Events in Social Networks

A Stochastic Actor-oriented Framework for Dynamic Event Processes in Social Networks

Christoph Stadtfeld

KIT Scientific Publishing

Christoph Stadtfeld

Events in Social Networks

A Stochastic Actor-oriented Framework
for Dynamic Event Processes in Social Networks

Events in Social Networks

A Stochastic Actor-oriented Framework
for Dynamic Event Processes in Social Networks

by
Christoph Stadtfeld

KIT Scientific Publishing

Dissertation, Karlsruher Institut für Technologie
Fakultät für Wirtschaftswissenschaften
Tag der mündlichen Prüfung: 30.11.2011
Referenten: Prof. Dr. Andreas Geyer-Schulz
Prof. Dr. Karl-Heinz Waldmann

Cover image: © Marish | Dreamstime.com
Cover layout: Lea Ellwardt
Chapter page image: cell-tower network in Boston visualized with visone

Impressum

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe
www.ksp.kit.edu

KIT – Universität des Landes Baden-Württemberg und nationales
Forschungszentrum in der Helmholtz-Gemeinschaft

KIT Scientific Publishing 2012
Print on Demand

ISBN 978-3-86644-792-9

Acknowledgements

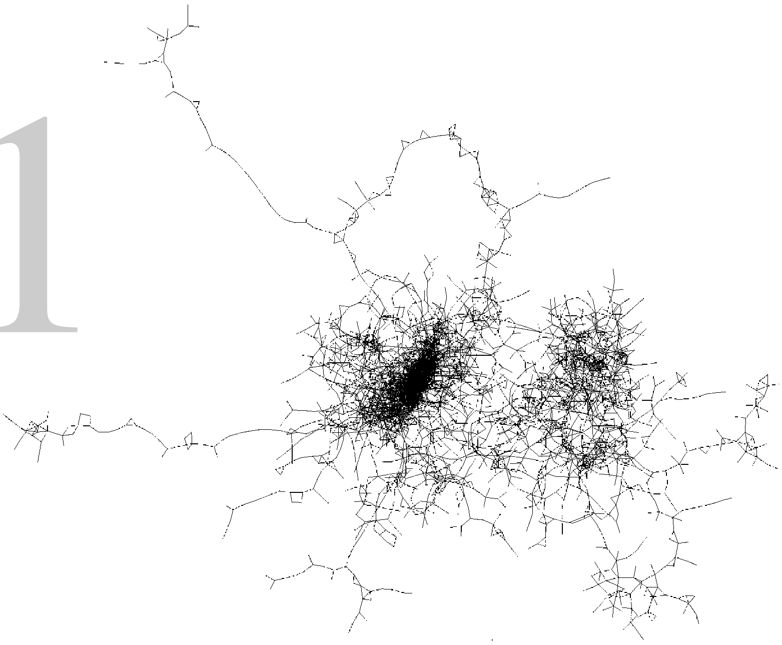
This dissertation is about dynamic processes in social networks. These processes are measured as sequences of events between people, such as phone calls and e-mails.

Writing this dissertation was similar to a dynamic, often non-deterministic process. It can partly be summarized by a sequence of events: conferences, papers, meetings, research visits, discussions and ideas. These events involved a number of inspiring people: my advisors, my collaborators and colleagues, the discussants of my work, the hosts of my research visits, my family and my friends. I thank you for you being such a great social network!

Christoph Stadtfeld
Karlsruhe, 30. November 2011

Contents

1 Introduction



1.1 Motivation

Interactions between people are ubiquitous. When people make phone calls, transfer money, connect on social network sites, or visit each other, these actions can be collected as dyadic, directed, relational *events*. Each of those events can be understood as driven by multiple individual decisions that at least partially involve rational considerations. As a whole, the many individually driven event decisions form social networks. In turn, these networks influence future event decisions. This book aims at developing models that allow to understand individual event decisions in the context of large social networks.

