Department of Statistics
Ludwig-Maximilians University Munich

Separable Temporal Exponential Random Graph Models

Master's Thesis

Author:
Sevag Kevork
M.Sc. Biostatistics

Supervisor:
Prof. Dr. Göran
Kauermann

Submission Date: August 24, 2017

Abstract

In this thesis we investigate international major conventional weapons networks between 1950-2013. After an introduction to statistical network theory and some descriptive analysis of the data at hand, we introduce a modelling approach based on previous work of *Krivitsky and Handcock* on *Separable Temporal Exponential Random Graph Models*, however, we deviate from their approach and use *Generalized Additive Mixed Models (GAMM)*, for estimating the independent formation and dissolution parts of the model. Thereby we estimate smooth effects for covariates and lag-network statistics with the approach of *Hastie and Tibshirani* on *Varying Coefficient Models* and use country specific random intercepts in modelling the trading dyads.
Using this approach, we find relations between the formation and dissolution of major arms trades over time and some of the considered covariates. Different ways of model evaluation are presented and applied to our estimated model.

Contents

1 Introduction

International arms trade is a topic of high public, political, and scientific interest. Decisions to trade weapons and equip other countries with arms can be highly controversial, be based on various, often case specific, political considerations, and can have severe consequences.

As discussed by Hough et al. [2015], there is nothing new about the trading with weapons in the global marketplace, but this trade has changed over the years in many ways. For example, in the decades immediately following the Second World War, developed nations' trade in arms with the developing world, often consisted of supplying already outdated military equipment left over from the preceding conflict. By the 1980s, states in the developing world were demanding and receiving some of the most sophisticated weapons systems available, as they do today.
Another important development is, that the trade's expansion in recent times has been greatly facilitated by the twin trends of globalisation and economic liberalisation. Just as barriers and restrictions in other areas of economic activity have been lowered and eased as part of a global trend towards embracing free markets and deregulation, the same is true in relation to the arms trade. Similarly, advances in information and communications technology have enabled transactions to be conducted faster and more easily, and improvements in transportation – land, sea and air – have greatly aided the physical delivery of weapons.

It is therefore of huge general interest to scientifically understand mechanisms, that are related to such decisions and find factors that are connected to the establishment of arms trade relations.

In this thesis, we focus on international trade of major conventional weapons (MCW) over a period of 60 years and analyze publicly available data of the *Stockholms International Peace Research Institute* [SIPRI, 2017] based on statistical network modelling. We choose such an approach because trade data (in particular at country level) exhibits a specific network structure, that it characterized by multiple, pairwise trade relations that may or may not be established between all the different pairs of actors. Standard statistical (regression) models can not be directly applied to datasets of such structure, since their main assumption is always based on conditional independence of the, to be modelled, responses, e.g. the binary information of the existence of a trade relation or their specific

trade value. This assumption is, in such a network context, highly questionable, since the existence of a trade relation between two nations is most probably directly related to other trade activities of the involved countries.
Section 2 of this thesis, gives a general introduction to network data and its definitions. Section 3 will give an introduction to the data set been used in this thesis, discussing the structure of it, and providing descriptive results of the arms trade data set. In Section 4 we are going to introduce the ERGM *exponential random graph model*, the TERGM *temporal ergm* and finally, the STERGM *separable temporal ergm*, and discuss about their properties, also giving a short introduction of *markov chain* properties. In Section 5 we are going to introduce classical statistical regression models, by taking a closer look at the estimated spline smoothers. In Section 6, we will motivate and introduce our modelling approach combining the definitions of STERGMs and classical regression models, and finally, in Section 7 we will present and interpret the results.

2 Network Analysis

In the following, we will give a short introduction into the statistical analysis of network data and define some basic terminology that will be used during the rest of this thesis. The section is mainly based on Kolaczyk and Csárdi [2014].
In its most general sense, a network is just a collection of somewhat interconnected things. To deal with such an abstract entity from a mathematical point of view, networks are in general identified with the structure of *graphs*.
Speaking of networks in general, it is possible to distinguish between *directed* and *undirected* networks. Undirected networks just focus on the existence of any relation between two actors (e.g. friendship networks for several people), whereas directed networks consider the direction of the relation between two actors. Since trade networks are in general *directed*, we will focus on such networks, beginning with the formal definition of a (directed) graph:

Definition 1: A graph G is defined as a pair of sets of *vertices* V and *edges* E, $G := (V, E)$. V is typically a finite set and $E \subset V \times V$. The elements of V are called *vertices* or *nodes*, while the elements of E are called *edges* or *ties*. When we think of a network as a graph the vertices represent the actors / things in the network and the edges indicate a (directed) relation / connection between two actors. Such relations can be binary or valued, depending on the context.

In our case, the elements of the set of *vertices* V are symbolizing the countries in the world, which we denote these as $v_i, v_j \in V$ with their indices i and j. In our context, the actors in the network are the countries, and the set of edges the relation between two countries (actors). This relation as mentioned above can be either directed or undirected. In the context of arms trade network, this relation indicates, whether a country i exports major conventional weapons to country j or not. That's why arms trade network is a directed network. The case that country i is selling weapons to country j, does not imply that j is also selling weapons to i.

As next, we are going to define some terms for a graph $G = (V, E)$, which are crucial in network analysis. For an edge $e_{ij} = (v_i, v_j) = (i, j)$ from actor i to j, we call v_i the *tail* and v_j the *head* of edge e_{ij}. We will also refer to v_i as the *sender* and to v_j as the *receiver* in the network. An important restriction that we make, is that a graph has no *loops* $e_{ij} = (v_i, v_i)$, i.e. edges with tail and head on the same vertex. This means, we are not going to pay attention to weapons produced for a country's own use. The number of actors in a network $N_V = |V|$ is usually called the *order* of a network, and the number of observed edges $N_E = |E|$ is called its *size*. A directed network of order N_V has $N := N_V^2 - N_V$ possible edges if we are not interested in *loops*. Based on those two statistics we can define the density of a network or graph:

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

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

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 edge, has density $\rho = 1$, while an *empty graph* is defined as a graph without any edges, $\rho = 0$.
Two vertices v_i and v_j are called *adjacent* if they are connected by an edge $e_{ij} = (v_i, v_j)$. The rather abstract structure of a graph (and consequently a network), can be completely specified by the definition of the so-called *adjacency matrix*:

Definition 3: Let $G = (V, E)$ be a finite, directed graph and $V = (v_1, \ldots, v_{N_V})$ be an enumeration of the set of vertices in G. The matrix $A := (a_{ij}) \in \mathbb{R}^{N_V \times N_V}$ is called the adjacency matrix. Its entries a_{ij} indicate the relation between v_i and v_j in the network.

If $(v_i, v_j) \in E$, $a_{ij} > 0$. For binary networks A is given by

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

$i, j \in \{1, \ldots, N_V\}$.
In the case of weighted / valued networks, the entries for observed edges a_{ij} get replaced by their specific values. Such values can, for example, specify the strength of the connection between two vertices.
So far, we mainly focused on statistics that characterize the structure of a whole observed network and described a way of storing network data. Focusing on the single actors in a network, one is often interested in their *centrality* or *popularity* within a network. Therefore we define:

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

$$\deg^{in}(v) := |\{(v_1, v_2) \in E : v_2 = v\}|$$
$$\deg^{out}(v) := |\{(v_1, v_2) \in E : v_1 = v\}|$$

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 hand, the *out-degree* is defined as the number of tails connected to v. Vertices with high in- or out-degrees can be seen as central, important actors in a network.

An another term introduced is the definition of a *dyad*. A dyad is a group of two actors and their relation. For directed networks we are going to differentiate between three kind of dyads: a dyad (ij) is *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 *asymmetric* if there is only one edge between the two actors, i.e. $e_{ij} \in E \vee e_{ji} \in E$ where $\vee$ is defined as *exclusive*. Lastly, a dyad is called *null* if there is no edge between two actors i and j i.e. $e_{ij}, e_{ji} \notin E$.

Most of the (observable) real-world networks are not static but evolving in time. New actors can join a network, edges can form, change in strength or disappear. Despite it is generally conceivable to observe networks in continuous time by registering any change in vertices or edges at its actual timepoint, it is often more realistic to observe a time-series of networks on a discrete grid. Data on such time evolving networks can then be stored in adjacency matrices A^t, $t = 1, \ldots, T$.

3 Data Structure and Descriptive Analysis

The international arms trade data for major conventional weapons was provided by the Stockholms International Peace Research Institute [SIPRI, 2017]. This initiative collects all officially registered arms trade activities of major conventional weapons between international actors and documents them in an online database. SIPRI has developed a unique system in order to measure the volume of international transfers of arms, which is listed in *trend indicator value* (TIV), a measure based on production costs see [Holtom et al., 2012]. There are many advantages using TIV instead of monetary value as cash flow, according to SIPRI using monetary value as cash flow such as USD would lead to distorted information. TIV has the advantage to be consistent over time, which makes it possible to compare the arms flow of different periods.

In our analysis we focus on all countries that have at least one registered trade activity between 1950 and 2013 and use any yearly, registered, trade between two countries as binary directed trade information. This leads to a raw data set of yearly, binary, pairwise information for 257 actors. When we examine our data set, one will recognize that not only countries are involved in the network, but also international organizations like the UN and NATO, extremist/terrorist groups like AL-Qaida, ISIS, and fiercely disputed regions like Eastern Ukraine or Nagorno-Karabakh can be involved in the network. Due to missing values in covariates (no informations for at least one covariate in all analyzed years), we focus our investigation of 218 countries, the excluded organizations and fiercely disputed regions are given as a list in appendix 9.

In the following section we present some basic descriptive statistics for the yearly trade networks in our considered dataset. In this connection the following questions are of interest:

- How the network changes over time?
- How is this change quantifiable?

One of the interesting issues is the change of density of the arms trade network over time. However, the number of nations differ over the entire period, some nations may not more exist that respectively been situated in a certain year. This problematic can be counteracted as follows: In each year, we only consider the nations they were active involved in the arms transfer network. This approach has the consequence, that the density values for each year will be calculated from networks with different number of

actors. The number of possible edges in a network grow exponentially (not linear) with increasing number of nodes (actors), that's why by comparison of the density values we should additionally take the size of the network into account for the concerned year. The corresponding time series are shown in Figure 1.

The left plot of Figure 1, shows the time series of the number of actors in the network for each year, as we can see, there is a steady increase of actors between 1950-1980. While in 1950, observing about 50 actors in the network, there are about 120 actors in 1980. On the following next ten years we have falling number of actors with the low points between 1991 and 1993, it is conspicuous, that this period has to do with the disintegration of the Soviet Union.

The right plot of Figure 1,shows the time-series of network's density. Even as we are careful while interpreting this plot, since the number of actors changes over time.
To illustrate our point of view, we consider the density of the arms trade network from year 1951-2006. We can recognize, that both of the networks have a density of about 0.03, respectively. Both networks are shown in Figure 2. We can see immediately, that both networks have fundamental different structure, because of the different number of actors in the networks respectively. In Figure 4 we visualize some arms trade networks through changing times.

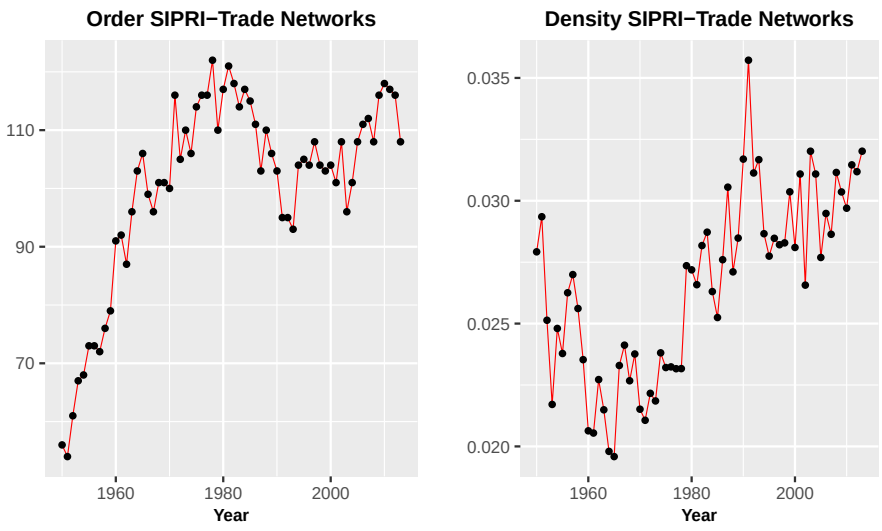


Figure 1: The number of the actors included in the arms trade networks (left) and the density of the networks (right) for the period 1950-2015

Figure 3 visualizes the size, i.e. the number of edges, of the yearly trade networks. It is therefore the time series of the yearly number of observed, registered arms trades between two nations.

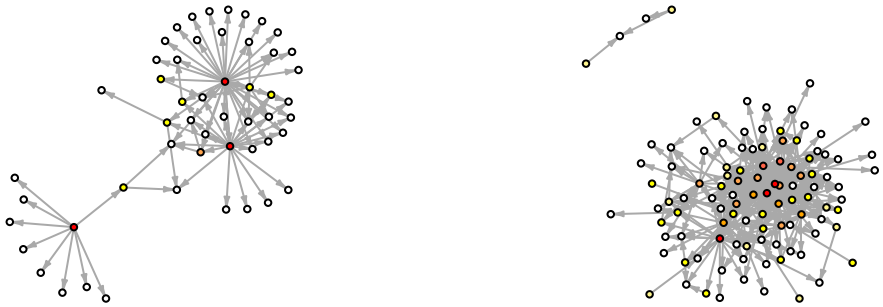


Figure 2: Arms trade network 1951 (left) and arms trade network 2006 (right), equal density with different number of actors

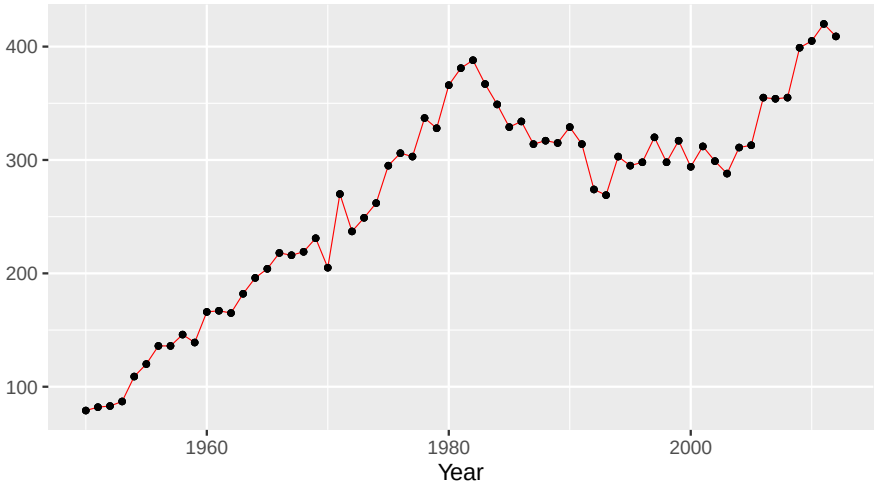


Figure 3: Size of the arms trade network for the period 1950-2012

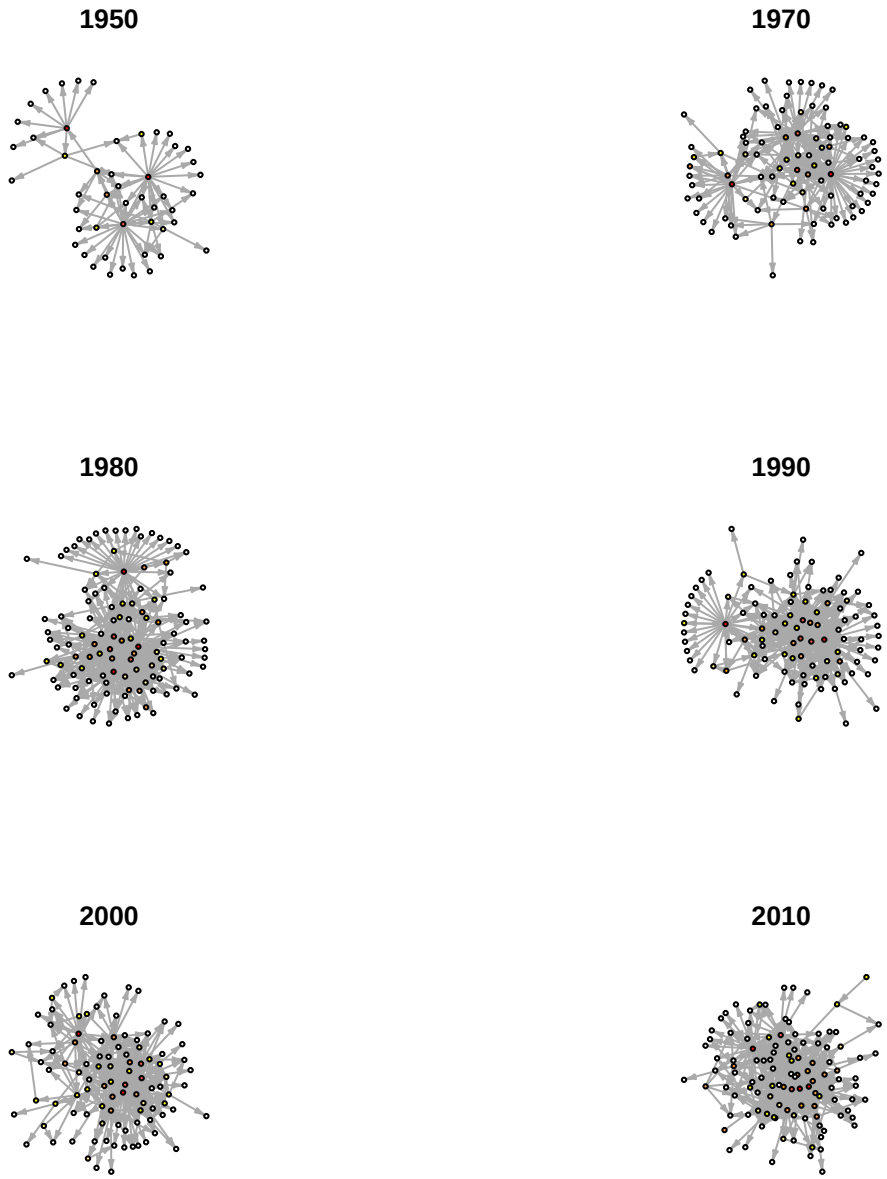


Figure 4: Arms Trade Network through changing times

3.1 Assignation of the key actors

In this section we will identify the countries, who played a central key role in arms trade network over the period 1950-2013. In order to take the disintegration of the Soviet Union into account, we will divide our entire period into periods from 1950-1991 and 1992-2013.

