since 1991	39	China	
10	Aruba	since 1986	40	Colombia	
11	Australia		41	Comoros	since 1975
12	Austria		42	Congo, Democratic Repubic of	since 1960
13	Azerbaijan	since 1991	43	Congo, Republic of	since 1960
14	Bahamas, the	since 1973	44	Cook Islands	since 1965
15	Bahrain	since 1971	45	Costa Rica	
16	Bangladesh	since 1971	46	Cote d'Ivoire	since 1960
17	Barbados	since 1966	47	Croatia	since 1991
18	Belarus	since 1991	48	Cuba	
19	Belgium		49	Cyprus	since 1960
20	Belize	since 1981	50	Cyprus, Northern	since 1983
21	Benin	since 1961	51	Czech Republic	since 1993
22	Bhutan		52	Czechosloviakia	until 1992
23	Biafra	1967-1970	53	Darfur	since 2003
24	Bolivia		54	Denmark	
25	Bosnia and Herzegovina	since 1992	55	Djibouti	since 1977
26	Botswana	since 1966	56	Dominica	since 1978
27	Brazil		57	Dominican Republic	
28	Brunei Darussalam		58	Ecuador	
29	Bulgaria		59	Egypt	
30	Burkina Faso	since 1960	60	El Salvador	

ID	Country	Years	ID	Country	Years
61	Equatorial Guinea	since 1968	96	Kenya	since 1963
62	Eritrea	since 1993	97	Kiribati	since 1979
63	Estonia	since 1991	98	Korea, North	
64	Ethiopia		99	Korea, South	
65	Fiji	since 1970	100	Kosovo	since 2008
66	Finland		101	Kuwait	since 1961
67	France		102	Kyrgyzstan	since 1991
68	Gabon	since 1960	103	Laos	
69	Gambia	since 1965	104	Latvia	since 1991
70	Georgia	since 1991	105	Lebanon	
71	German Democratic Republic	1949-1990	106	Lesotho	since 1966
72	Germany		107	Liberia	
73	Ghana	since 1957	108	Libya	since 1951
74	Greece		109	Liechtenstein	
75	Grenada	since 1974	110	Lithuania	since 1990
76	Guatemala		111	Luxembourg	
77	Guinea	since 1958	112	Macedonia, FYROM	since 1991
78	Guinea-Bissau	since 1973	113	Madagasacar	since 1960
79	Guyana	since 1966	114	Malawi	since 1964
80	Haiti		115	Malaysia	since 1957
81	Honduras		116	Maldives	since 1965
82	Hungary		117	Mali	since 1960
83	Iceland		118	Malta	since 1964
84	India		119	Marshall Islands, the	since 1986
85	Indonesia		120	Mauritania	since 1960
86	Iran		121	Mauritius	since 1968
87	Iraq		122	Mexico	
88	Ireland		123	Micronesia	since 1986
89	Israel		124	Moldova	since 1991
90	Italy		125	Monaco	
91	Jamaica	since 1962	126	Mongolia	
92	Japan		127	Montenegro	since 2006
93	Jordan		128	Morocco	since 1956
94	Katanga		129	Mozambique	since 1975
95	Kazakhstan	since 1991	130	Myanmar	