In recent years, there has been an increasing interest in the analysis of event data. A major reason is the growing availability of this type of data: With the emergence of computer-mediated communication, digital data storage, Web 2.0 technologies and social network sites, a variety of application scenarios and data sources with new and interesting research questions emerged. However, so far research in the field of social network analysis lacks the appropriate tools for the study of actor-driven events in social networks.

The research questions we tackle in this work are of methodical, computational, and substantive nature: First, we are interested in how event decisions can be modeled as a stochastic process. Second, a way to efficient estimation of the new class of models is proposed. Third, we apply the new event framework to two extensive case studies.

1. How can event streams be modeled as a stochastic process that takes different levels of decision making into account?

2. If such a model is defined, how can it be estimated with a reasonable computational effort?

3. Given an event model and its efficient implementation – can we fit models on real event streams, and thereby help to understand dynamic choice behavior in more detail?

First, our methodical contribution is to provide a new class of event models. Current methods lack tools for modeling large event data sets from an actor-oriented perspective. They are either designed for the analysis of dynamic network data measured at discrete points in time or they are not based on actor-oriented frameworks. Similar to the ideas of stochastic actor-oriented models (Snijders, 2005), we model events as a two-level decision process embedded in a Markov process, i.e. a continuous-time Markov chain. Next, we introduce extensions that allow modeling individual event choices more precisely by including additional decision levels. The core of the model is the choice of event receivers. For example, receivers may be communication partners in the case of communication events. The choice of receivers is a sub-model defined as a multinomial logit model with a variety of possible independent variables. It is possible to test, for instance, whether network structures or actor attributes influence the decisions to choose a certain receiver.

Second, the computational contribution is to present a robust estimation algorithm and to provide a new software tool. To date, there is no tool that is tailored to the specific demands of analyzing event stream data in large datasets. The proposed software tool applies a maximum likelihood estimation of independent network variables which is shown to be appropriate for the estimation of the model parameters. Through the use of preprocessing and heuristics, it is possible to estimate structural effects for long event data sets with a large number of actors involved.

Finally, the substantive contribution is related to two specific application scenarios. To our knowledge, these are two of the first empirical studies modeling communication event data. The successful implementation of model and software could give rise to empirical research from other disciplines that investigate decisions in large networks, such as sociology, biology, economics, and marketing. First, the communication choices of private messages in a question and answer community are analyzed. It is shown that actors both communicate in stable groups, and with others they are related to in a functional way. Over the communities's life time, the relevance of in-group communication does, however, increase. Second, by processing the phone call event stream of the MIT Reality Mining dataset, we show that structural effects in the communication network, the time spans since the last communication and spatial distance are all important predictors of the choice of communication partners.

In many scenarios it can be valuable to understand the dynamics of event streams from an actor point of view. In marketing research, for example, the flow of information among potential costumers could be analyzed to gain a deeper understanding of the dynamics of providing product recommendations. In marketing, it can also be valuable to understand the decision behavior of single actors, rather than that of the group, to identify individual roles. In organization research, it could be of interest to test whether formally defined network structures are reflected in informal communication behavior. For providers of social network sites it can be highly interesting to understand how individual decisions drive global dynamics, like community growth and content creation.

1.2 Structure of the Book

This book presents a new framework for the analysis of event stream data in social networks. It is structured into seven chapters.

Chapter 2 reviews relevant literature that is related to the proposed event framework. After a brief positioning of this work in the social networks literature, four relevant threads of literature are discussed. The development of *multinomial choice models* in econometrics is discussed first. The most important publication in this field stems from McFadden (1974). His article first introduced multinomial choice models, a discovery that was crowned with the Nobel Memorial Prize in Economic Sciences in 2001 for Daniel McFadden. The next three threads of literature regard structural network models: *Exponential random graph models*

(ERGMs) have been constantly refined since the 1970's. Based on the introduction of dependence assumptions by Besag (1974) and Frank and Strauss (1986), this class of models provides a tool for probability distributions over graphs that are specified with structural effects that are often assumed to be driven by actor decisions. A dynamic and actor-oriented advancement of ERGMs is the class of *stochastic actor-oriented models* (Snijders, 2001, 2005). Here, structural network decisions are embedded in a stochastic process for which different parameters can be estimated separately. The underlying data consists of panel data. Event-based models have been increasingly discussed in the last years. One important thread of work is the introduction of the *relational event framework* (Butts, 2008), which is an adaption of classical event history modeling.