One measurement criteria for the relevance of a nation in arms trade network is the accumulated TIV over the observed period, distinguishing between import and export TIVs.

In Table 1 and Table 2 we list the top 10 supplier and recipient nations over the periods 1950-1991 and 1992-2013, respectively. It is conspicuous the big dominance of some nations in export of major conventional weapons. From 1950-1991, hold the USA and the Soviet Union together almost twice so high export TIV as all the nations together. We have a similar picture for USA and Russia according their export TIV for the period 1992-2013, although with Germany and France, we have two nations with more engagement in arms export.

The picture changes for the top 10 recipient nations. The biggest recipients for the period 1950-1991 are India and Germany, and for 1992-2013 India and China, however the import TIVs are distributed more evenly on several nations.

	Country	**TIV**
1	Soviet Union	453346.43
2	United States	399666.61
3	United Kingdom	106784.90
4	France	76981.93
5	Germany	38859.38
6	China	29080.99
7	Czechoslovakia	28850.87
8	Italy	19376.97
9	Switzerland	9978.70
10	Netherlands	9484.59

	Country	**TIV**
1	United States	185312.88
2	Russia	109621.57
3	Germany	39139.71
4	France	32751.90
5	United Kingdom	28070.28
6	China	17952.36
7	Netherlands	10997.93
8	Italy	10211.12
9	Ukraine	9849.26
10	Israel	9211.44

Table 1: The left table lists the top 10 supplier nations for the period 1950-1991 and the right the top 10 supplier nations for the period 1992-2013

	Country	TIV
1	India	65237.31
2	Germany	51418.49
3	Iraq	44838.28
4	Japan	42891.52
5	Egypt	40913.41
6	Iran	39710.55
7	Poland	37521.23
8	Syria	35549.60
9	China	34102.77
10	German Democratic R.	30829.97

	Country	TIV
1	India	43094.62
2	China	38575.73
3	South Korea	26143.63
4	Turkey	25115.59
5	Saudi Arabia	22778.87
6	Greece	19915.27
7	United Arab Emirates	18311.78
8	Egypt	17326.19
9	Japan	16822.00
10	Pakistan	16553.95

Table 2: The left table lists the top 10 recipient nations for the period 1950-1991 and the right the top 10 recipient nations for the period 1992-2013

Another measurement criteria for the relevance of a nation in arms trade network is its out- and in-degree over the observed period, i.e. it presents the number of different selling / buying countries over a given observation period. It equals the nation-specific out- or in-degree of the binary trade network, aggregated over the observation period. The results are shown in Table 3 and Table 4.

	Country	Has Supplied
1	United States	127
2	France	109
3	United Kingdom	106
4	Germany	83
5	Italy	83
6	Canada	77
7	Soviet Union	69
8	Switzerland	59
9	Netherlands	51
10	Sweden	47

	Country	Has Supplied
1	United States	115
2	France	95
3	Russia	87
4	Germany	84
5	Italy	78
6	United Kingdom	73
7	Ukraine	69
8	Israel	67
9	Canada	58
10	China	58

Table 3: The left table lists the top 10 supplier nations according to their out-degree for the period 1950-1991 and the right the top 10 supplier nations for the period 1992-2013

	Country	Was Supplied
1	Iran	23
2	Iraq	23
3	Egypt	21
4	India	20
5	Indonesia	20
6	Morocco	20
7	Nigeria	18
8	Soudan	17
9	Thailand	17
10	Argentina	16

	Country	Was Supplied
1	USA	26
2	Indonesia	23
3	UAE	23
4	Malaysia	22
5	Pakistan	22
6	Thailand	21
7	Brazil	20
8	Peru	20
9	India	19
10	Iraq	19

Table 4: The left table lists the top 10 recipient nations according to their in-degree for the period 1950-1991 and the right the top 10 recipient nations for the period 1992-2013

Actually it is difficult to make up the central key actors in the international arms trade. According to the underlying criteria we choose, we obtain different list of samples.

3.2 Assignation of the relevant effects

In this section we will identify some network characteristic structures such as: Do we have reciprocative trade? How many nations do we have that only buy weapons but dont sell? Which characteristic structures change over time and which stay stable?
Our goal, is implying the detected structures in the form of network statistics in our modelling approach in section 6, to generate the most suitable model we can.

Figure 5 indicates the proportion of onesided and reciprocative edges to the overall existing edges. We can see, that in each observed year, the proportion of onesided edges clearly predominate the proportion of reciprocative edges. This characteristical structure is identical for the entire period and thats why should be considered in our modelling approach. Another peculiarity in the arms trade network for the observed period is that major part of the nations exclusively import major conventional weapons, without taking part in exporting them.

Figure 5: Proportion of onesided (red) and reciprocative (blue) edges to the overall existing edges

The left plot of Figure 6 depicts the average out-degree distribution over the entire period, we can see that the arms trade network exhibits a large proportion of actors with out-degree of nulls over the entire period, the left plot depicts the average in-degree distribution, conspicuous is especially the large average in-degree=1 actors. Obviously it is characteristic for the arms trade network that actors are supplied from one nation. Figure 7 indicates the number of nations with an out-degree of 0,1,2,3 and in-degree of 0,1,2,3 for the period of 1950-2013, respectively.

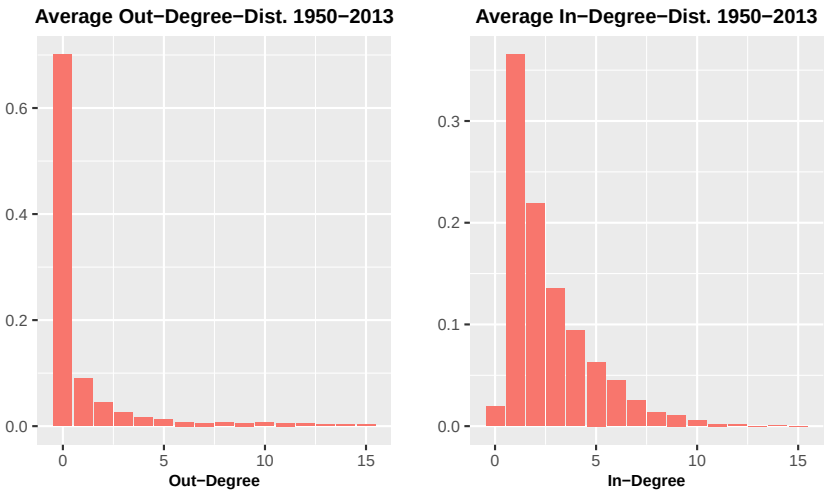


Figure 6: Average out-degree distribution (left) and the average in-degree distribution for the period 1950-2013

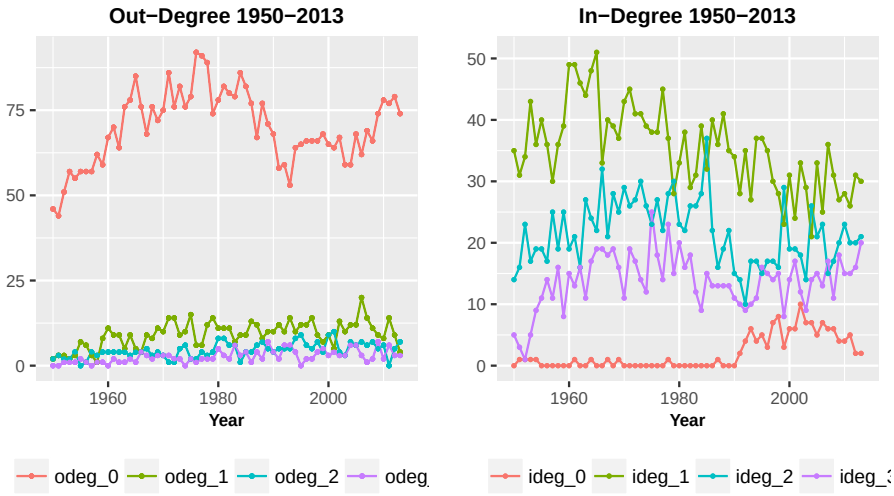


Figure 7: Time series of in- and out- degree values over the entire period

4 Network Modelling Approaches

In this section we will introduce some network modelling approaches, such as the family of exponential graph models, that will be of use for our proposed modelling framework introduced in Section 6.

4.1 The Exponential Random Graph Model

The ERGM Robins et al. [2007] takes the adjacency matrix of an observed network Y^{obs} as the realization of a matrix valued random variable Y. Recall definition (3) in Section 2 a network of N_V nodes is defined as adjacency matrix $A = Y = (y_{ij}) \in \mathbb{R}^{N_V \times N_V}$, where $y_{ij} \in \{0, 1\}$ for all $i, j \in \{1, \dots, N_V\}$. Where $y_{ij} = 1$ means that there is an edge going from i to j, while $y_{ij} = 0$ indicates that this edge does not exist. Since our model does not involve loops, we have $y_{ii} = 0$ for all $i \in \{1, \dots, N_V\}$. So we can define Y

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

as the set of all existing networks of order N_V without loops, that depends i.a. on the number of specific configurations or patterns in the edges of the networks (and potentially additional covariates) and thereby considers potential dependence structures in the occurence of edges. With this definition, we can define Y as a matrix valued

random variable, with the probability function

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

where

- $\theta \in \mathbb{R}^q$ is a q-dimensional vector of parameters
- $\Gamma(y)$ is a q-dimensional vector of statistics based on the adjacency matrix y
- $\kappa(\theta, y) := \sum_{y \in \mathcal{Y}(N)} exp(\theta^T \cdot \Gamma(y))$ is a normalizing factor that ensures that (1) is a legitimate probability distribution

Specification of Y, including the number of vertices, n, is an important yet often overlooked aspect of model (1). At its largest, for a fixed n, Y may contain up to $N = 2^{n(n-1)}$ networks, a very large number even for moderate-sized n, which makes calculation of $\kappa(\theta, y)$ the primary barrier to inference using this model.

An alternative specification of the model (1) clarifies the interpretation of the coefficients. To articulate this alternative, we first introduce the notion of a vector of *changestatistics*. Such a vector is a function of three things: A particular choice $\Gamma(.)$ of statistics defined on a network, a particular network y, and a particular pair of different vertices (i, j). We define the vector of change statistics as

$$\delta_\Gamma(y)_{ij} = \Gamma(y_{ij}^+) - \Gamma(y_{ij}^-)$$

where y_{ij}^+ and y_{ij}^- represent the networks realized by fixing $y_{ij} = 1$ or $y_{ij} = 0$, respectively, while keeping all the rest of the network exactly as in y itself. In other words, $\delta_\Gamma(y)_{ij}$ is the change in the value of the network statistic $\Gamma(y)$ that would occur if y_{ij} were changed from 0 to 1 while leaving all of the rest of y fixed.

In terms of the change statistic vector, model (1) may be shown to imply the following distribution of the Bernoulli variable Y_{ij}, conditional on the rest of the network:

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

$$\Rightarrow logit(\mathbb{P}_{\theta,y}(Y_{ij}=1|Y^c_{ij}=y^c_{ij}))=\theta^T\cdot\delta_\Gamma(y)_{ij} \tag{2}$$

When the network statistics involve covariates X in addition to y, we may add X to the notation and write $\delta_\Gamma(y,X)_{ij}$.

Equation (2) reveals two facts: First, the probability on the left hand side depends on y^c_{ij} only through the change statistics $\delta_\Gamma(y)_{ij}$, not on $\Gamma(y^+_{ij})$ or $\Gamma(y^-_{ij})$ themselves. In many cases, it is much easier to calculate $\delta_\Gamma(y)_{ij}$ than it is to calculate $\Gamma(y^+_{ij})$ or $\Gamma(y^-_{ij})$, and this fact can lead to efficient computational algorithms.

Second, Equation (2) says that each component of the δ vector may be interpreted as the increase in the conditional log-odds of the network, per unit increase in the corresponding component of $\Gamma(y)$, resulting from switching a particular Y_{ij} from 0 to 1 while leaving the rest of the network fixed at Y^c_{ij}.

Exponential family random graph models (ERGMs) are a natural way to represent dependencies in cross-sectional graphs and dependencies between graphs over time, particularly in a discrete-time context. Hanneke define and describe a *Temporal ERGM (TERGM)* postulating an exponential family model for the transition probability from at time t to a network at time $t+1$. In section 4.2 we review discrete-time ERGM-based network models, and discuss the specification of such models.

4.2 The Temporal Exponential Random Graph Model

As discussed by Hanneke et al. [2010], we consider a discrete-time dynamic network model in which the network at time t is a single draw from an ERGM conditional on the network at time $t-1$. Specifically, one way to simplify a statistical model for evolving

networks, is to make a *Markov Chain* assumption (see section 4.2.1) on the network from one time step to the next. If Y^t is the matrix representation of a single-relation network at time t, then we might make the assumption that Y^t is independent of $Y^1, \ldots, Y^{t-2}$ given Y^{t-1}. Put another way, a sequence of network observations $Y^1, \ldots, Y^t$ has the property that

$$\mathbb{P}(Y^2, Y^3, \ldots, Y^t | Y^1) = \mathbb{P}(Y^t | Y^{t-1}) \cdot \mathbb{P}(Y^{t-1} | Y^{t-2}) \cdots \mathbb{P}(Y^2 | Y^1)$$

We recall again definition (3) in Section 2 a network of N_V nodes is defined as adjacency matrix $A = Y = (y_{ij}) \in \mathbb{R}^{N_V \times N_V}$, where $y_{ij} \in \{0, 1\}$ for all $i, j \in \{1, \ldots, N_V\}$. Where $y_{ij} = 1$ means that there is an edge going from i to j, while $y_{ij} = 0$ indicates that this edge does not exist. Since our model does not involve loops, we have $y_{ii} = 0$ for all $i \in \{1, \ldots, N_V\}$. Further let $Y^t \in \mathcal{Y}$ be a random variable representing the state of the network at the discrete time point t and $y^t \in \mathcal{Y}$ be its realization.

The one-step transition probability from y^{t-1} to y^t is then defined to be

$$\mathbb{P}(Y^t = y^t | Y^{t-1} = y^{t-1}; \theta) = \frac{exp(\theta^T \cdot \Gamma(y^t, y^{t-1}))}{\sum_{y \in \mathcal{Y}(N)} exp(\theta^T \cdot \Gamma(y^{t-1}))}, \quad y^t, y^{t-1} \in \mathcal{Y} \tag{3}$$

where

- $\theta \in \mathbb{R}^q$ is a q-dimensional vector of parameters
- $\Gamma() : \mathcal{Y}^2 \to \mathbb{R}^p$ is the sufficient statistic for the transition from network y^{t-1} at time $t-1$ to network y^t at time t with $q \leq p$
- $\kappa(\theta, y^{t-1}) := \sum_{y \in \mathcal{Y}(N)} exp(\theta^T \cdot \Gamma(y^{t-1}))$ is a normalizing factor that ensures that (3) is a legitimate probability distribution

TERGMs are a natural elaboration of the traditional ERGM framework. They are essentially stepwise ERGM in time.