ID	Country	Years	ID	Country	Years
131	Namibia		166	Sierra Leone	since 1961
132	Nauru	since 1968	167	Singapore	since 1965
133	Nepal		168	Slovakia	since 1993
134	Netherlands		169	Slovenia	since 1991
135	New Zealand		170	Solomon Islands	since 1978
136	Nicaragua		171	Somalia	since 1960
137	Niger	since 1960	172	Somaliland	since 1991
138	Nigeria	since 1960	173	South Africa	
139	Niue	since 1974	174	South Ossetia	since 1990
140	Norway		175	South Sudan	since 2005
141	Oman		176	Soviet Union	until 1991
142	Pakistan		177	Spain	
143	Palau	since 1994	178	Sri Lanka	
144	Palestine	since 1988	179	Sudan	since 1956
145	Panama		180	Suriname	since 1975
146	Papua New Guinea	seit 1975	181	Swaziland	since 1968
147	Paraguay		182	Sweden	
148	Peru		183	Switzerland	
149	Philippines, the		184	Syria	
150	Poland		185	Taiwan	
151	Portugal		186	Tajikistan	since 1991
152	Qatar		187	Tanzania	since 1961
153	Romania		188	Thailand	
154	Russia	since 1992	189	Timor-Leste	since 2002
155	Rwanda	since 1962	190	Togo	since 1960
156	Saint Kitts and Nevis	since 1983	191	Tonga	since 1970
157	Saint Lucia	since 1979	192	Trans-Dniester	since 1990
158	Saint Vincent and the Grenadines	since 1979	193	Trinidad and Tobago	since 1962
159	Samoa	since 1962	194	Tunisia	since 1956
160	San Marino		196	Turkey	
161	Sao Tome and Principe	since 1975	197	Turkmenistan	since 1991
162	Saudi Arabia		197	Tuvalu	since 1978
163	Senegal	since 1960	198	Uganda	since 1962
164	Serbia	since 1992	199	Ukraine	since 1991
165	Seychelles	since 1976	200	United Arab Emirates	since 1971

ID	Country	Years	ID	Country	Years
201	United Kingdom		216	Viet Nam, South	until 1976
202	United States		217	Western Sahara	since 1976
209	Uruguay		218	Yemen	since 1990
210	Uzbekistan	since 1991	219	Yemen, North	until 1990
211	Vanuatu	since 1980	220	Yemen, South	until 1990
212	Vatican (Holy See)		221	Yugoslavia, SFRo	until 1992
213	Venezuela		222	Zambia	since 1964
214	Viet Nam	since 1976	223	Zanzibar	since 1963
215	Viet Nam, North	until 1976	224	Zimbabwe	

9.4 List of Excluded Countries

We decided to exclude a handful of countries and areas from our networks, even though they are among the countries for which SIPRI gathered the data. A key reason is that these countries and areas are not recognized as independent, sovereign states by the majority of other states. Furthermore, the data sets used in this paper which were not created by SIPRI are usually missing data for these entities.

1	Abkhazia	6	Palestine
2	Aruba	7	Somaliland
3	Northern Cyprus	8	South Ossetia
4	Darfur	9	Trans Dniester
5	Niue	10	Zanzibar

9.5 The Arms Trade Network in the Course of Times

In the following section, we visualize the arms trade networks for the years 1950, 1970, 1990 and 2013. The threshold was set at one million TIV. White nodes indicate no out-degree, yellow nodes indicate $0 <$ out-degree <5, and orange nodes indicate out-degree ≥ 5.