Chapter 3 introduces the new, event-based framework[1]. To reduce the complexity of the explanations, a basic sub-model is presented. As a continuing example, a short phone-call event stream is modeled step by step. Event stream data is defined and rules are introduced that allow to translate event streams into graphs that represent social networks. Similar to stochastic actor-oriented models, the event framework is defined as a two-level decision process. General activity of actors and decisions regarding communication partners are estimated separately – both decisions are embedded in a Markov process. The second decision (on communication partners) is then discussed in more detail. It is defined as a multinomial choice model, which is exemplarily specified with two dyadic independent variables. The log-likelihood function, its concavity and a straightforward estimation algorithm are presented. The interpretation of estimates is discussed. The new event framework is compared to the structural network models introduced in chapter 2.

Chapter 4 extends the basic event framework and introduces a range of possible new specifications[2]. The need for new specifications can result from additional information in the data set that could be exploited, or from new research questions that could be answered with the model based on an advanced data collection. New information can be reflected in an extended state space of the stochastic process. Additional parameterizations, new decision levels, new process transition rates and individualized models can then be applied to the model. It follows a detailed discussion of extended specifications of the multinomial sub-model, which describes the choice of communication partners. To describe this choice, a wide range of independent variables can be defined: Endogenous structures in the event graph, structures in other graphs, actor attributes, weighted statistics and combinations of the previously mentioned structures are possible specifications. Some of the proposed extended specifications are applied in the two subsequent chapters 5 and 6.

Chapter 5 introduces an analysis of communication via private messages in a German-speaking question and answer web community[3]. We examined whether private commu-

[1] An earlier version of this chapter is published in Stadtfeld et al. (2010). New parts of this chapter have been written during a research visit at the University of Melbourne (funded by Karlsruhe House of Young Scientists, KHYS) in collaboration with Garry Robins and Philippa Pattison.

[2] Parts of this chapter are published in Stadtfeld (2010).

[3] This chapter is published in the journal *Social Networks* (Stadtfeld and Geyer-Schulz, 2011). An earlier

nication events depend on previous communication patterns and on the actors' affiliation with questions and answers in the community. The latter is understood as functional communication. As the analyzed event data was logged over many months, the same model is applied on different phases of the community development. Thereby, changes in communication patterns over time can be revealed. It seems that private communication within communities gets more relevant compared to purely functional communication over time.

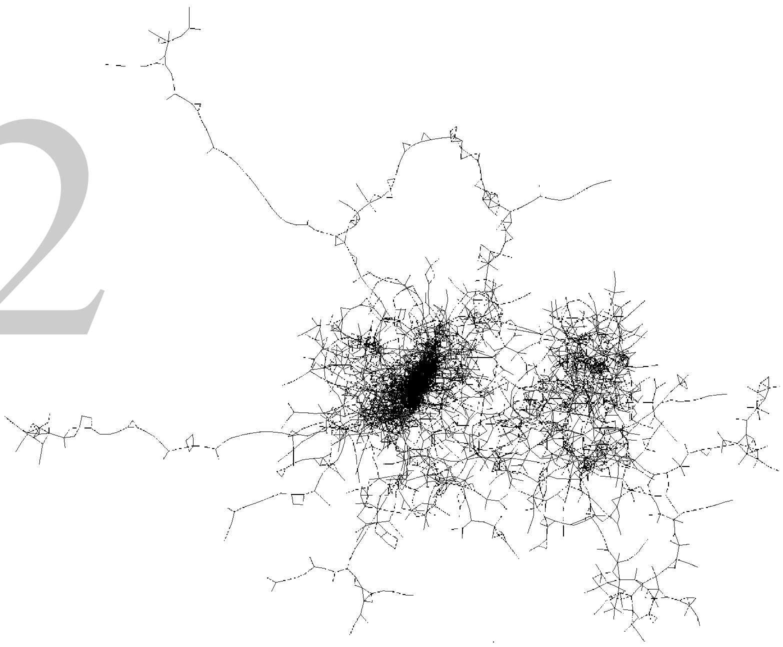
Chapter 6 introduces an analysis of the MIT Reality Mining dataset[4]. The dynamic communication patterns of phone calls among a group of students are analyzed. It is shown that the choice of phone call receivers is significantly influenced by spatial distance, time effects and structures in a communication graph that represents previous communication.