The class of models specified by (3) is very broad and a key component of model specification is the selection of Γ. However, the choices in this dynamic situation are richer and can be any valid network statistics evaluated on y^t especially those that depend on y^{t-1}. Hanneke et al. [2010] focused on a choice of Γ that had the property of *conditional dyadic independence* that

$$\mathbb{P}(Y^t = y^t | Y^{t-1} = y^{t-1}; \theta) = \prod_{(i,j) \in \mathbb{Y}} \mathbb{P}(Y^t_{i,j} = y^t_{i,j} | Y^{t-1} = y^{t-1}; \theta)$$

the distribution of Y^t in which edge states are independent, but only conditional on *the whole of* Y^{t-1}, and where $\mathbb{Y}$ is the set of potential edges between them.
However, caution must be used in interpreting their parameters. Consider the simplest such statistic, the edge count

$$\Gamma(y^t, y^{t-1}) = |y^t|$$

That means, a higher coefficient on Γ will, for any y^{t-1}, produce a Y^t distribution in which networks with more edges have a higher probability. But, this term would accomplish it in two ways simultaneously: it would both increase the weight of those networks in which more edges were formed on previously empty dyads and increase the weight of those networks in which more extant edges were preserved (fewer dissolved). That is, it would both increase the *incidence* (the rate at which new edges are formed) and increase the *duration* (how long they tend to last once they do).
Hanneke et al. [2010] gave an example of a statistic that controls the rate of evolution of the network: a measure of *stability*. This statistic counts the number of edge variables whose states did not change between time steps, which is then divided by the maximum number of edges an actor could have (a constant):

$$\Gamma(y^t_{i,j}, y^{t-1}_{i,j}) = \frac{1}{n-1} \sum_{(i,j)\in\mathbb{Y}} (y^t_{i,j} y^{t-1}_{i,j} + (1 - y^t_{i,j})(1 - y^{t-1}_{i,j}))$$

That means, a higher coefficient on it will slow the evolution of the network down and a lower coefficient will speed it up. From the point of view of *incidence* and *duration*, it will do so in two ways: a higher coefficient will result in networks that have fewer new edges formed and fewer extant edges dissolved, *incidence* will be decreased and duration will be increased.

The two-sided nature of these effects tends to muddle parameter interpretation, but a more substantial issue arises if selective mixing statistics, like those described by Koehly and Pattison [2005] are used. The coupling between the *incidence* of edges and their *duration* not only makes such terms problematic to interpret, but has a direct impact on modeling.
In section 4.3 we will describe and motivate the concept of *separability* of formation and dissolution in a dynamic network model, and describe the *Separable Temporal Exponential Random Graph Model (STERGM)*.

4.2.1 Markov Chains

As discussed by Fahrmeir et al. [1981] discrete-time stochastic process is a sequence of random variables $\mathbf{X} = \{X_n : n \in I\}$ where I is a discrete index set, and to be the set of nonnegative integers $I = \{0, 1, 2, \ldots\}$, so

$$\mathbf{X} = \{X_n : n = 0, 1, 2, \ldots\}$$

We will denote the state space of $\mathbf{X}$ by S, where S is the set of all possible values of any of the X_i's. The state space is also assumed to be discrete, and we let $|S|$ denote the number of elements in S, called the *cardinality* of S. So S could be ∞ or some finite positive integer.

A discrete-time stochastic process $\mathbf{X}$ is said to be a *Markov Chain* if it has the *Markov Property*:

Markov Property (version 1):
For any $s, i_0, \ldots, i_{n-1} \in S$ and any n ≥ 1,

$$\mathbb{P}(X_n = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = s | X_{n-1} = i_{n-1})$$

In words, the distribution of X_n given the entire past of the process only depends on the immediate past. Note that, we are not saying that, for example X_{10} and X_1 are independent. They are not. However, given X_9, for example, X_{10} is conditionally independent of X_1. Graphically, we may imagine being on a particle jumping around in the state space as time goes on to form a (random) sample path. The Markov property is that the distribution of where I go to next depends only where I am now, not on where I have been. This property is a reasonable assumption for many real-world processes.

Note that, as with the notation of independence, in applied modeling the Markov property is not something we usually try to prove mathematically. It usually comes into the model as an *assumption*, and its validity is verified either empirically by some statistical analysis or by underlying a-priori knowledge about the system being modeled.
A useful alternative formulation of the Markov property is:

Markov Property (version 2):
For any $s, i_0, \ldots, i_{n-1} \in S$ and any n ≥ 1 and m ≥ 0

$$\mathbb{P}(X_{n+m} = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) = \mathbb{P}(X_{n+m} = s | X_{n-1} = i_{n-1})$$

In words, this says that the distribution of the process at any time point in the future given the *most recent* past is independent of the earlier past. We should prove that the versions of the Markov property are equivalent, because version 2 appears on the surface to be more general. We do this by showing that each implies the other. It is clear that version 2 implies version 1 by setting $m = 0$. We can use conditioning and induction argument to prove that version 1 implies version 2, as follows.

Version 2 is certainly true for $m = 0$ (it is exactly version 1 in this case). The induction hypothesis is to assume that version 2 true holds for some arbitrary fixed m and the induction argument is to show that this implies it must also hold for $m + 1$. If we condition on X_{n+m} then

$$\begin{aligned}
&\mathbb{P}(X_{n+m+1} = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&= \sum_{s' \in S} \mathbb{P}(X_{n+m+1} = s | X_{n+m} = s', X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&\qquad \times \mathbb{P}(X_{n+m} = s' | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}).
\end{aligned}$$

For each term in the sum, for the first probability we can invoke version 1 of the Markov property and for the second probability we can invoke the induction hypothesis, to get

$$\begin{aligned}
&\mathbb{P}(X_{n+m+1} = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&= \sum_{s' \in S} \mathbb{P}(X_{n+m+1} = s | X_{n+m} = s', X_{n-1} = i_{n-1}) \\
&\qquad \times \mathbb{P}(X_{n+m} = s | X_{n-1} = i_{n-1}).
\end{aligned}$$

Note that in the sum, in the first probability we left the variable X_{n-1} in the conditioning. We can do that because it doesn't affect the distribution of X_{n+m+1} conditioned on X_{n+m}. The reason we leave X_{n-1} in the conditioning is so we can use the basic property that

$$\mathbb{P}(A \cap B | C) = \mathbb{P}(A | B \cap C) \cdot \mathbb{P}(B | C)$$

for any events A, B and C. With $A = \{X_{n+m+1} = s\}$, $B = \{X_{n+m} = s'\}$ and $C = \{X_{n-1} = i_{n-1}\}$. we have

$$\begin{aligned}
&\mathbb{P}(X_{n+m+1} = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&= \sum_{s' \in S} \mathbb{P}(X_{n+m+1} = s, X_{n+m} = s' | X_{n-1} = i_{n-1}) \\
&\qquad = \mathbb{P}(X_{n+m+1} = s | X_{n-1} = i_{n-1}).
\end{aligned}$$

So version 2 holds for $m + 1$ and by induction for all m.

4.3 The Separable Temporal ERGM

Intuitively, those processes and factors that result in edges being formed, are not the same as those that result in edges being dissolved.
Furthermore, it is often the case in practice that information about cross-sectional properties of a network has a different source from that of the information about its longitudinal properties (i.e *duration*), and it is useful to be able to consider them separately Krivitsky and Handcock [2014].
Thus, it is useful for the parameterization of a model to allow seperate control over *incidence* and *duration* of edges and separate interpretation.

Consider a class of discrete-time models for network evolution, which assumes that these processes are separable from each other within a time step. We consider in this section a sub-class of models based on the ERGM family.

We represent networks as sets of edges, so given $y, y' \in \mathcal{Y}$, the network $y \cup y'$ has the edge (i, j) if, and only if, (i, j) exists in y or y' or both, the network $y \cap y'$ has (i, j) if, and only if, (i, j) exists in both y and y', and the network $y \setminus y'$ has (i, j) if, and only if, (i, j) exists in y but not in y'. The relation $y \supseteq y'$ holds, if, and only of, y has all of the edges that y' does, and conversely for $y \subseteq y'$.

$y_{i,j}^{t-1}$	$\rightarrow$	$(y_{i,j}^{+}, y_{i,j}^{-})$	$\rightarrow$	$y_{i,j}^{t}$
0	$\rightarrow$	(0,0)	$\rightarrow$	0
0	$\rightarrow$	(1,0)	$\rightarrow$	1
1	$\rightarrow$	(1,0)	$\rightarrow$	0
1	$\rightarrow$	(1,1)	$\rightarrow$	1

Table 5: Possible transitions of a single edge variable

Consider the evolution of a random network at time $t-1$ to time t, and define two intermediate networks, the *formation network* Y^{+}, considering of the initial network Y^{t-1} with edges formed during the time step added and the *dissolution network* Y^{-}, considering of the initial network Y^{t-1} with edges dissolved during the time step removed (with y^{+} and y^{-} being their realization, respectively). Then given y^{t-1}, y^{+}, and y^{-}, the network y^{t} may be evaluated via a set of operation, as

$$y^{t} = y^{+} \setminus (y^{t-1} \setminus y^{-}) = y^{-} \cup (y^{+} \setminus y^{t-1}) \tag{4}$$

Since it is the networks y^{t-1} and y^t that are actually observed, y^+ and y^- may be regarded as latent variables, but it is possible to recover them given y^{t-1} and y^t, because an edge variable can only be in one of four states given in Table 5. Each possibility has a unique combination of edge variable states in y^{t-1} and y^t, so observing the network at the beginning and the end allows the two intermediate states to be determined as $y^+ = y^{t-1} \cup y^t$ and $y^- = y^{t-1} \cap y^t$.
If Y^+ is conditionally independent of Y^- given Y^{t-1} then

$$\begin{aligned} \mathbb{P}(Y^t = y^t | Y^{t-1} = y^{t-1}; \theta) &= \mathbb{P}(Y^+ = y^+ | Y^{t-1} = y^{t-1}; \theta) \times \\ &\quad \mathbb{P}(Y^- = y^- | Y^{t-1} = y^{t-1}; \theta) \end{aligned} \tag{5}$$

We refer to the two factors on the right hand side as the *formation model* and the *dissolution model*, respectively. Suppose that we can express $\theta = (\theta^+, \theta^-)$ where the formation model is parametrized by θ^+ and the dissolution model by θ^-.

Definition: We say that a dynamic model is *seperable* if Y^+ is conditionally independent of Y^- given Y^{t-1} and the parameter space of θ is the product of the individual parameter spaces of θ^+ and θ^-.

We refer to such a model as separable, because it represents an assumption that during a given discrete time step, the process by which the edge form does not interact with the process by which they dissolve (Krivitsky and Handcock [2014]): both are separated (in the conditional independence sense) from each other conditional on the state of the network at the beginning of the time step.

For the sake of completeness, we will further introduce how the *formation model* and the *dissolution model* can in the ERGM context be modeled, however for our final model approach introduced in section 6 will not be relevant.

For the components of the seperable model, we specifically model the formation model

$$\mathbb{P}(Y^+ = y^+ | Y^{t-1} = y^{t-1}; \theta^+) = \frac{exp(\theta^+ \cdot \Gamma^+(y^+, y^{t-1}))}{\delta_{\Gamma^+}(\theta^+, y^{t-1})}; \quad y^+ \in \mathcal{Y}^+(y^{t-1})$$

and the dissolution model

$$\mathbb{P}(Y^- = y^- | Y^{t-1} = y^{t-1}; \theta^-) = \frac{exp(\theta^- \cdot \Gamma^-(y^-, y^{t-1}))}{\delta_{\Gamma^-}(\theta^-, y^{t-1})}; \quad y^- \in \mathcal{Y}^-(y^{t-1})$$

with their normalizing constants $\delta_{\Gamma^+}(\theta^+, y^{t-1})$ and $\delta_{\Gamma^-}(\theta^-, y^{t-1})$ summing over $\mathcal{Y}^+(y^{t-1})$ and $\mathcal{Y}^-(y^{t-1})$, respectively.

Thus, the STERGM class is a subclass of a first order Markov TERGM of Hanneke et al. [2010] described in Section 4.2. However, the essential issue is the specification of models within these classes. What is gained is ease of specification, tractability of the model, and substantial improvement in interpretability.
In the next Section we are going to introduce classical regression models, in order to come to our final model selection combining both *STERGM* and classical regression definitions.

5 Classical Regression Models

In this Section we are going to discuss the *generalized linear model*, that generalizes linear regression by allowing the *linear model* to be related to the response function via a specific link function. Then we will introduce some techniques for editing nonparametric functions, the so-called smoothing splines, which create approximate functions to capture important patterns in the data. Furthermore, we are going to discuss the *generalized additive model*, as introduced by Hastie and Tibshirani [1987], is a generalized linear model with a linear predictor including a sum of smooth functions of covariates. At last we are going to introduce the *generalized additive mixed models*, which we will combine its definition with the definition of *STERGM* to come to our final model approach the so-called *Separable Temporal Logistic Additive Mixed Models*.

5.1 The Logit Model

In our context we are interested in binary response variable Y_{ij}, for which an edge exists or does not exist between two actors. We denote the random variable with capital letters, while the specific realizations are denoted by lower-case characters. Binary logistic regression is a type of regression analysis where the response variable is a dummy variable [0: no edge between two actors ; 1: an edge between two actors]. The aim of binary regression is to model and estimate the effects of given covariates X_{ij}.

Why we should not use for the modeling the probability of the occurence of an edge with the *linear model*?

Consider the *linear model*:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + \epsilon_{ij} \tag{6}$$

with the linear predictor:

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} = x'_{ij}\beta \tag{7}$$

where:

- y_{ij} is the response dummy variable, $= 1$ edge occurs, $= 0$ if not
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the coefficients on the covariates
- x_{ij} are the covariates
- $\epsilon_{ij} \sim N(0, \sigma^2)$ is the error term

Use of the *linear model* generally gives us the correct answers in terms of the sign and significance level of the coefficients. The predicted probabilities from the model are usually where we run into trouble. There are 3 problems with using the *linear model*:

1. The error terms are heteroscedastic (heteroscedasticity occurs when the variance of the response variable is different with different values of the independent variables): $var(\epsilon_{ij}) = \pi_{ij}(1 - \pi_{ij})$, where π_{ij} is the probability that $event = 1$. Since π_{ij} depends on x_{ij} the *classical regression assumption* that the error term does not depend on the x_{ij} is violated
2. ϵ_{ij} is not normally distributed because π_{ij} takes only two values, violating another *classical regression assumption*
3. The predicted probabilities can be greater than 1 or less than 0, which can be a problem if the predicted values are used in a subsequent analysis. This amounts to an interpretation that a high probability of the *event* occuring is considered a sure thing

The *logit model* solve these problems, the common way to fit a model with binary response is to link the probability π_{ij} to the linear predictor η_{ij} through

$$\begin{aligned} \pi_{ij} &= \mathbb{P}(y_{ij} = 1 | x_{ij1}, \ldots, x_{ijp}) \\ &= h(\eta_{ij}) = h(\beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp}) \end{aligned} \tag{8}$$

where $h(.)$ is the *response function* with a distribution from the exponential family, such that for any β and any x_{ij} one gets $h(\eta) \in [0, 1]$. Since we can assume that $h(.)$ is strictly monotonically increasing function, so there exists an inverse function $g(.) = h^{-1}(.)$, called a *link function*, and can be written as

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

choosing the logistic distribution function

$$F(\eta) = \frac{exp(\eta)}{1 + exp(\eta)}$$

we get the logit model

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

which yields, for the link function

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

by multiplying the $g(.)$ function with the exponential function yields

$$\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = exp(\beta_0) \cdot exp(\beta_1 x_{ij1}) \cdots exp(\beta_p x_{ijp}) \tag{9}$$

This expression defines a multiplicative model for the odds. For example, if we were to change the j-th predictor by one unit while holding all other variables constant, we would multiply the odds by $exp(\beta_j)$. To see this point suppose the linear predictor is $x'_{ij}\beta$ and we increase x_{ij} by one, to obtain $x'_{ij}\beta + \beta_j$. Exponentiating we get $exp(x'_{ij}\beta)$ times $exp(\beta_j)$. Thus, the exponentiated coefficient $exp(\beta_j)$ represents an odd ratio. Translating the results into multiplicative effects on the odds, or odds ratios, is often helpful, because we can deal with a more familiar scale while retaining a relatively simple model.
Solving for the probability function of Y_{ij} we get

$$\mathbb{P}(Y_{ij} = 1 | X_{ij} = x_{ij}) = h(\eta_{ij}) = \frac{exp(\eta_{ij})}{1 + exp(\eta_{ij})} \tag{10}$$

Then, we can compute the odds of occurence of edge (i, j), conditional on the covariates:

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

This implies the equation:

$$logit(\mathbb{P}(Y_{ij} = 1|X_{ij} = x_{ij})) = \eta_{ij} \tag{11}$$

Thus, the effect of the j-th predictor on the probability π_{ij} depends on the coefficient β_j and the value of the probability.