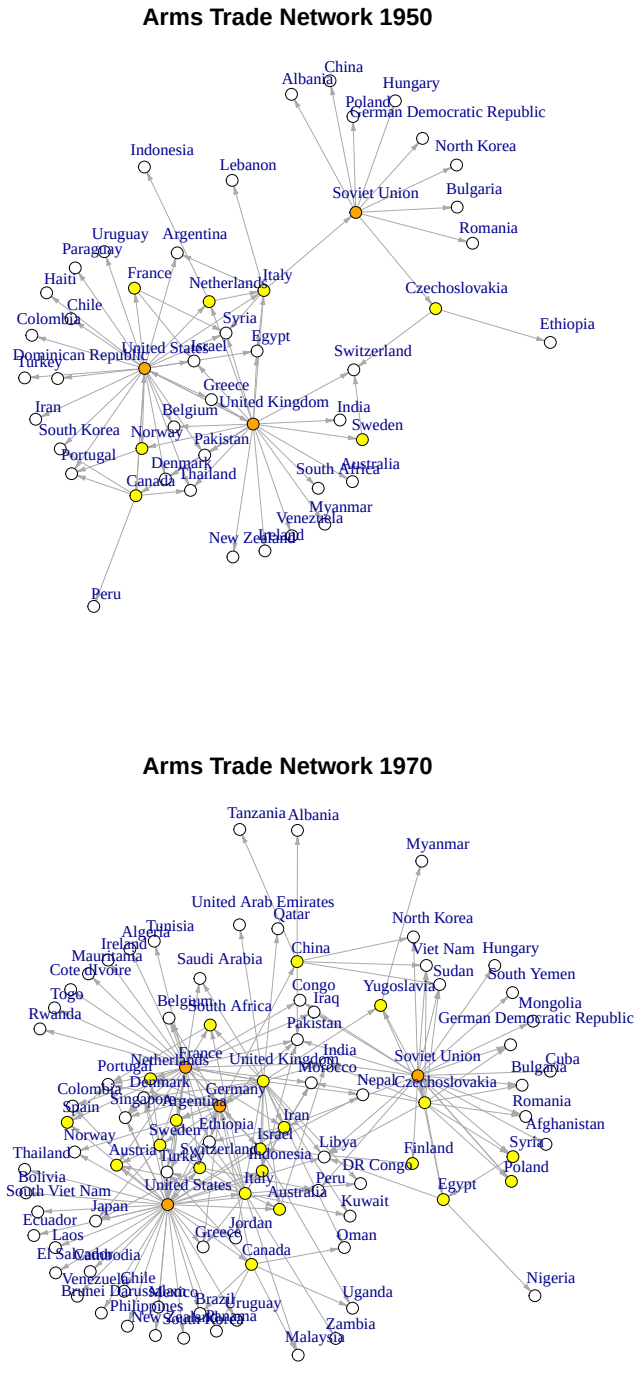
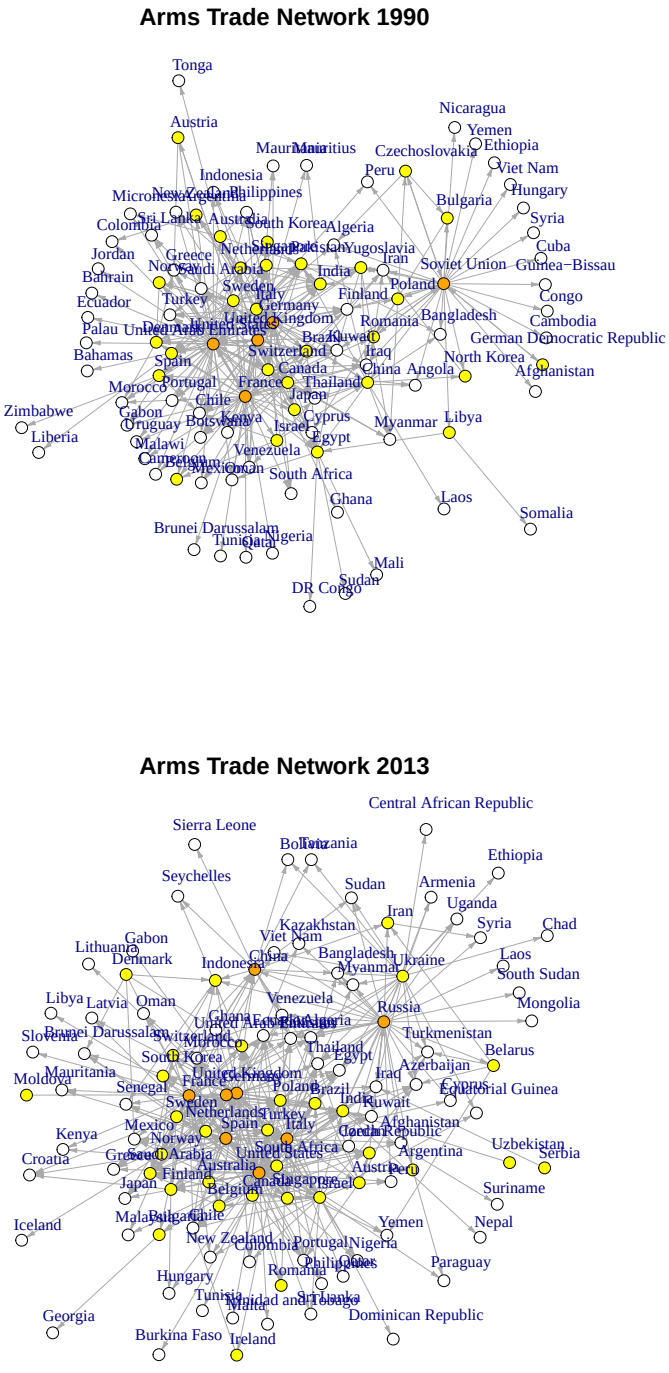