Chapter 7 summarizes the findings of this book and gives an outlook on future research questions.

version is published online (Stadtfeld and Geyer-Schulz, 2010).

[4] This chapter is published in the proceedings of the *Third IEEE Conference on Social Computing 2011* (Stadtfeld et al., 2011).

2 Related Work



The statistical assessment of social network data was discussed first in the 1930's, see Moreno and Jennings (1938). Networks or their representation as graphs can be understood to be realizations of stochastic processes that are driven by individual sub-processes. Erdős and Rényi (1959, 1960) introduced the concept of random graphs in which edges are assumed to be created by stochastic processes. Since their fundamental publications, research into statistical properties of (social) networks has been performed in many different scientific disciplines. Regarding social networks research, Robins (2009, section II) states "that history is rather sparsely scattered across various literatures and different eras and, as a result, even today we see that older techniques and approaches are reinvented and repackaged as new". A short introduction to the history of social network analysis is provided by Freeman (2004). An excellent overview of basic concepts and the different fields of research within social network analysis until the mid 1990's is given by Wasserman and Faust (1994). Brandes and Erlebach (2005); Newman (2010) also give good overviews, including some newer developments.

The research presented in this book connects to, builds on and details on techniques of social network analysis. More specifically, we are interested in the statistical evaluation of social network structures in dynamic environments. The data type generated in dynamic environments is often described as event stream data. Events are observations of individual behavior in social networks. The behavioral patterns observed are understood to be driven by rational choices (see Coleman (1990)). An important (econometric) framework for multinomial choice processes based on rational choice theory was developed by Daniel McFadden (McFadden, 1974). In section 2.1, the origins and extensions of his Nobel prize-winning choice framework are explored. It is the core of the stochastic actor-oriented framework for event data introduced in this book.

In the three following sections it is focused on structural network models for the analysis of cross-sectional data, panel data, and event stream data.

First, section 2.2 describes the class of *exponential random graph models* (ERGMs). The development of this class of models was mainly driven by theories and research problems of social sciences (although some of the basic ideas were developed by biologists). ERGMs are a key tool for understanding structural effects in social networks, which were collected at a single point in time (i.e., cross-sectional data). Although the non-dynamic nature of the data does not allow to study actual dynamic effects, ERGMs are usually used to explain outcomes of unobserved, dynamic decision processes. However, the limitations of cross-sectional data sets make such an interpretation difficult. Before moving to dynamic models, the basic idea of ERGMs, its dependence assumptions, different estimation procedures and new specifications will be introduced.

Second, the class of *stochastic actor-oriented models* (SAOMs) reviewed in section 2.3 is closely related to ERGMs. SAOMs may be understood to be a dynamic, actor-oriented extension of the class of ERGMs. SAOMs were developed to study structural effects in social network panel data. Similarly to ERGMs, SAOMs allow to evaluate the influence of endogenous (i.e., preferences for structural configurations in the social network) and exogenous effects that take into account actor attributes, for example. However, these effects are modeled over time. SAOMs can be extended to model the co-evolution of social selection and influence processes.

Third, the increasing availability of event stream data gave rise to new statistical models. These models are mainly based on traditional event history modeling – a technique commonly used in social sciences to evaluate state changes of individuals in a population over time. The *relational event framework* is an adaptation of this idea to relational event data. This type of data consists of dyadic, often directed social interactions between actors. The observations are at least ordered and often even timestamped. Unlike SAOMs, the relational event framework is not explicitly defined as an actor-oriented model. The basic ideas of this approach are explained in section 2.4 together with possible extensions.