5.2 Smoothing Splines

In this section we will introduce some techniques for editing nonparametric functions. It is obvious to assume that the relation between response and the covariates will not be linear. Therefore, we will need some nonparametric functions, such as smoothing splines 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 covariates $X_{ij1}, \ldots, X_{ijp}$. Even though there are several smoothing technique approaches, in this thesis we will discuss the techniques of spline smoothers, such as the *Polynomial Splines*, *B-Splines*, and *P-Splines*. This section is mainly based on Fahrmeir et al. [2013], Hastie and Tibshirani [1987], and Wood [2011].

5.2.1 Polynomial Splines

We assume by a given data in the form of (y_{ij}, x_{ij}), $i, j = \{1, \ldots, N_V\}, i \neq j$, where y_{ij} are the observations of the response variable and x_{ij} are the corresponding metric covariates. Taking y_{ij} as dyads in a network on N_V nodes we obtain $N = N_V^2 - N_V$ observations. We can assume that the response variable can be described by a function $f(\cdot)$ and an error term ϵ_{ij}.

$$y_{ij} = f(x_{ij}) + \epsilon ij \tag{12}$$

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

$$f(x_{ij}) = \gamma_0 + \gamma_1 x_{ij} + \ldots + \gamma_l x_{ij}^l$$

where $l \in \mathbb{N}$ and $\gamma_k \in \mathbb{R}$, $k \in \{0, \ldots, l\}$. This approach can be realized by the least square method. However, in most cases a pure polynomial approach does not provide satisfying results. An example illustrating this approach is given in Fahrmeir et al. [2013]. While polynomials with low degrees do not capture the true relation of the data sufficiently,

polynomials with high degrees provide wiggly fits of the data.

In order to find a way out of this dilemma, Fahrmeir et al. [2013] suggest to divide the codomain into m parts $c = k_0 < \cdots < k_m = d$, and capture the relation between the x and y on each interval $[k_j, k_{j+1})$, $j \in \{0, \ldots, m-1\}$ with a l-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 $[k_j, k_{j+1})$ but still provide continuous transitions will be introduced in the next chapters.

5.2.2 B(asic)-Splines

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 f B-splines is a construction to guarantee that piecewise estimated functions on knots $k_1, \ldots, k_{m-1}$ are composed in a sufficient, $(l-1)$-times differentiable way. In order to estimate $f(.)$ with B-splines, we have to represent the smooth function in such a way that $y = f(x) + \epsilon$ becomes a linear model. This is done by choosing specific *basic functions* $B_1(x), \ldots, B_d(x)$, $d = m + l - 1$. Thus, we can write

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x) \tag{13}$$

B-splines are defined as non-zero functions on only a few intervals $[k_i, k_p]$, $i, p \in \{0, \ldots, m\}$, with $i \neq p$. Which results good numerical properties Fahrmeir et al. [2013]. Let

$$B_j(x) = \begin{cases} f(x) & , if\ x \in [k_j, k_{j+l+1}) \\ 0 & , else \end{cases}$$

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

$$\sum_{j=1}^{d} B_j(x) = 1$$

The function $f(.)$ is composed of $l + 1$ polynomial pieces of degree l, which are put together in a $l - 1$-times differentiable way. Figure 8 illustrates single B-Spline

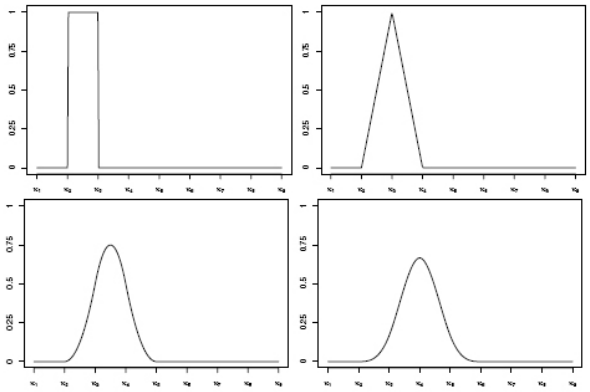


Figure 8: One single B(asic)-Spline basis function of degree $l = 0, 1, 2, 3$ at equidistant knots illustrated by Fahrmeir et al. [2013]

basis function of degree $l = 0, 1, 2, 3$ on equidistant knots as the results from these considerations.

All B-Spline basis functions are built for visualization of polynomial splines based on the underlying knots. The complete B-Spline basis of degree $l = 0, 1, 2, 3$ are depicted at equidistant knots in Figure 9.

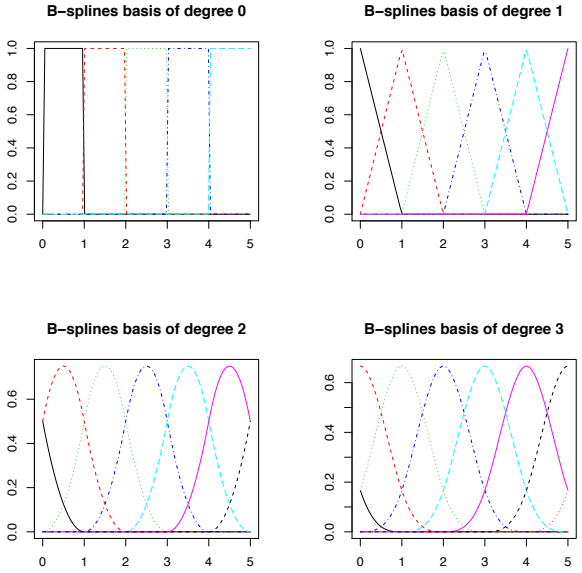


Figure 9: B(asic)-Spline basis function of degree $l = 0, 1, 2, 3$ at equidistant knots

By looking at the basis functions in 9 with a single B-spline degree, we can verify the

actual definition of linear B-spline basis functions

$$B_j^1(x) = \frac{x - k_j}{k_{j+1} - k_j}\mathbb{I}_{[k_j, k_{j+1}]}(x) + \frac{k_{j+2} - x}{k_{j+2} - k_{j+1}}\mathbb{I}_{[k_{j+1}, k_{j+2}]}(x)$$

where the 1 in $B_j^1(x)$ points out the linear form of the piecewise defined polynomials. We van come to the conclusion that $B_j^1(x)$ consists of two linear pieces.
In general, B-splines basis functions for higher degrees can be defined recursively

$$B_j^l(x) = \frac{x - k_j}{k_{j+1} - k_j}B_j^{l-1}(x) + \frac{k_{j+l+2} - x}{k_{j+l+2} - k_{j+1}}B_{j+1}^{l-1}(x)$$

Due to the linear form of (13) we can define X and γ as

$$X = \begin{bmatrix} B_1(x_{12}) & \dots & B_d(x_{12}) \\ \vdots & \dots & \vdots \\ B_1(x_{(N_V-1)N_V}) & \dots & B_d(x_{(N_V-1)N_V}) \end{bmatrix}, \gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_d \end{bmatrix}$$

thus, we can write (12) in linear form

$$y = X\gamma + \epsilon \tag{14}$$

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

$$\hat{\gamma} = (X'X)^{-1}X'y \tag{15}$$

As already mentioned above, the design matrix X holds some beneficial characteristics. The most important one stermsfrom the local definition of the basis functions, which mainly yield matrix entries of 0. The only non-zero entries occur along the diagonal of the matrix. These kind of matrices are helpful, since solving (15) with these matrices is numerically efficient.

However, the parameter vector γ can not be interpreted in a reasonable way. Instead, we are interested in the form of the estimated function $\hat{f}(.)$, which is a result of $\hat{\gamma}$

$$\hat{f}(x) = B\hat{\gamma}$$

where $B = (B_1(x), \dots, B_d(x))$
There are several reasons why B-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, but how can we control the

wiggliness of a smoother? A common technique is by controlling the degree of smoothing by *penalized B-splines*.

5.2.3 P(enalized)-Splines

P(enalized)-splines differ from the methods discussed in the previous section, since instead of minimizing

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

we are going to minimize

$$||y - B\gamma||^2 + \lambda \int_C f''(x)^2 dx \tag{16}$$

with regard to γ, where C is the codomain of x and $f''(x)$ is the second derivative of function $f(x)$. The second derivative of a function yields information about a functions curvature see O'Sullivan [1986], and therefore by minimizing (16) 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 wiggly models, $\lambda \to \infty$ leads to the linear regression of the data. See section 5.2.4 for the discussion finding a fitting smoothing parameter λ, but for now we treat λ as given.

We can show that we can write the penalty in (16) as

$$\int_C f''(x)^2 dx = \gamma^T S \gamma \tag{17}$$

where $S \in \mathbb{R}^{d \times d}$ is a matrix that can be expressed by the basis functions $B_j(x)$. Recall that we define function $f(x)$ as

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x)$$

which yields

$$f''(x) = \gamma^T B''(x)$$

for the second derivative. Since $f''(x)$ is a scalar and scalars are their own transpose we can write

$$\begin{aligned}\int_C f''(x)^2 dx &= \int \gamma^T B_j''(x) B_j''(x)^T \gamma \, dx \\ &= \gamma^T \underbrace{\int_C B_j''(x) B_j''(x)^T dx}_{:=\mathrm{S}} \gamma\end{aligned}$$

As a consequence, instead of minimizing (16) we can minimize

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

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

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

and further we get for the first and second derivation

$$\frac{\partial LS(\gamma)}{\partial \gamma} = -2B'y + 2B'B\gamma + 2\lambda S\gamma \tag{18}$$

$$\frac{\partial^2 LS(\gamma)}{\partial \gamma \partial \gamma'} = 2B'B + 2\lambda S \tag{19}$$

$B'B + \lambda S$ is positive definite and therefore invertible see Fahrmeir et al. [2013]. We get a solution to our minimization problem by solving equation (18) for γ, and this yields the least square estimator for γ

$$\hat{\gamma} = (B'B)^{-1} B'y \tag{20}$$

However these derivatives lead to rather complex systems of equations. Eilers and Marx [1996] suggest a simple approximation of the derivatives, which can be used for the construction of the penalty terms. Instead of (16) we are going to minimize

$$||y - B\gamma||^2 + \lambda \sum_{j=3}^{d} (\Delta^2 \gamma_j)^2 \tag{21}$$

Besides easy computation, this approach has the advantage of being able to penalize linear B-splines in a reasonable way. The spline functions estimated in section 6 are going to apply this approximation.
After having discussed how to estimate γ, we will discuss in the next section, how we can establish an appropriate smoothing parameter λ.

5.2.4 Choice of the Smoothness Parameter λ

The optimal choice of the smoothness parameter λ is an important aspect. The smoothness parameter λ controls the smoothness of estimated functions and ensures a suitable compromise between bias and variability of an estimator. For $\lambda \to \infty$ exists a widely linear estimation of the function $f(x)$. Contrary to $\lambda \to 0$ exists a quite rough estimation of the function $f(x)$.

The problem occurs that bias and variability of a smoothness method are simultaneously depended on the smoothness parameter λ and both cannot be minimized at the same time. Therefore, a suitable equalization must be found.

On the one hand, the Mean Squared Error (MSE) is a good possibility:

$$\begin{aligned} MSE(\hat{f}(x)) &= \mathbb{E}\Big[\Big(\hat{f}(x) - f(x)\Big)^2\Big] \\ &= \underbrace{\Big(\mathbb{E}\Big[\hat{f}(x) - f(x))\Big]\Big)^2}_{\text{bias}} + \underbrace{Var(\hat{f}(x))}_{\text{variability}} \end{aligned}$$

The MSE is added additively by the squared bias and the variance. Finally, the λ is taken where the MSE is minimal.

On the other hand, there is the Cross-Validation (CV) to find the optimal smoothness parameter λ. Respectively one observation is deleted in cross validation. Within the next step the smoothness parameter λ is estimated with the remaining $n-1$ observations. Finally, $f(x_i)$ is predicted for the deleted observation. Denoted by $\hat{f}^{(-i)}(x_i)$ is the estimation, which occurs without the observation (x_i, y_i) and receives Cross-Validation criterion Stone [1974]:

$$CV = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \hat{f}^{(-i)}(x_i)\Big)^2$$

The minimization of the CV criterion leads in the sense of prediction error to an optimal λ.

A further alternative method to achieve the optimal smoothness parameter λ is by the Akaikes Information Criterion (AIC) Akaike [1974]:

$$AIC = n.log(\hat{\sigma}^2) + 2(df + 1)$$

where $\hat{\sigma}^2 = \sum(y_i - \hat{f}(x_i)/n$. The AIC has to be minimized concerning the smoothness parameter.

5.3 The Additive Model

The class of *additive models* is an extension of linear models, where the 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

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + \epsilon_{ij}$$

with parameters $\beta_0, \ldots, \beta_p$ and i.i.d. $\epsilon_{ij} \sim N(0, \sigma^2)$. For additive models, we generalize the linear predictor with smooth functions $f(.)$

$$y_{ij} = \beta_0 + f_1(x_{ij1}) + \ldots + f_q(x_{ijq}) + \epsilon_{ij} \tag{22}$$

Since the arms trade model will also include binary covariates as well, which do not have to be smoothed, we can rewrite our equation above as

$$y_{ij} = f_1(x_{ij1}) + \ldots + f_q(x_{ijq}) + Z_{ij}\beta + \epsilon_{ij} \tag{23}$$

where $\beta' = (\beta_1, \ldots, \beta_p)$ is a vector of parameters and $Z_{ij} = (x_{ij1}, \ldots, x_{ijp})$ 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, 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.

5.4 The Generalized Additive Model

A generalized additive model Hastie and Tibshirani [1987] is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. 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. Therefore, we extend the linear predictor (7) with smooth functions $f_1(.), \ldots, f_q(.)$ to

$$\begin{aligned} y_{ij} = \eta_{ij} &= f_1(x_{ij1}) + \ldots + f_q(x_{ijq}) + \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_k x_{ijk} + \epsilon_{ij} \\ &= f_1(x_{ij1}) + \ldots + f_q(x_{ijq}) + \eta_{ij}^{lin} + \epsilon_{ij} \\ &= \eta_{ij}^{add} + \epsilon_{ij} \end{aligned} \tag{24}$$

where $f_1(x_{ij1}), \ldots f_q(x_{ijq})$ are smooth functions of the metric covariates $x_1, \ldots, x_q$, and the errors ϵ_{ij} are independent of the x_{ij}, with $\mathbb{E}(\epsilon_{ij}) = 0$ for all $N_V^2 - N_V$ observations. $\beta' = (\beta_1, \ldots, \beta_p)$ and $Z_{ij} = (x_{ij1}, \ldots, x_{ijp})$ are defined as in the previous section. The smooth functions are estimated in a nonparametric fashion. A generalized additive model differs from an additive model. Its additive predictor is linked with the expected value by a known smooth monotonic link-function.

The smooth functions $f_1(x_{ij1}), \ldots f_q(x_{ijq})$ are represented by a linear combination of basic-functions

$$f_j = \sum_{l=1}^{dj} \gamma_{jl} B_l(x_j), \quad j = 1 \ldots q$$

There are different types of basic-functions for $B_l, l = 1 \ldots d_j$. Common examples are B-Splines or P-Splines see Section 5.2.

A covariate can always be represented by

$$f_j = Z_j \gamma_j$$

with the coefficient vector $\gamma_j = (\gamma_1 \ldots \gamma_j)$ and the design matrix Z_j. The additive model in matrix notation

$$y = Z_1 \gamma_1 + \ldots Z_q \gamma_q + X\beta + \epsilon$$

The estimation occurs either with the penalized least squares criterion for normal distributed response

$$PKQ(\lambda) = (y - Z\gamma)^T (y - Z\gamma) + \lambda \gamma^T K \gamma$$

Thereby denotes Z a matrix whose entries are the basic-functions evaluated at the observations

$$Z = \begin{pmatrix} B_1^l(z_1) & \ldots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \ldots & B_d^l(z_n) \end{pmatrix}.$$

Simple GAMs are estimated with the penalized least-squares estimator

$$\hat{\gamma} = (Z^T Z + \lambda K)^{-1} Z^T y$$

or with the Fisher Scoring algorithm Fahrmeir et al. [2013]. Generalized additive models require more complex methods as the backfitting algorithm Hastie and Tibshirani [1987]. For a more detailed overview of generalized additive models see Hastie and Tibshirani [1987] and Fahrmeir et al. [2013].