Figure 26: The arms trade networks for 1950 and 1970. White nodes indicate no out-degree, yellow nodes indicate 0 < out-degree <5, and orange nodes indicate out-degree $\geq$ 5



Data Source SIPRI

Figure 27: The arms trade networks for 1990 and 2013. White nodes indicate no out-degree, yellow nodes indicate 0 < out-degree <5, and orange nodes indicate out-degree $\geq$ 5

Bibliography

[1] Anders Akerman and Anna Larsson Seim. The global arms trade network 1950–2007. *Journal of Comparative Economics*, 42(3):535–551, 2014.

[2] L. D. Andersen and A.J.W. Hilton. Generalized latin rectangles 1: Construction and decomposition. *Discrete Mathematics*, 31(2):125–152, 1980.

[3] Carl de Boor. *A practical guide to splines: With 32 figures*, volume v. 27 of *Applied mathematical sciences*. Springer, New York, rev. ed edition, 2001.

[4] Jurgen Brauer. Arms industries, arms trade and developing countries. *Handbook of Defense Economics*, 2007.

[5] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 1985.

[6] Michael Brzoska. The economics of arms imports after the end of the cold war. *Defence and Peace Economics*, 2004.

[7] Margherita Comola. Democracies, politics, and arms supply. *Review of International Economics*, 20(1):150–163, 2012.

[8] S. J. Cranmer and B. A. Desmarais. Inferential network analysis with exponential random graph models. *Political Analysis*, 19(1):66–86, 2011.

[9] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge and New York, 1997.

[10] B. A. Desmarais and S. J. Cranmer. Consistent confidence intervals for maximum pseudolikelihood estimators, 2010.

[11] B. A. Desmarais and S. J. Cranmer. Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications*, 391(4):1865–1876, 2012.

[12] Reinhard Diestel. *Graphentheorie*. Springer-Lehrbuch Masterclass. Springer, Heidelberg [u.a.], 4. aufl edition, 2010.

[13] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on statistics and applied probability*. Chapman & Hall, New York, 1994.

[14] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with b -splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[15] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression: Modelle, Methoden und Anwendungen*. Statistik und ihre Anwendungen. Springer, Berlin and Heidelberg, 2. aufl edition, 2009.

[16] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

[17] Douglas M. Gibler. *International military alliances from 1648 to 2000*. CQ and Eurospan, Washington and D.C and London, 2004.

[18] Chong Gu. *Smoothing spline ANOVA models*. Springer series in statistics. Springer, New York, 2002.

[19] Mark S. Handcock. *Assessing degeneracy in statistical models of social networks: https://www.csss.washington.edu/Papers/wp39.pdf*. 2003.

[20] Steve Hanneke, Wenjie Fu, and Eric P. Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 2009.

[21] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*. Springer series in statistics. Springer, Berlin, 2004.

[22] Jenine K. Harris. *An introduction to exponential random graph modeling*, volume 173 of *Quantitative applications in the social sciences*.

[23] Trevor Hastie and Robert Tibshirani. *Generalized additive models*, volume 43 of *Monographs on statistics and applied probability*. Chapman and Hall, London and New York, 1st ed edition, 1990.

[24] Paul Holtom, Mark Bromley, and Verena Simmel. Measuring international arms transfers. *SIPRI Fact Sheet*, 2012.

[25] David R. Hunter. Curved exponential family models for social networks. *Soc Networks. 2007 Mar and 29(2): 216–230.*, 2007.

[26] David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

[27] David R. Hunter and Mark S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

[28] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. In *Journal of Statistical Software*, volume 24(3), pages 1–29, 2008.