2.1 Multinomial Choice Models

Multinomial choice models are applied in economics and related disciplines to describe discrete, individual choices. Given a set of alternatives, a probability is assigned to each possible choice. This probability depends on attributes describing the concrete choice and the actor. As in linear regression models, these attributes are understood to be independent variables that determine the outcome of a dependent variable. In multinomial logit models, however, the dependent variable is not measured on a continuous scale, but as the probability of a choice of an item in a discrete set of alternatives. A linear model might describe the *money of an actor spent for cars per year* predicted by the actor's income. A multinomial model, on the other hand, might describe the choice of a *specific car model* predicted by the income (actor variable) and the color (choice variable). In this book, we are interested in explaining *social choices*. As long as these choices can be understood as optimizations of individual utility functions, the same classes of multinomial choice models as in economics can be applied.

In the following, the ideas of multinomial logit models as introduced by McFadden (1974) shall be outlined. There are two ways to approach this class of models, which slightly differ in their parameterization but can be combined (Cramer, 2003, p.128f). First, this class of models can be understood as a generalization of the binomial logit model (general logit model). Second, it can be constructed from rational choice theory and understood to be a random utility maximization (conditional logit model). The two approaches will be discussed in this order. In later chapters, both models will be referred to as multinomial logit models, but an underlying actor-oriented utility maximization will be assumed.

The simplest multinomial logit model is a binary model (a logistic regression model). In this class of models, the dependent variable has only two different states. For example, it may be modeled whether an actor is likely to buy a car or to accept a friendship request on a social networks platform (or not). The probability in case of the friendship request might depend on the similarity in gender of the two actors involved. In case of the car choice, the actor's income or socio-economic status may be independent variables. In the following, we refer to the car example. It is also used for explaining logit models by Cramer (2003).

Let Y_i be a random variable that describes the outcome of the i-th observation in the sample. It can be interpreted as the outcome of a stochastic decision process. It is either 1 (if actor i buys a car) or 0; $Y_i \in \{1,0\}$. The probability of buying a car depends on the independent variable x_i that represents the income of actor i. The univariate probabilities are given by the logistic function:

$$P(Y_i = 1|x_i) = \frac{\exp(\beta x_i)}{1+\exp(\beta x_i)} \tag{2.1}$$

$$P(Y_i = 0|x_i) = \frac{1}{1+\exp(\beta x_i)} \tag{2.2}$$

Parameter β is unknown and estimated based on the observations in the sample $Y_1, Y_2, \ldots,$ for example, by applying a likelihood maximization. If $\hat{\beta}$ is the maximum likelihood estimator and positive, then it may be inferred (on a certain level of significance) that actors with a high income are more likely to buy a car. If it is negative, the opposite effect is true. If it is zero, we can infer that income has no effect on the decisions to buy a car. In this case, both probabilities are equal: $P(Y_i = 1|x_i) = P(Y_i = 0|x_i) = \frac{1}{2}$. A detailed introduction to binary logit models can be found in Hosmer and Lemeshow (2000, pp.1–46).

The logistic regression model can be understood to be a choice between two alternatives (either a car is bought, or not). The standard multinomial logit model is a generalization of this binary model. An important assumption is the *independence of irrelevant alternatives* (IIA). It states that the ratio of two discrete choice probabilities is independent of the existence of further choice alternatives. This assumption was introduced by Luce (1959) as the *choice axiom*. Luce formulated it as the irrelevance of additional choices on choice probabilities that are based on *constant* utility functions (without an error term). The generalized IIA feature assumes choice probabilities based on random utility functions (Cramer, 2003, p.129).