5.5 The Generalized Additive Mixed Model

Before we introduce *the generalized additive mixed model* (GAMM), we will give a short introduction to *linear mixed models* (LMM), and also to *generalized linear mixed models* (GLMM) for a better understanding. This Section is mainly based on Wood [2006].

5.5.1 The Linear Mixed Model

The linear mixed model can conveniently be written as

$$y = XB + Zb + \epsilon, \quad b \sim N(0, \psi_\theta), \quad \epsilon \sim N(0, \Lambda\sigma^2) \tag{25}$$

where ψ_θ is a positive definite covariance matrix for the random effects b, and Z is a matrix of fixed coefficients describing how the response variable, y, depends on the random effects. ψ_θ depends on some parameters, θ, which will be the prime target of statistical inference about the random effects. Finally, Λ is a positive definite matrix which usually has a simple structure depending on few or no unknown parameters.

We could combine the residual vector and random effects into a single, non-independent, variable-variance residual vector, $e = Zb + \epsilon$. It is obvious that e is a zero mean multivariate normal vector, and its covariance matrix must be $Z\psi_\theta Z^T + I\sigma^2$. Hence (25) can be written as:

$$y = X\beta + e, \quad e \sim N(0, \Sigma_\theta \sigma^2)$$

where $\Sigma_\theta = Z\psi_\theta Z^T/\sigma^2 + I$, and the subscript, θ, emphasizes the dependence of Σ_θ on the covariance parameter vector, θ. So if θ were known then we could estimate β using the methods of *least squares criterion*.

5.5.2 The Generalized Linear Mixed Model

The generalized linear mixed model (GLMM) follow from *linear mixed model*. Let $\mu^b \equiv \mathbb{E}(y|b)$. Then a GLMM has the form

$$g(\mu^b) = X_i\beta + Z_i b, \quad b \sim N(0, \psi_\theta) \quad \textit{and} \quad y_i|b \sim \textit{exponential family distribution}$$

where g is a monotonic link function, and the covariance matrix, ψ_θ, of the random effects, is parametrized in terms of a parameter vector θ. The $y_i|b$ are independent.

The likelihood for a GLMM is obtained by considering the joint distribution of the response and the random effects:

$$f_{\beta,\theta,\phi}(y, b) \propto |\psi_\theta|^{-1/2} exp\left(log\, f(y|b) - \frac{1}{2} b^T \psi_\theta^{-1} b\right)$$

where $f(y|b)$ is the joint distribution of the response conditional on the random effects. The marginal distribution of y, and hence the likelihood, is obtained by integrating out the random effects

$$L(\beta, \theta) \propto |\psi_\theta|^{-1/2} \int exp\left(l(\beta, b) - \frac{1}{2} b^T \psi_\theta^{-1} b\right) db$$

where $l(\beta, b)$ is considered as function of β and b the likelihood of the GLM that would result from treating both β and b as fixed effects.

5.5.3 The Generalized Additive Mixed Model

A GAMM is just a GLMM in which part of the linear predictor is specified in terms of smooth functions of covariates. Extending the Equation (24) we get for the generalized additive mixed model the following form

$$y_{ij} = \eta_{ij}^{lin} + f_1(x_{ij1}) + \ldots + f_q(x_{ijq}) + Z_i b + \epsilon_{ij} \tag{26}$$

where y_{ij} is the response variable, η_{ij}^{lin} is the linear predictor $\beta_0 + \beta_1(x_{ij1}) + \ldots \beta_p(x_{ijp})$ with $\beta' = (\beta_0, \ldots, \beta_p)$ vector of fixed parameters, and $x_{ij1}, \ldots, x_{ijp}$ the covariates, which are assumed to be linear. $f_1, \ldots, f_q$ are smooth functions of the metric covariates $x_{ij1}, \ldots, x_{ijq}$, Z_i is a row of a random effects model matrix, $b \sim N(0, \psi_\theta)$ is a vector of random effects coefficients, with unknown positive definite covariance matrix ψ_θ, with parameter θ, $\epsilon \sim N(0, \Lambda)$ is a residual error vector, with i^{th} element ϵ_i, and the covariance matrix Λ, which is usually assumed to have some simple pattern.

The generalization from GLMs to GAMs required the development of theory for penalized regression as described in Section 5.2, in order to avoid overfitting, but GLMM methods require no adjustment in order to cope with GAMMs, it is possible to write any of the penalized regression smoothers considered, as components of a mixed model, while treating their smoothing parameters as variance component parameters, to be estimated by Likelihood methods.

6 Modelling of Arms Trade Networks

To model the international major conventional weapons trade network, we have yearly, pairwise, binary trading information of major conventional weapons and, additionally, covariate informations for 218 countries from $1950 - 2013$ available. Given the trade data, it is possible to evaluate the (*indirectly* observed) formation y^+ and dissolution y^- networks for the years $1951 - 2012$. The formation networks can be evaluated as $y_{ij}^{+t} = y_{ij}^t | y_{ij}^{t-1}$ and the dissolution networks as $y_{ij}^{-t} = y_{ij}^t \cap y_{ij}^{t-1}$. To estimate the models, it is convenient to create two long response vectors by connecting all yearly, rowwise stacked, binary formation / dissolution adjacency matrices (removing the missing entries of the main diagonals). For modelling the formation and dissolution processes, we condition on the previous observed trading networks and therefore, it is possible to compute arbitrary statistics of the previous observed networks for each trading dyad and use them as covariates in the logistic regression models. Focusing only on the network y^{t-1} for modelling the formation and dissolution networks at timepoint t (which equals a Markov-1 assumption see Section 4.2.1) some reasonable statistics for modelling y_{ij}^{+t} or y_{ij}^{-t} are e.g. the lag-1 trading information of actor i to actor j, i.e. y_{ij}^{t-1}, the lag-1 reciprocal trade information y_{ji}^{t-1}, the number of export activities (the *out-degree*) of actor i in the last year, i.e. $\sum_k y_{ik}^{t-1}$, and the number of import activities of actor j in the last year $\sum_k y_{kj}^{t-1}$. Those statistics are often referred to as (lag-1) *two-node* statistics, since they are based on counts or the existence of (specific) pairwise trade patterns. Furthermore, it is possible to include lag-1 *three-node* statistics as covariates in the model: We are going to consider the number of transitive ("indirect") trade relations between two nations in the last year, i.e. $\sum_k y_{ik}^{t-1} y_{kj}^{t-1}$, the number of last years reverse transitive trade relations, $\sum_k y_{jk}^{t-1} y_{ki}^{t-1}$, the number of shared suppliers in the last year, $\sum_k y_{ki}^{t-1} y_{kj}^{t-1}$, and the number of shared customers in the last year, $\sum_k y_{ik}^{t-1} y_{jk}^{t-1}$. Figure 10 illustrates these described patterns, for each trading dyad (i, j) we count their observed numbers at timepoint $t-1$ and used them as a covariates for modelling the trading dyad at timepoint t.

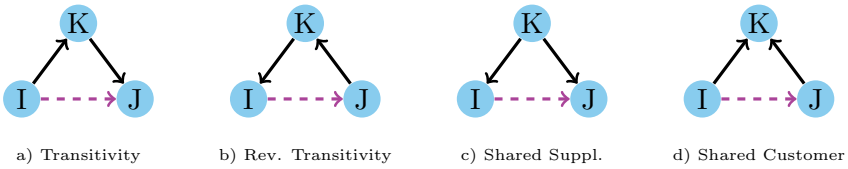


Figure 10: Three-Node Statistics

Additionally, it is possible to include sender- and receiver-specific covariates, e.g. the log-GDPm or polity scores of the countries, as well as binary information whether the sender and / or receiver have defence agreements in the specific years. A third type of covariates are observation (or dyad-) specific information, e.g. the year of the specific trade observations. Table 6 presents all covariates used in our modelling of the small arms trade formation and dissolution networks and their specific types.

Description	Abbreviation	Type
lag-1 trade y_{ij}^{t-1}	trade.bin.lag	binary
lag-1 reciprocal trade y_{ji}^{t-1}	trade.bin.recip.lag	binary
lag-1 defence alliances y_{ij}^{t-1}	daml	binary
intra-state conflict sender in year t	coml.sender	quasi-continuous
intra-state conflict receiver in year t	coml.receiver	quasi-continuous
lag-1 number of transitive trade pattern	trade.lag.trans	quasi-continuous
lag-1 number of reverse transitive trade patterns	trade.lag.revtrans	quasi-continuous
lag-1 number of shared suppliers	trade.lag.samesource	quasi-continuous
lag-1 number of shared buyers	trade.lag.samebuyer	quasi-continuous
lag-1 in-degree receiver	in.deg.lag.rec	quasi-continuous
lag-1 out-degree sender	out.deg.lag.sen	quasi-continuous
year of observation / modelled network	year	quasi-continuous ($[1950, 2013]$)
log GDPm sender in year t	logGDPm.sender	continuous
log GDPm receiver in year t	logGDPm.receiver	continuous
MC military capability of sender	mc.sender	continuous
MC military capability of receiver	mc.receiver	continuous
absolute difference of polity score of sender and receiver in year t	abs.polity.diff	quasi-continuous ($[0, 20]$)

Table 6: Covariates used in the formation and dissolution model, their abbreviations and types

As discussed before, we are going to use Mixed Additive Logistic Regression for analyzing the formation and dissolution processes in the international major conventional weapons trade networks from 1951 until 2012. The basic idea of generalized additive models is to extend the classic framework of generalized linear models by estimating effects of continuous covariates nonparametrically. Instead of assuming linear effects for the continuous covariates in the linear predictor, nonlinear, smooth effects are estimated for (at least some of) those continuous covariates, typically via a representation of the effects through smoothing splines. This already leads to a very versatile class of models, however, it is, like in the case of generalized linear models, possible to additionally incorporate random effects into models to account for e.g. a longitudinal or clustered structure in the data. Such models are called generalized additive mixed models (*GAMMs*), see Section 5.5 or Wood [2006].

In our application case of modelling the entries of the formation and dissolution adjacency matrices of the international MCW trades, we are going to estimate two seperate mixed additive logistic regression models with dummy effects for each binary, and non-linear, smooth effects for each (quasi-)continuous covariate from Table 6. Additionally, we incorporate gaussian random intercepts for the specific sender and receiver countries of the modelled trading dyads. Estimating effects for the lag-network statistics, thereby, aims at considering the special network structure of the data, while the sender- and receiver-specific covariates are included to analyze potential relations between the observed countries' characteristics and the formation and dissolution processes. By, including additional random intercepts for sender and receiver countries, we take the countries' different degrees of involvement into the international MCW trade networks into account.

The covariate *lag-1 trade* plays a somewhat special role in the formation and dissolution models. Due to the definition of the formation and dissolution networks, the entries of the formation and dissolution adjacency matrices are already predetermined for one of the groups of the (binary) covariate, each: The formation network y^{+t} at timepoint t was defined as the network that contains all edges of the observed network $\mathbf{y}^{t-1}$ and additionally all the newly formed edges at timepoint t. Consequently, all entries of the formation models' adjacency matrix y_{ij}^{+t} with lag-1 trade, i.e. $y_{ij}^{t-1} = 1$, are per definition 1. The dissolution model, by contrast, was defined as the network that consists out of the edges of the observed network $\mathbf{y}^{t-1}$, but with all ties that are dissolved at timepoint t removed. All entries y_{ij}^{-t} of the dissolution networks' adjacency matrix at timepoint t without lag-trade, i.e. with $y_{ij}^{t-1} = 0$, are consequently predetermined to be 0. Instead of focusing on all entries of the years' formation and dissolution adjacency matrices, we can therefore focus on two distinct subsets of our dataset in modelling the formation and

dissolution processes. The inclusion of a categorical covariate into a logistic regression model, that doesn't feature observations of both response classes in one of its categories, leads to so-called *(quasi-) complete separation*. The estimation of an effect for the class that separates the response leads to huge (in absolute value) point, estimates with inflated standard errors. The estimation algorithm (in general based on likelihood optimization), doesn't converge for the specific parameter, the best possible estimate for it would be $+/-\infty$. All other estimated parameters, however, are still valid estimates for the effects of the other covariates, and describe their effects in all classes of the categorical covariate that contain both levels of the response. By anticipating the separation of the responses induced through the definition of the formation / dissolution networks, we gain stability in the estimation algorithms and achieve more clarity in the description our modelling approach. In the formation model we look at all trade relations that did not exist in the last year and model their log-odds to form, whereas, in the dissolution model we focus on all edges that did exist in the last year and model their log-odds to persist.
Summarizing it in a formal way, our assumptions in modelling the international major conventional arms trade networks from 1951 until 2012 are the following:

1. The processes of the formation of new edges from timepoint $t-1$ to t and the persistence of edges that existed at $t-1$ in t do not interact with each other, i.e. the conditional probability of observing a network $\mathbf{y}^t$ at timepoint t equals the product of the conditional probabilities of the associated formation and dissolution networks.

2. For each of the processes we assume that the relevant (for the formation network all entries Y_{ij}^{+t}, with $y_{ij}^{t-1}=0$, and for the dissolution network all entries Y_{ij}^{-t} with $y_{ij}^{t-1}=1$) entries $Y_{ijt}^{+/-}$ of the adjacency matrices are conditionally independent realizations of a Bernoulli trial with a specific, to be modelled, success probability, i.e.
$$Y_{ijt}^{+/-}|\mathbf{y}^{t-1},\, \mathbf{x}_{ijt},\, b_{sen,\,i}^{+/-},\, b_{rec,\,j}^{+/-} \overset{ind}{\sim} B(1,\pi_{ijt}^{+/-}),$$
with
$$(b_{sen,\,i}^{+/-}, b_{rec,\,j}^{+/-}) \sim N(\mathbf{0}, \mathrm{diag}(\tau_{sen}^{2\,+/-}, \tau_{rec}^{2\,+/-})),$$
and we model the log-odds of the success probabilities through a linear predictor $\eta_{ijt}^{+/-}$, i.e.
$$\log\Big(\frac{\pi_{ijt}^{+/-}}{1-\pi_{ijt}^{+/-}}\Big) = \eta_{ijt}^{+/-}.$$
The linear predictors of the formation and dissolution model contain the same covariates, presented in Table 6, but it is of course possible to estimate different

effects on the formation of new and the persistence of existing ties within the two independent models. The linear predictor of our formation model is:

$$\begin{aligned}\eta^{+}_{ijt} = {} & \beta^{+}_{0} + \beta^{+}_{recip} y^{t-1}_{ji} + \beta^{+}_{daml} y^{t-1}_{ij} + \mathrm{f}^{+}_{out.deg}\Big(\sum_k y^{t-1}_{ik}\Big) + \mathrm{f}^{+}_{in.deg}\Big(\sum_k y^{t-1}_{kj}\Big) + \mathrm{f}^{+}_{trans}\Big(\sum_k y^{t-1}_{ik} y^{t-1}_{kj}\Big) \\ & + \mathrm{f}^{+}_{rev.trans}\Big(\sum_k y^{t-1}_{jk} y^{t-1}_{ki}\Big) + \mathrm{f}^{+}_{shared.supp}\Big(\sum_k y^{t-1}_{ki} y^{t-1}_{kj}\Big) + \mathrm{f}^{+}_{shared.cust}\Big(\sum_k y^{t-1}_{ik} y^{t-1}_{jk}\Big) \\ & + \mathrm{f}^{+}_{intra.conf.sen}\big(x^{t}_{i,\,intra.conf.sen}\big) + \mathrm{f}^{+}_{intra.conf.rec}\big(x^{t}_{j,\,intra.conf.rec}\big) + \mathrm{f}^{+}_{gdpm.sen}\big(x^{t}_{i,\,gdpm}\big) \\ & + \mathrm{f}^{+}_{gdpm.rec}\big(x^{t}_{j,\,gdpm}\big) + \mathrm{f}^{+}_{mil.cap.sen}\big(x^{t}_{i,\,mc.sen}\big) + \mathrm{f}^{+}_{mil.cap.rec}\big(x^{t}_{j,\,mc.rec}\big) \\ & + \mathrm{f}^{+}_{abs.polity.diff}\big(x^{t}_{ij,\,abs.pol.diff}\big) + \mathrm{f}^{+}_{out.deg\times year}\Big(\sum_k y^{t-1}_{ik},\ t\Big) + \mathrm{f}^{+}_{in.deg\times year}\Big(\sum_k y^{t-1}_{kj},\ t\Big) \\ & + \mathrm{f}^{+}_{year}(t) + b^{+}_{sen,\,i} + b^{+}_{rec,\,j},\end{aligned}$$