[29] Christoph Jansen and Christian Schmid. *Eine statistische Analyse des Netzwerks des internationalen Waffenhandels von 1950-2012: Bericht im Rahmen des statistischen Consultingprojekts.* 2014.

[30] Dorothea Jansen. *Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele.* Lehrbuch. VS, Verl. für Sozialwiss., Wiesbaden, 3., überarb. aufl edition, 2006.

[31] Florian Johannsen and Inma Martinez-Zarzoso. Gravity of arms, http://works.bepress.com/inma_martinez_zarzoso/31/, 2014.

[32] Eric D. Kolaczyk. *Statistical analysis of network data: Methods and models.* Springer series in statistics.

[33] Philip Leifeld, S. J. Cranmer, and B. A. Desmarais. Estimating temporal exponential random graph models by bootstrapped pseudolikelihood: R package vignette for xergm 1.2, 2014.

[34] Monty G. Marshall. Polity iv project: Political regime characteristics and transitions, 1800-2013: http://www.systemicpeace.org/polity/polity4.htm, 2014.

[35] T. Mayer and S. Zignago. Notes on cepiis distances measure: the geodist database, 2011.

[36] P. McCullagh and John A. Nelder. *Generalized linear models*, volume 37 of *Monographs on statistics and applied probability.* Chapman and Hall, London and New York, 2nd ed edition, 1989.

[37] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Communications and control engineering series. Cambridge University Press, Cambridge and New York, 2nd ed edition, 2009.

[38] Matthew Moore. Arming the embargoed: A supply-side understanding of arms embargo violations. *Journal of Conflict Resolution*, 2010.

[39] Finbarr O'Sullivan. [a statistical perspective on ill-posed inverse problems]: Rejoinder. *Statistical Science*, 1(4):523–527, 1986.

[40] Richard Perkins and Eric Neumayer. The organized hypocrisy of ethical foreign policy: Human rights, democracy and western arms sales. *Geoforum*, 41(2):247–256, 2010.

[41] R Core Team. R: A language and environment for statistical computing, 2013. ISBN 3-900051-07-0.

[42] Garry Robins, Tom A.B. Snijders, Peng Wang, Mark S. Handcock, and Philippa E. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

[43] G. Rodriguez. Lecture notes on generalized linear models: http://data.princeton.edu/wws509/notes/, 2007.

[44] David Ruppert, M. P. Wand, and Raymond J. Carroll. *Semiparametric regression.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge and New York, 2003.

[45] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap.* Springer series in statistics. Springer Verlag, New York, 1995.

[46] J. David Singer, Stuart Bremer, and John Stuckey. Capability distribution, uncertainty, and major power war, 1820-1965. *Bruce Russett (ed) Peace and War and and Numbers and Beverly Hills: Sage and 19-48*, 1972.

[47] Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.

[48] Tom A.B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 2002.

[49] Stockholm International Peace Research Institute. http://www.sipri.org/.

[50] The Maddison-Project. http://www.ggdc.net/maddison/maddison-project/home.htm, 2013.

[51] Grace Wahba. *Spline models for observational data: Based on a series of 10 lectures at Ohio State University at Columbus, Mar. 23-27, 1987*, volume 59 of *CBMS-NSF Regional Conference series in applied mathematics.* Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[52] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8 of *Structural analysis in the social sciences*. Cambridge University Press, Cambridge and New York, 1994.

[53] Stanley Wasserman and Philippa Pattison. Logit models and logistic regression for social networks: An introduction to markov graphs and p-star. *Psychometrica Vol.61*, 1996.

[54] Spencer L. Willardson. *Under the influence of arms: the foreign policy causes and consequences of arms transfer: http://ir.uiowa.edu/etd/2660/*. PhD thesis, University of Iowa, 2013.

[55] Simon N. Wood. *Generalized additive models: An introduction with R*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, FL, 2006.

Declaration of Authorship

I hereby confirm that I have written the present thesis independently and without illicit assistance from third parties and using solely the aids mentioned.

Munich, May 19, 2015 .....

(Christian Schmid)