and the linear predictor of the dissolution model is:

$$\begin{aligned}\eta^{-}_{ijt} = {} & \beta^{-}_{0} + \beta^{-}_{recip} y^{t-1}_{ji} + \beta^{-}_{daml} y^{t-1}_{ij} + \mathrm{f}^{-}_{out.deg}\Big(\sum_k y^{t-1}_{ik}\Big) + \mathrm{f}^{-}_{in.deg}\Big(\sum_k y^{t-1}_{kj}\Big) + \mathrm{f}^{-}_{trans}\Big(\sum_k y^{t-1}_{ik} y^{t-1}_{kj}\Big) \\ & + \mathrm{f}^{-}_{rev.trans}\Big(\sum_k y^{t-1}_{jk} y^{t-1}_{ki}\Big) + \mathrm{f}^{-}_{shared.supp}\Big(\sum_k y^{t-1}_{ki} y^{t-1}_{kj}\Big) + \mathrm{f}^{-}_{shared.cust}\Big(\sum_k y^{t-1}_{ik} y^{t-1}_{jk}\Big) \\ & + \mathrm{f}^{-}_{intra.conf.sen}\big(x^{t}_{i,\,intra.conf.sen}\big) + \mathrm{f}^{-}_{intra.conf.rec}\big(x^{t}_{j,\,intra.conf.rec}\big) + \mathrm{f}^{-}_{gdpm.sen}\big(x^{t}_{i,\,gdpm}\big) \\ & + \mathrm{f}^{-}_{gdpm.rec}\big(x^{t}_{j,\,gdpm}\big) + \mathrm{f}^{-}_{mil.cap.sen}\big(x^{t}_{i,\,mc.sen}\big) + \mathrm{f}^{-}_{mil.cap.rec}\big(x^{t}_{j,\,mc.rec}\big) \\ & + \mathrm{f}^{-}_{abs.polity.diff}\big(x^{t}_{ij,\,abs.pol.diff}\big) + \mathrm{f}^{-}_{out.deg\times year}\Big(\sum_k y^{t-1}_{ik},\ t\Big) + \mathrm{f}^{-}_{in.deg\times year}\Big(\sum_k y^{t-1}_{kj},\ t\Big) \\ & + \mathrm{f}^{-}_{year}(t) + b^{-}_{sen,\,i} + b^{-}_{rec,\,j},\end{aligned}$$

in general the same, but with all effects β^{+}, f^{+}, and b^{+} replaced by their counterparts β^{-}, f^{-}, and b^{-}.

Hence, we have a class of generalized additive mixed model in the form of

$$\eta^{+/-} = \eta^{lin(+/-)} + f_1(X_1) + \ldots + f_q(X_q) + Zb \tag{27}$$

where $\eta^{+/-}$ is the response variable for the formation and the dissolution models respectively, η^{lin} is the linear predictor $\beta_0 + \beta_1(x_1) + \ldots \beta_p(x_p)$ with $\beta' = (\beta_0, \ldots, \beta_p)$ vector of fixed parameters, and $x_1, \ldots, x_p$ the covariates, which are assumed to be linear. $f_1, \ldots, f_q$ are smooth functions of the metric covariates $x_1, \ldots, x_q$, Z is a row of a random effects model matrix, $b \sim N(0, \psi_\theta)$ is a vector of random effects coefficients.

We can expand our model (27) with a special model (see Hastie and Tibshirani [1993]) in which the coefficients are allowed to change smoothly with the value of other variables,

which we call *effect modifiers*. Then, a *varying-coefficients* model with random effects has the form

$$\eta^{+/-} = \eta^{lin(+/-)} + f_{year}(T) + \delta_1(T)X_1 + \ldots + \delta_q(T)X_q + Zb \tag{28}$$

where T changes the coefficients of the metric covariates $X_1, \ldots X_q$ through the functions $\delta_1(), \ldots, \delta_q()$. The dependence of $\delta_j()$ on T implies a special kind of interaction between each T and X_j. In our case T is the metric covariate *year*.

7 Results

After introducing our proposed modelling approach, we can continue with the results of applying it to the observed major conventional arms trade data. For all our computations we used the statistical programming language **R**, version 3.3.0 [R Core Team, 2013]. In addition to its basic functionality we used some supplementary software packages for our evaluations: For the handling of network data and the computation of network statistics the package **igraph**, version 1.0.1 [Csardi and Nepusz, 2006], for the estimation of the additive mixed logistic regression models for the formation and dissolution processes the package **mgcv**, version 1.8.15 [Wood, 2011], and for the ROC-curves in subsection 7.2 the package **pROC**, version 1.8 [Robin et al., 2011].

7.1 Results Network Model

This subsection is splitted into two parts, at first we present the results of the formation model, afterwards the results of the dissolution model. In the following subsection we are going to present different ways to evaluate the model and assess its fit to the observed network data. We estimated the formation and dissolution model by using the function **bam()** of the package **mgcv**, parameter estimation and (data driven) selection of the smoothing parameters was achieved by optimizing the *Restricted-Maximum-Likelihood (REML)*. For representation of the smooth effects we used (univariate) cubic p-splines with second-order difference penalty and 30 equidistant knots. The smooth interaction surface was estimated via bivariate tensor-product p-spline with 10^2 knots, again with cubic basis functions and second-order difference penalty. To check the sensitivity of our model on the hyperparameters setting, we estimated the model with more flexible settings (more knots), however the results stay, in general, the same. We therefore conclude that the described setting provides enough flexibility to capture the general structure in data and the penalization (smoothness selection) works reliable.

7.1.1 Formation Model

Table 7 presents the estimated intercept and the dummy effects of all binary covariates in the formation model, Fig. 11 shows the estimated varying-coefficient smooth effects of the (quasi-) continuous covariates and the distributions of the predicted random intercepts. The solid lines in Fig. 11 represent the estimated smooth effects, the light-blue area represents $+/-2$ times the estimated standard errors. For the interpretation of those estimated parameters it is useful to remind oneself what we try to describe with the formation model: For each trading dyad without observed trade at timepoint $t-1$ we model the log-odds of a trade at timepoint t, i.e.

$$\frac{P(y_{ij}^t = 1)}{1 - P(y_{ij}^t = 1)}$$

through the specified linear predictor. To interpret the estimated intercept $\hat{\beta}_0^+ = -15.365$ we can use the response function of the logistic regression model:

$$\frac{1}{1 + exp(-(-15.43))} \approx 0.0000002$$

equals the predicted probability of a new tie to form with all other components of the linear predictor set to zero. This gives a reasonable first impression of the general probability of trading ties to arise, since the parametric coefficients all refer to binary covariates (and consequently the value 0 of the covariates is in the observed part of the covariate space) and the other estimated / predicted (smooth and random) effects are all zero for at least one somewhat reasonable value of the covariates.

The standard procedure of interpreting the p-th estimated (parametric) effect in logistic regression models is not to interpret the respective $\hat{\beta}_p$ directly, but $exp(\hat{\beta}_p)$ since it represents the *Odds Ratio*, the ratio of the odds of two observations that differ (only) in the p-th covariate by one unit. This can also be described as the multiplicative effect on the odds of the observation with $x_p = 0$ if x_p is instead 1, what is a useful interpretation for effects of binary covariates. Since we have random intercepts for the sender and receiver country of the trading dyads included in our models it is necessary to consider that the estimated effects in the mixed additive logistic regression model are observation-specific (in our case dyad-specific) effects, i.e. conditional on the predicted random intercepts of the respective sender and receiver countries. The correct interpretation of the estimated effect for a reciprocal trade (a trade from country j to i) at timepoint $t-1$ is therefore: Focusing on a specific trading dyad, the odds of a trade at timepoint t between country i and j are approximately $exp(0.396) \approx 1.5$ times higher if there was a trade from j to i at

timepoint $t-1$ compared to the situation in which no trade from j to i was observed (and consequently no trade at all in the dyad (i,j)). This effect has to be interpreted *ceteris paribus*, i.e. with all other covariates fixed.
For the defence alliance dummy-effect (daml) presented in Table 7 we can interpretate it in the following way: Focusing on a specific trading dyad, the odds of a trade at timepoint t between country i and j are approximately $exp(1.548) \approx 4.7$ times higher if there was a defence alliance between the countries i and j at timepoint $t-1$, compared to the situation in which no defence alliance was observed. This effect has to be also interpreted *ceteris paribus*, i.e. with all other covariates fixed.

	Estimate	Std.Error	z.value	p-value
(Intercept)	-15.43	0.408	-37.73	0.000
trade.bin.recip.lagTRUE	0.396	0.103	3.85	0.000
daml1	1.548	0.057	27.09	0.000

Table 7: Parametric Coefficients Formation Model

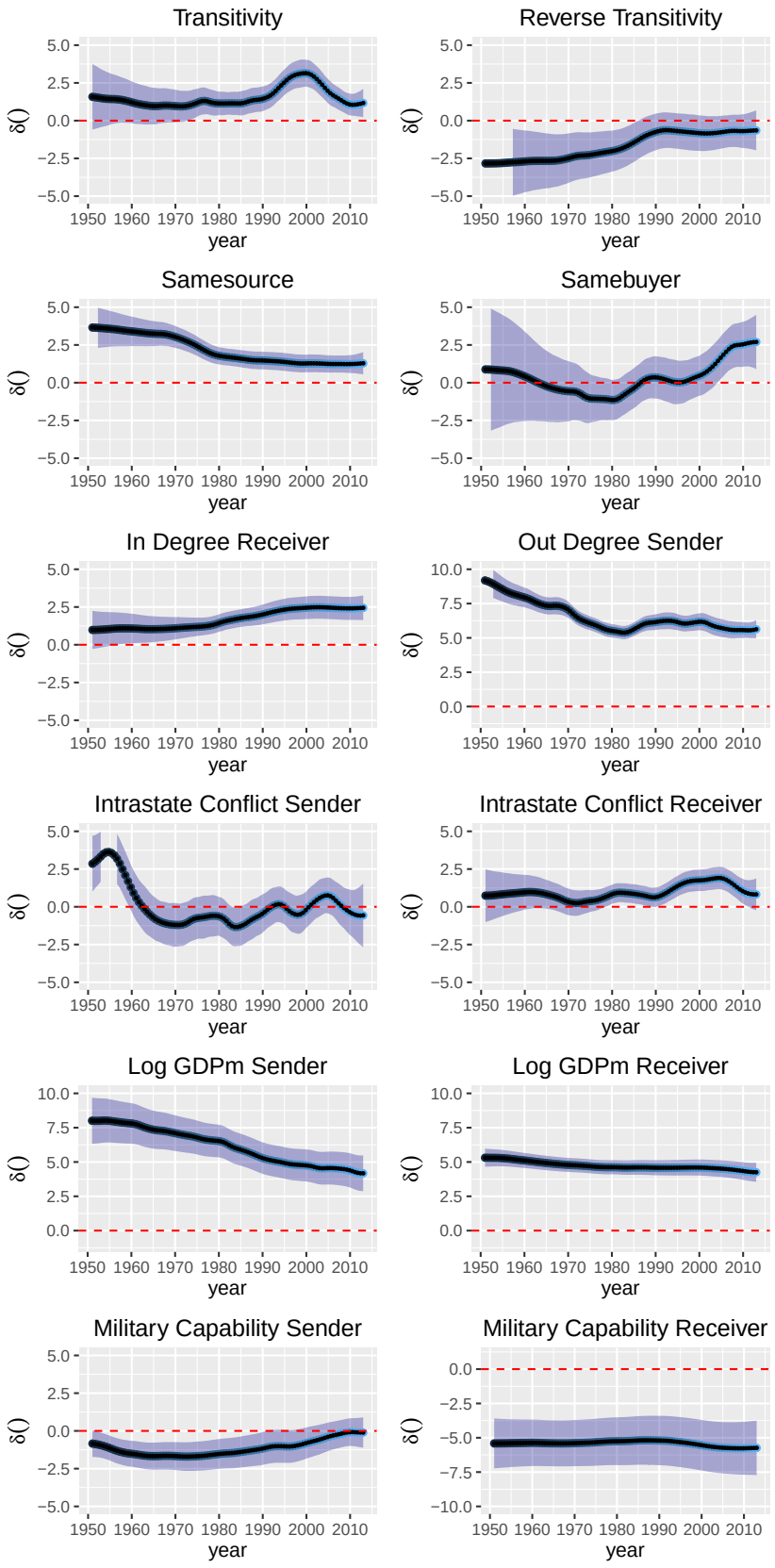
The smooth effect of *year* in Fig. 11, has to be interpreted as in the nonparametric context, i.e. the estimated odds ratio of a newly formed trade for a specific pair of sender and receiver countries for year 2013 and 1978 is approximately

$$\frac{\frac{P(trade|year=2013,\, b_{sen,\,i},\, b_{rec,\,j})}{1-P(trade|year=2013,\, b_{sen,\,i},\, b_{rec,\,j})}}{\frac{P(trade|year=1978,\, b_{sen,\,i},\, b_{rec,\,j})}{1-P(trade|year=1978,\, b_{sen,\,i},\, b_{rec,\,j})}} \approx exp(1.7-0) \approx 5.5,$$

i.e. the odds of a new trade between country i and j are ceteris paribus approximately 5.5 times higher in 2013 than in 1978.

For the interpretation of the varying-coefficient smooth effects (the first 12 plots) in Fig. 11 has to be interpreted in the following way: for a given constant value of the metric covariate (notice that the codomain of the metric covariates have to be on the same scale in order to get meaningful interpretations):

- $f_{year}(t)$ is the nonlinear effect of year
- $\delta_{ij1}(t)\, x_{ij1}$ is a function of from year varying effect, for a given constant value of x_{ij1}
- $\beta_0 + f_{year}(t_l) + \delta_{ij1}(t_l)\, x_{ij1}$ is the odds for a trading dyad in the formation model


Transitivity
Reverse Transitivity
Samesource
Samebuyer
In Degree Receiver
Out Degree Sender
Intrastate Conflict Sender
Intrastate Conflict Receiver
Log GDPm Sender
Log GDPm Receiver
Military Capability Sender
Military Capability Receiver
δ()
year
1950 1960 1970 1980 1990 2000 2010

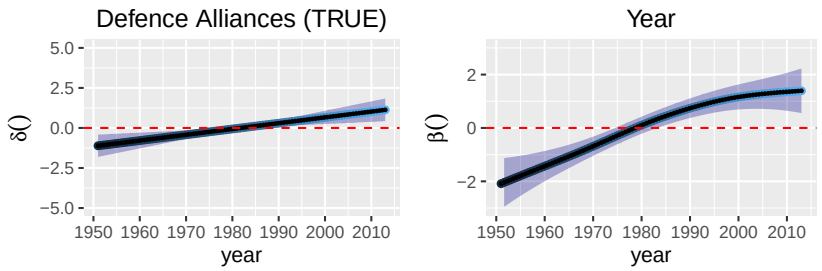


Figure 11: Varying-Coefficient Smooth Effects Formation Model

The first six smooth plots in Fig. 11 show the estimated varying-coefficient smooth effects of the quasi-continuous lag-network statistics in the formation model (see Section 6 for details).

Let us consider the first plot the *transitivity*, we see that the *transitivity* pattern of the network has a positive effect over the entire period, for example in year 1951, we have an effect of approximately $exp(1.5) \approx 4.5$ and in 2013 an effect of $exp(1.24) \approx 3.49$, i.e. the *transitivity* is positively related in 1951 and 1997 to the odds of formation (forming new trading ties), on the contrary the *reverse transitivity* is negatively related to the odds of formation in the entire period. For the covariate *samesource* we have at the beginning an effect of approximately $exp(3.5)$, and an effect of $exp(1)$ in the last period of our observation. For *samebuyer* we have positive effects from 1951 till 1960, having negative effects from 1960 till 1990, and getting positive in the last period, *in degree* of the receiver has more or less constant positive effect over the entire period, but we have a big effect of senders' *out degree* at the beginning of our period, becoming constant from 1980 till 2013, for the senders' and receivers' *intrastate conflicts* we have more or less constant effects over the entire period. For the *gdpm* of the sender and receiver we have big effects, although for the senders' *gdpm* we have a decreasing positive effect from $exp(7.5)$ in 1951 to $exp(3.5)$ in 2013. For the covariate *military capability* for both sender and receiver we have negative effects. For *defence alliances* we have a negative effect at the beginning of the period compared to *no defence alliances*, becoming slightly positive at the end of the period compared to *no defence alliances*.

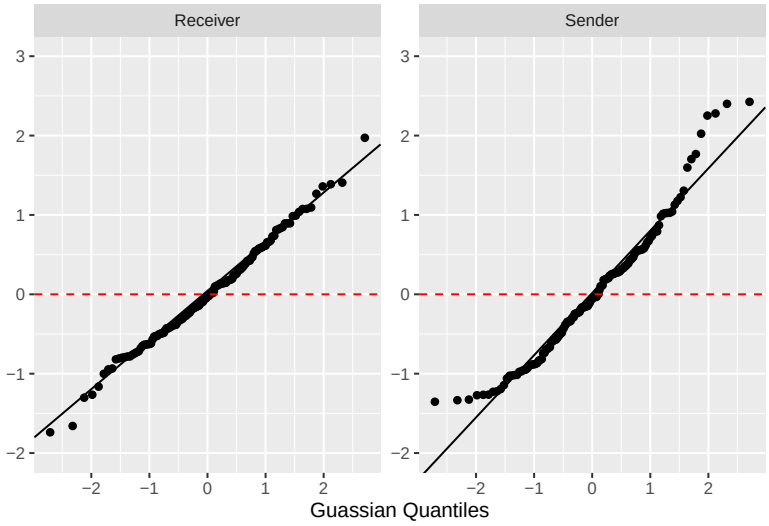


Figure 12: Guassian Quantiles for the Random Intercepts Formation Model

The presented QQ-plots of the predicted random effects look quite nice, the assumption of gaussian random intercepts for sender and receiver countries seems to be quite reasonable.

7.1.2 Dissolution Model

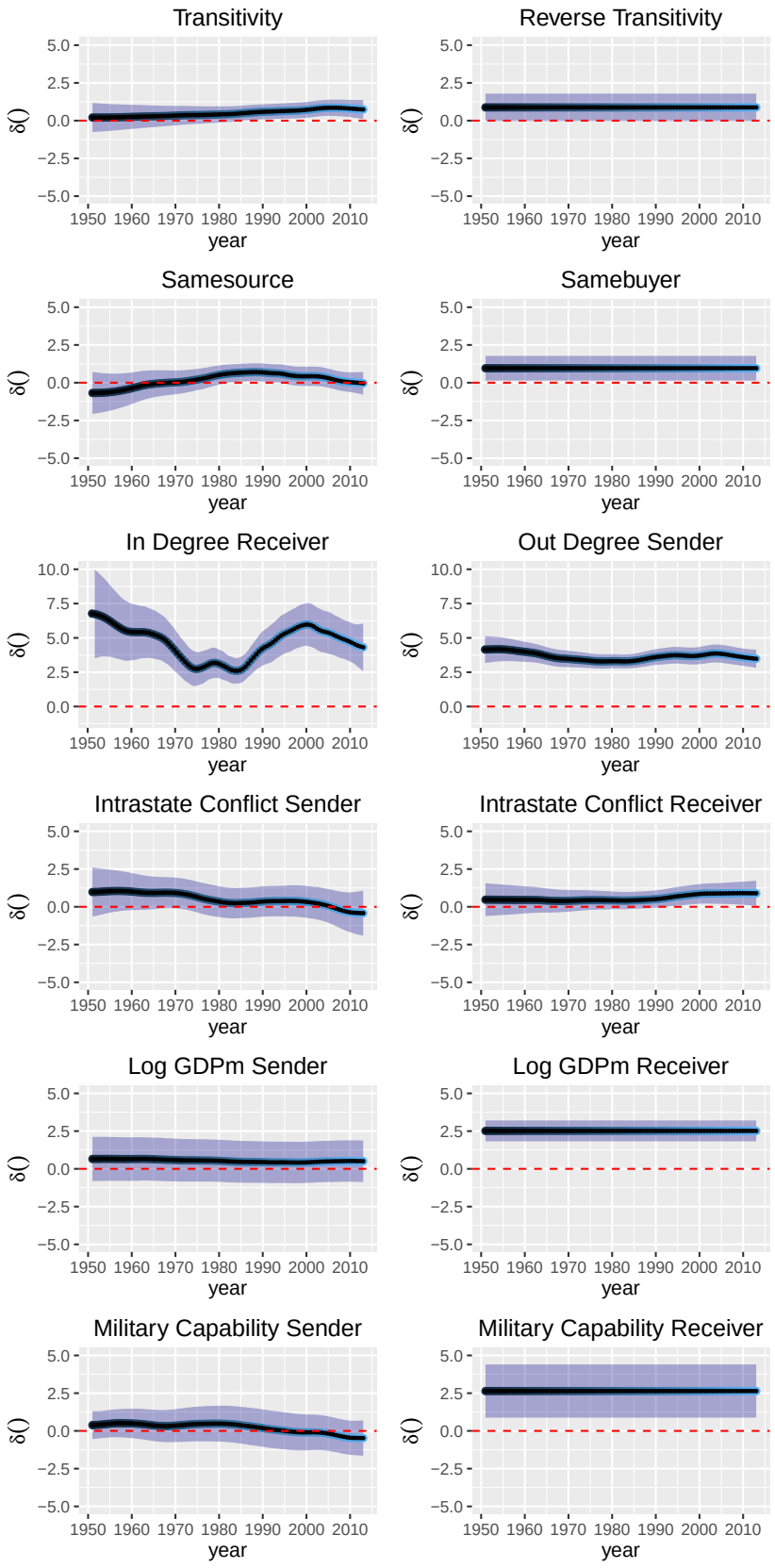
Table 8 presents the estimated intercept and the dummy effects of all binary covariates, while Fig. 13 shows the estimated coefficient-varying smooth effects and Fig. 14 the distribution of the predicted random intercepts in the dissolution model. For interpretation purposes it is again helpful to remind what is being modelled with the dissolution model: For each trading dyad with an observed trade relation in $t-1$ we are modelling the odds of the trade relation to persist at timepoint t.

	Estimate	Std.Error	z.value	p-value
(Intercept)	-3.380	0.542	-6.228	0.000
trade.bin.recip.lagTRUE	0.181	0.109	1.658	0.05
daml1	0.445	0.069	6.408	0.000

Table 8: Parametric Coefficients Dissolution Model

The estimated intercept $\hat{\beta}_0^- = -3.38$ indicates a quite low probability of trading relations to persist, with all other terms of the linear predictor being zero, it equals $\frac{1}{1+exp(-3.38)} \approx 0.967$.
Focusing on a specific trading dyad, the odds of a trade to persist at timepoint t between country i and j are approximately $exp(0.181) \approx 1.19$ times higher if there was a trade from j to i at timepoint $t-1$ compared to the situation in which no trade from j to i was observed. This effect has to be interpreted *ceteris paribus*, i.e. with all other covariates fixed.
Furthermore, the odds of a trade to persist at timepoint t between country i and j are approximately $exp(0.445) = 1.56$ higher if there was a defence alliance agreement between those countries, compared to countries with no defence alliance agreements.

The interpretation of the smooth effects in Fig.13 have an analogous approach as in Section 7.1.1 discussed above. Whereby these effects are small, and in the most cases constant over time.


Transitivity
Reverse Transitivity
Samesource
Samebuyer
In Degree Receiver
Out Degree Sender
Intrastate Conflict Sender
Intrastate Conflict Receiver
Log GDPm Sender
Log GDPm Receiver
Military Capability Sender
Military Capability Receiver
δ()
year
1950 1960 1970 1980 1990 2000 2010

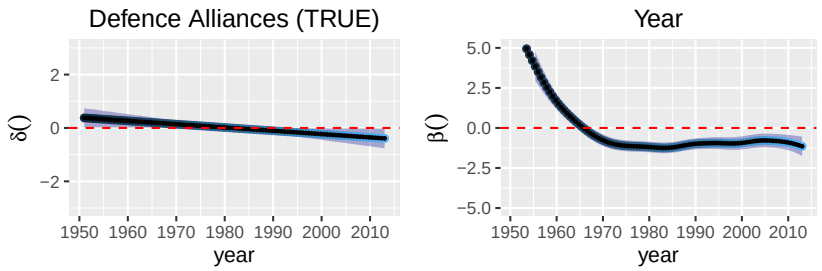


Figure 13: Varying-Coefficient Smooth Effects Dissolution Model

The QQ-plots in Fig.14 of the predicted random effects do not exhibit strong deviations from the assumption of gaussian distributions, so the assumption of gaussian random intercepts for sender and receiver countries seems to be quite reasonable..

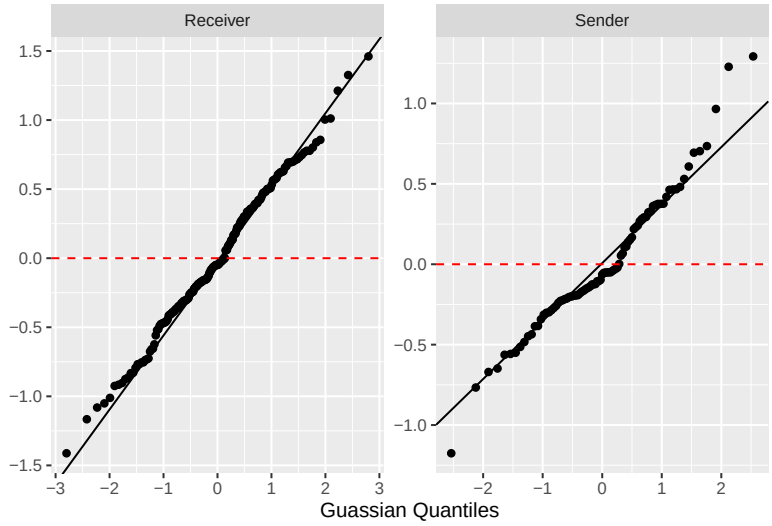


Figure 14: Guassian Quantiles for the Random Intercepts Dissolution Model

7.2 Evaluation

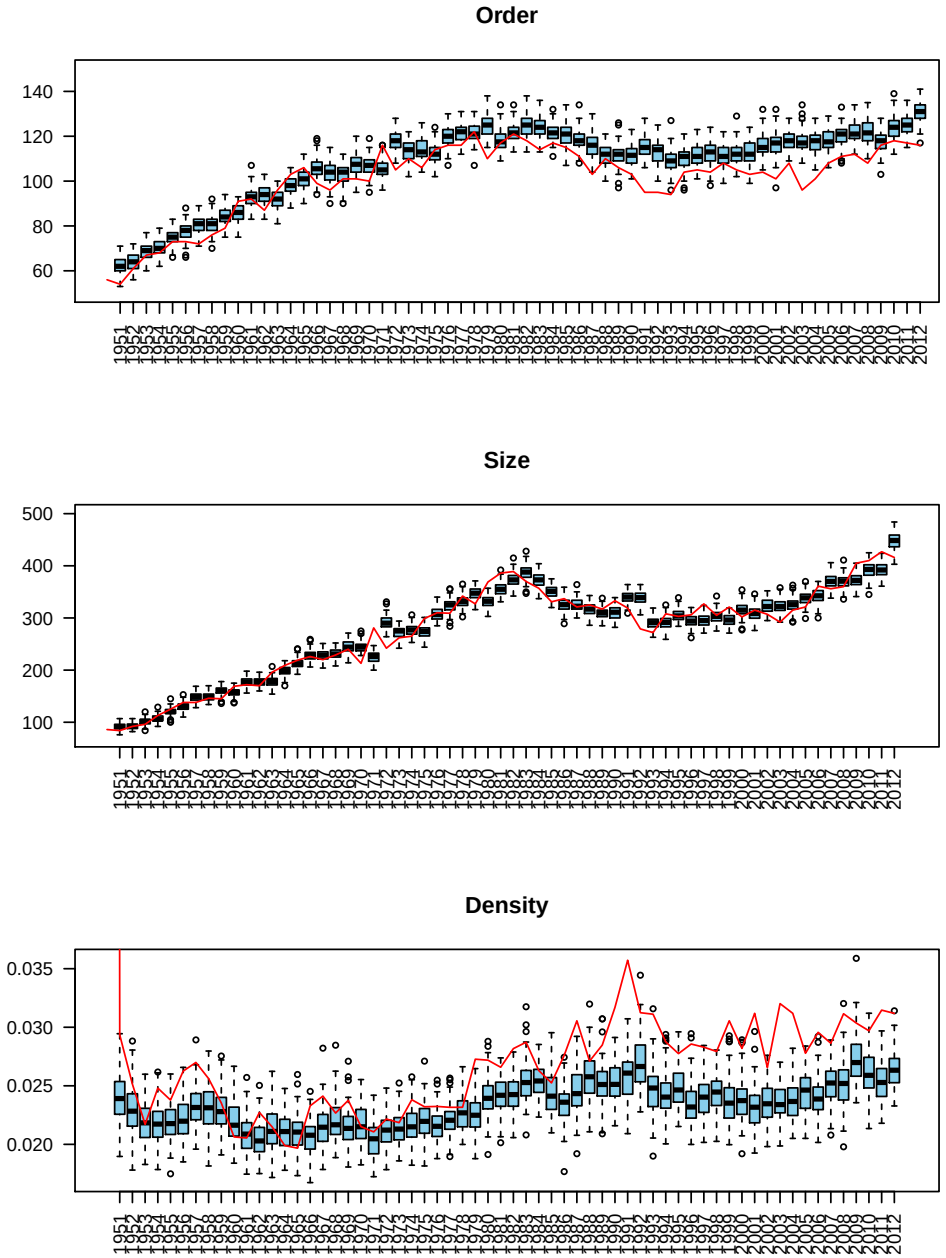
One of the major aims in modelling networks is to capture the specific network structure of the observed data within the specified and estimated model. A common way to evaluate how well this was achieved is to sample several networks based on the estimated model and compare those sampled networks to the original observed one based on their characteristics / network statistics. In our application case we are following Hanneke et al. [2010] and perform a *cross-validation style* evaluation: For each timepoint t we estimate our model based on all observed networks except the networks of timepoint t and $t+1$. Based on this model we sample C networks of international major conventional arms trade in year t and calculate various network statistics. We then compare the distribution of the network statistics of those sampled networks with the network statistics on the actual observed network for all timepoints t. The sampling of the c-th network for timepoint t based on a separable logistic network regression model, which can be achieved with the following pseudo code:

1. Sample Formation Network $\mathbf{y}^{+t,\,c}$:
   - If $y_{ij}^{t-1} == 1$ set $y_{ij}^{+t,\,c} = 1$
   - If $y_{ij}^{t-1} == 0$ sample $y_{ij}^{+t,\,c}$ from $B(1, \hat{\pi}_{ijt}^{+,\,-(t,t+1)})$
2. Sample Dissolution Network $\mathbf{y}^{-t,\,c}$:
   - If $y_{ij}^{t-1} == 0$ set $y_{ij}^{-t,\,c} = 0$
   - If $y_{ij}^{t-1} == 1$ sample $y_{ij}^{-t,\,c}$ from $B(1, \hat{\pi}_{ijt}^{-,\,-(t,t+1)})$
3. Evaluate $\mathbf{y}^{t,\,c} = \mathbf{y}^{+t,\,c} \setminus (\mathbf{y}^{t-1} \setminus \mathbf{y}^{-t,\,c})$,

with $\hat{\pi}_{ijt}^{+,\,-(t,t+1)}$ and $\hat{\pi}_{ijt}^{-,\,-(t,t+1)}$ being the predicted success probabilities for dyad $(i,\,j)$ in year t of the formation or dissolution model fitted on all data except the data of timepoint t and $t+1$.

Fig. 15 presents the results of the described evaluation approach based on $C = 1000$ network samples per year. For each year we computed the order, the size, the density statistic and the average in-degree of the vertices for each of the sampled and the observed networks. Additionally, we analyze the reciprocal structure in the networks by calculating the fraction of trading dyads with ties in both directions of all dyads with at least one tie, and the transitive structure of the networks by calculating the fraction of all triads with

three pairwise (undirected) trade relations of all triads with at least two pairwise trade relations. The yearly distribution of the statistics of the sampled networks is presented by boxplots, the time-series of network statistics on the observed trading networks is presented by a red line.



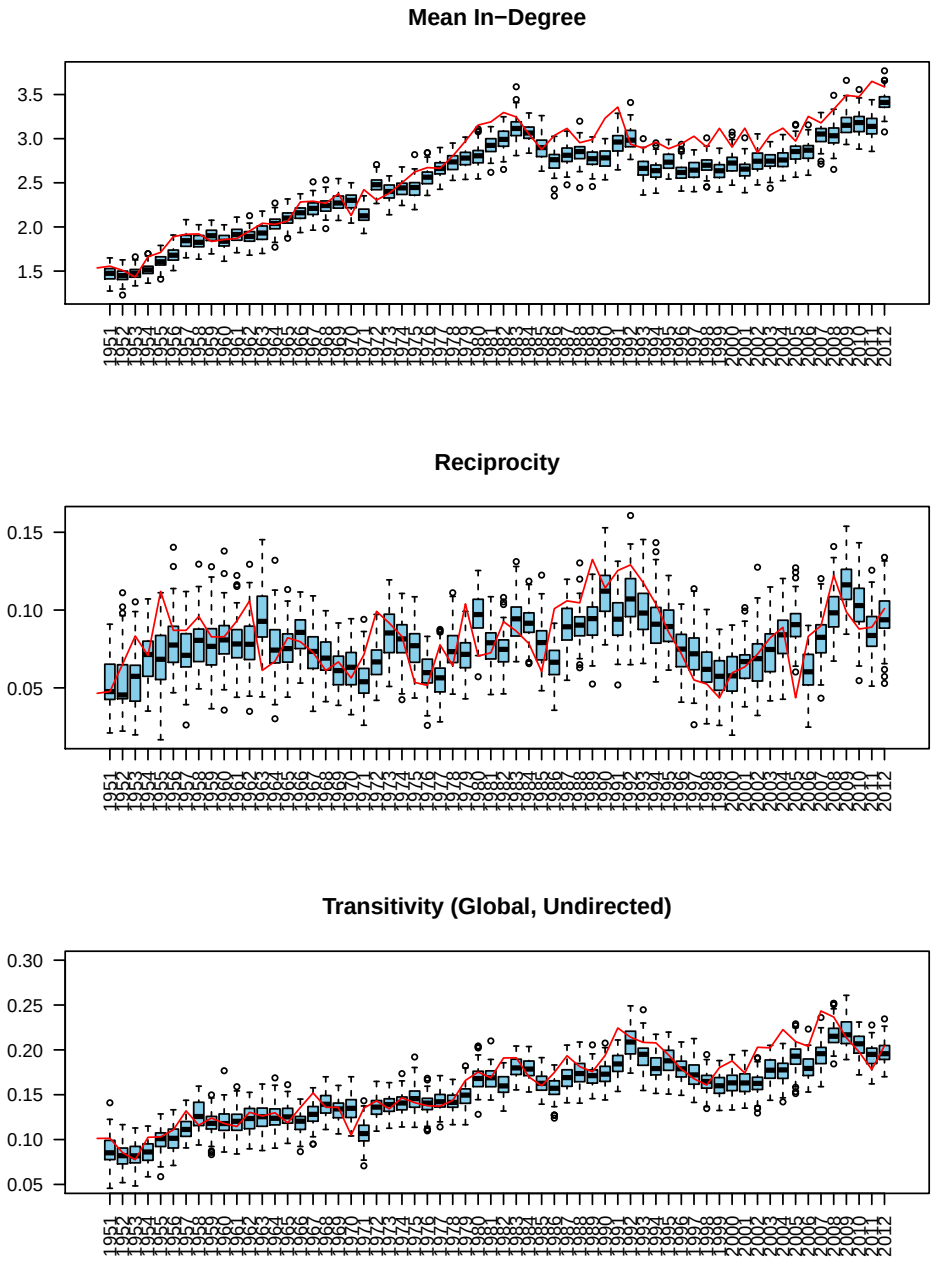


Figure 15: CV-Style Evaluation: Yearly distributed of network statistics for $C = 1000$ networks sampled from the estimated full model. Time-series of network statistics for observed networks (red)

In general it seems like sampling from the models produces trading networks that show a

quite similar structure to the true, observed ones. We seem to be able to capture the general time trends in order, size, in-degree, reciprocity and transitivity.

Looking at the order statistic we perform better in the earlier part of our observation period (till 1991), and then in the last part of our observation, it makes sense since the order of our networks grows continuously, we overestimate the actual number of involved actors in the early 90s, but we capture it at the end of our observation. For the size statistic, seems we are able to estimate the number of trading ties well. The other statistics seem also able to deliver good estimates, except the density, which is comprehensible thus density is related to the possible number of edges in the network, thats why it is obvious that we underestimate it especially after the 90s in our observation period.

Instead of looking at the network structure of sampled networks from our model, our second evaluation approach focuses on the two independent formation and dissolution parts of the model and evaluates their discriminatory power in their respective responses. For each year we estimate, again, the formation and dissolution model on all data except from year t and $t+1$ and calculate ROC-curves for the relevant responses (formation / dissolution adjacency matrix entries without / with lag-1 trade) of year t. ROC-curves are a well established method to evaluate binary classifiers, the idea is to plot the sensitivity (true positive rate) against $1 -$ specificity (false positive rate) for various classification thresholds q (assign class 1 if $\hat{\pi} > q$). The higher the sensitivity, while keeping the false positive rate low, the better the investigated classifier. The *area under the curve* (AUC) is then often used as a general performance measure of the classifier and can be interpreted as the probability that the model assigns a higher class-1 probability to an observation from class 1 than to an observation of class 0 for any pair of observations uniformly sampled from the two classes.

Fig. 16 presents the yearly ROC curves of the formation and dissolution models, we can see that the formation model seems to fit the data better than the dissolution model, however both models seem to be quite good in discriminating between the response groups with AUC scores of around 0.9 for the formation and around 0.7 for the dissolution models. All years' AUC values of the formation and dissolution model are presented in Table 9 in the Appendix.

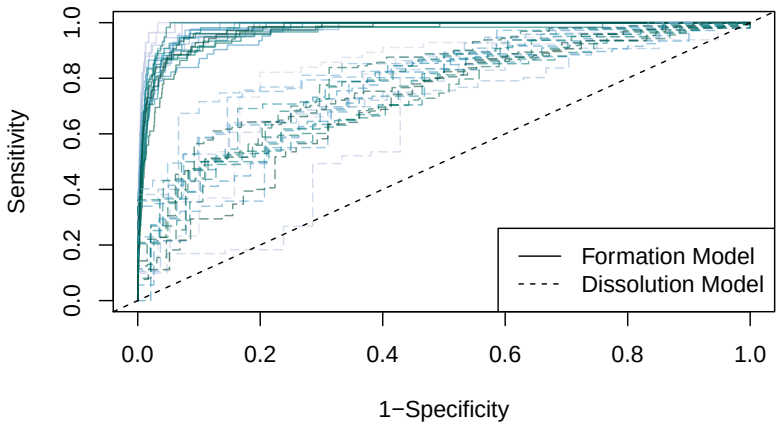


Figure 16: Yearly CV ROC-Curves: Formation and Dissolution Models

In our last evaluation step we now focus on extreme residuals of our estimated formation and dissolution models. This can either help us to detect "unusual observations" based on our models, or to find limitations of our model by investigating the observations with big residuals based on expert knowledge and thereby detecting factors that lead to those observations but are not considered within the models. For this purpose it is adequate to look at the *raw residuals* of the logistic regression models, i.e. the difference of the observed values (binary adjacency matrix entries $y_{ijt}^{+/-}$) and the predicted success probabilities $\pi_{ijt}^{+/-}$. Table 9 and Table 10 present the five biggest / smallest raw residuals of the formation model, they can be interpreted as being the most unexpectedly observed trade relations and, on the contrary, the most probable but not observed trade relations, based on the estimated formation model.

raw.resd	year	gdpm.s	gdpm.r	polity.sen.	polity.rec.	sender	receiver
0.99	1964	0	9142	0	2	Zimbabwe	Zambia
0.99	2010	0	0	0	0	Montenegro	Serbia
0.99	2010	0	34578	0	2	Serbia	Cambodia
0.99	2012	0	19375	0	5	Serbia	DR Congo
0.99	2004	17959	2158	7	-5	Georgia	Gambia

Table 9: Biggest residuals formation model: unexpectedly emerged trade relations

In Table 9 we see, for instance, that a trade relation between Zimbabwe and Zambia was observed in 1964, while we assigned, based on the formation model, a probability close to zero for it to arise. The smallest residual of the formation model, presented in Table 10, belongs to the trading dyad Germany and United States in the year 2005. We assign a probability of approximately 0.84 for observing a trade, however no trade between those actors was registered during that year. Those presented observations are, however, only the most extreme examples. To get the full picture it would be necessary to have a look at many more big, in absolute value, residuals. The formation model has e.g. 1807 observations with resulting raw residuals between 0.95 and 1, it would now be possible to analyze those trading dyads based on the involved actors, their observed network statistics, and arbitrary other factors.

raw.resd	year	gdpm.s	gdpm.r	polity.sen.	polity.rec.	sender	receiver
-0.84	2005	2614550	12564300	10	10	Germany	USA
-0.77	2007	1784665	13149344	10	10	Italy	USA
-0.70	1981	6027685	91430	10	-10	USA	Iraq
-0.70	1984	997087	73362	10	-9	UK	Iraq
-0.69	2004	2499782	12196382	10	10	Germany	USA

Table 10: Smallest residuals formation model: most probable, not emerged trade relations

Table 11 and 12 present the 5 biggest / smallest residuals of the dissolution models, they can be interpreted as the most unexpectedly persistent trade relations and the most unexpectedly dissolved trade relations.

raw.resd	year	gdpm.s	gdpm.r	polity.sen.	polity.rec.	sender	receiver
0.94	1979	46993	0	-2	0	Singapore	Brunei Darussalam
0.92	1977	498522	14770	-10	-6	Saudi Arabia	North Yemen
0.92	1982	21584	1919	-7	-7	Libya	Central African Rep.
0.91	1985	54048	24697	10	-6	New Zealand	North Yemen
0.91	1991	438040	0	10	0	Australia	Tonga

Table 11: Biggest residuals dissolution model: unexpectedly persistent trade relations

raw.resd	year	gdpm.s	gdpm.r	polity.sen.	polity.rec.	sender	receiver
-0.99	2002	160371	5110871	10	-7	Israel	China
-0.99	2003	11789128	2326411	10	9	USA	India
-0.98	2013	621234	4661976	10	9	Netherlands	India
-0.98	1997	9869378	3859120	10	-7	USA	China
-0.97	1984	6626666	73362	10	-9	USA	Iraq

Table 12: Smallest residuals dissolution model: unexpectedly dissolved trade relations

In Table 11 we observe, that a trade relation between Singapore and Brunei Darussalam in 1979 has been persisted, although based on the dissolution model we assigned a probability close to zero to persist. The smallest residual of the dissolution model, in Table 12, has the trading dyad between Israel and China in year 2002, where we assign a probability of approximately 0.99 for persisting a trade, however this trade was dissolved.

8 Summary and Outlook

The proposed *Separable Temporal Logistic Network Regression* approach seems to perform quite well on modelling the discrete time-series of major conventional weapons trade networks. Under the assumption of independent formation and dissolution processes and conditionally independent Bernoulli events for the entries of the two yearly adjacency matrices, it is possible to use the very flexible and established class of mixed additive logistic regression for modelling the binary trade networks. Our presented and evaluated model seems to fit the observed data well, networks sampled based on the model show similar characteristics than the originally observed ones. It exhibits some interesting relations between sender- and receiver-specific covariates, and the odds of e.g. forming new trade relations. Additionally, we are able to consider the network structure of previous years in our model, thereby we found i.a. a strong positive relation between reciprocal trade in the last year and the formation and persistence of trade relations.

However, due to the flexibility of the used additive logistic regression model and its *ready to use*-implementation in the R-package **mgcv**, it would be easily possible to extend and further adapt the presented model. This could e.g. be done by incorporating additional covariates. To us, the used assumptions in modelling seem to be quite reasonable and the proposed modelling strategy could in general be used for different application scenarios as well. The assumption of independent formation and dissolution processes becomes probably more and more reasonable, the denser the grid of (discrete) observed networks is. Nevertheless, a comparison of our modelling approach with other approaches for modelling time-evolving networks would be very interesting in itself and could, additionally, serve as a sensitivity analysis for our substantive results.

References

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

L. Fahrmeir, H. L. Kaufmann, and F. Ost. *Stochastische Prozesse*. Hanser München, 1981.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression - Models, Methods, Applications*. Springer, 2013.

S. Hanneke, W. Fu, and E. P. Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4, 2010.

T. Hastie and R. Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

P. Holtom, M. Bromley, and V. Simmel. *Measuring international arms transfers*. Stockholm International Peace Research Institute, 2012.

P. Hough, S. Malik, A. Moran, and B. Pilbeam. *International Security Studies: Theory and Practice*. Routledge, 2015.

L. M. Koehly and P. Pattison. Random graph models for social networks: Multiple relations or multiple raters. *Models and methods in social network analysis*, pages 162–191, 2005.

E. D. Kolaczyk and G. Csárdi. *Statistical Analysis of Network Data with R*. Springer, 2014.

P. N. Krivitsky and M. S. Handcock. A separable model for dynamic networks. *Journal of the Royal Statistical Society, Series B*, 76(1):29–46, 2014.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518, 1986.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(77), 2011.

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

SIPRI. Stockholm international peace research institute. https://www.sipri.org, 2017.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

S. Wood. *Generalized Additive Models: An Introduction in R*. Chapman and Hall/CRC, 2006.

S. Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, 2011.

List of Figures and Tables

List of Figures

List of Tables

9 Appendix

List of Countries / Actors

In the following table, all countries for which the *major conventional weapons*-data was gathered are listed. The entry in the 'Years' column indicates the period within the corresponding country is included into the network. A blank entry denotes that the corresponding country existed during the whole period ($1950 - 2013$).

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		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	since 1973	44	Cook Islands	since 1965
15	Bahrain	since 1971	45	Costa Rica	
16	Bangladesh	since 1971	46	Cote dIvoire	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	
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	Nation	Jahre	ID	Nation	Jahre
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-untilsau	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	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	Nation	Jahre	ID	Nation	Jahre
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	
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	
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		179	Sudan	since 1956
145	Panama		180	Suriname	since 1975
146	Papua New Guinea	since 1975	181	Swaziland	since 1968
147	Paraguay		182	Sweden	
148	Peru		183	Switzerland	
149	Philippines		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	Nation	Jahre	ID	Nation	Jahre
201	United Kingdom		210	Viet Nam, South	until 1976
202	United States		211	Western Sahara	since 1976
203	Uruguay		212	Yemen	
204	Uzbekistan	since 1991	213	Yemen, North	
205	Vanuatu	since 1980	214	Yemen, South	
206	Vatican (Holy See)		215	Yugoslavia, SFRo	until 1992
207	Venezuela		216	Zambia	since 1964
208	Viet Nam	since 1976	217	Zanzibar	since 1963
209	Viet Nam, North	until 1976	218	Zimbabwe	

List of excluded countries

In the following table, we decided to exlude some countries from our network data, thus these countries listed below do not have any trade informations in the whole period $(1950 - 2013)$

1	Abkhazia	11	Niue
2	Andorra	12	Saint Lucia
3	Antigua and Barbuda	13	San Marino
4	Cook Islands	14	Sao Tome and Principe
5	Darfur	15	Somaliland
6	Dominica	16	South Ossetia
7	Kosovo	17	Trans-Dniester
8	Liechtenstein	18	Vatican (Holy See)
9	Monaco	19	Viet Nam, North
10	Nauru	20	Zanzibar

Table of AUC values

	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
AUC Form	0.99	0.99	0.98	0.98	0.97	0.98	0.97	0.98	0.98	0.96
AUC Diss	0.82	0.75	0.66	0.76	0.73	0.85	0.83	0.75	0.77	0.74
	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
AUC Form	0.97	0.97	0.98	0.96	0.98	0.96	0.96	0.97	0.97	0.98
AUC Diss	0.75	0.79	0.75	0.73	0.79	0.73	0.79	0.77	0.70	0.76
	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
AUC Form	0.97	0.97	0.97	0.97	0.96	0.98	0.97	0.96	0.96	0.96
AUC Diss	0.74	0.81	0.77	0.77	0.75	0.74	0.74	0.74	0.74	0.76
	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
AUC Form	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.96	0.96	0.96
AUC Diss	0.76	0.74	0.77	0.76	0.78	0.74	0.79	0.75	0.78	0.80
	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
AUC Form	0.96	0.94	0.94	0.95	0.95	0.95	0.96	0.95	0.96	0.95
AUC Diss	0.77	0.70	0.80	0.78	0.76	0.77	0.72	0.76	0.77	0.76
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
AUC Form	0.95	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.95
AUC Diss	0.76	0.77	0.77	0.78	0.77	0.75	0.77	0.77	0.74	0.72
	2011	2012								
AUC Form	0.96	0.95								
AUC Diss	0.77	0.70								

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, July 20, 2017 .....

(Sevag Kevork)