If $P_S(a)$ denotes the positive probability of choosing alternative a out of a set of alternatives $S = \{a, b, c, \ldots\}$, then the choice axiom states (see Luce (1959, p.9)) that

$$\frac{P_S(a)}{P_S(b)} = \frac{P_{\{a,b\}}(a)}{P_{\{a,b\}}(b)} \tag{2.3}$$

although generally $P_S(a) < P_{\{a,b\}}(a)$. From this axiom, Luce derived that a general form

$$P_S(a) = \frac{v(a)}{\sum_{b \in S} v(b)} \tag{2.4}$$

must exist with v being a positive, real-valued function (Luce, 1959, p.23).

Let us consider the car choice example again. After deciding to buy a car at all, an actor chooses one out of three car types. We closely follow the introduction in Hosmer and Lemeshow (2000, p.261f). The random variable Y_i now has three possible realizations: $Y_i \in \{0,1,2\}$. The probability of a specific realization depends on a *vector x*, including p covariates (like income) and one intercept (as in linear regression models) with $|x| = p+1$. Following the choice axiom, the ratio of two probabilities can be denoted by the ratio of two binomial logit models:

$$\begin{aligned} v_1(x) :=& \frac{P(Y_i = 1|x)}{P(Y_i = 0|x)} \\ =& \exp(\beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p) \\ =& \exp(\beta_1^T x) \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} v_2(x) :=& \frac{P(Y_i = 2|x)}{P(Y_i = 0|x)} \\ =& \exp(\beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p) \\ =& \exp(\beta_2^T x). \end{aligned} \tag{2.6}$$

Any outcome $Y_i = 0$ can be defined as having the null parameter vector β_0. From $\sum_{j=1}^{3} P(Y_i = j) = 1$, it follows that

$$P(Y_i = 0|x) = \frac{1}{1 + v_1(x) + v_2(x)} \tag{2.7}$$

$$P(Y_i = 1|x) = \frac{v_1(x)}{1 + v_1(x) + v_2(x)} \tag{2.8}$$

$$P(Y_i = 2|x) = \frac{v_2(x)}{1 + v_1(x) + v_2(x)}. \tag{2.9}$$

With $v_0(x) = 1$, it follows that

$$P(Y = j|x) = \frac{\exp(\beta_j^T x)}{\sum_{k=1}^{3} \beta_k^T x} \tag{2.10}$$

which is the basic form of a standard multinomial logit model. It can be seen that each choice has the same vector of covariates x – this is different from the conditional logit models that will be explained below. Having the car choice example in mind, such a model would explain *which* factors drive a certain categorization. For example, the influence of income, gender or education on the choice of cars could be tested. The standard multinomial model does, however, not aim at explaining *why* specific choices are made, which would be an actor-oriented explanation.

The class of conditional logit models aims at explaining why certain choices are made by evaluating the effect of choice covariates on the probability distributions. This class of models was introduced by McFadden (1974) based on previous research into random utility maximization (RUM) models. RUM models were originally developed in psychology as extensions of stimuli-response models and introduced to economics by Marschak (1960). A good overview of the origins can be found in McFadden (2001). In contrast to the standard multinomial logit model, these models are based on individual utility functions. The utility functions are assumed to be stochastic (this is different to *constant utility models* studied by Luce (1959)). Choices are assumed to be rational. The stochastic nature is assumed due to missing information about the actor who makes the choice observed, or about the alternatives the actor has (Manski, 1977, p.229).

McFadden (1974, p.108) defines the individual utility of a choice j as

$$U_j = V(x_j) + \varepsilon(x_j) \tag{2.11}$$

where x_j is a vector of attributes of j. The alternative j is in a set of possible choices S. McFadden additionally introduces actor covariates that will not be considered here. Function V is non-stochastic, the mentioned stochastic uncertainty due to missing information is represented by function ε. The probability of choosing i over j (with attribute vectors x_i,x_j) equals $P(U_i > U_j)$ and depends on both the deterministic and the stochastic part of the utility function 2.11 (see Cramer (2003, p.131)). With certain assumptions regarding the distribution of the ε functions (independently and identically distributed with a type I extreme value distribution of standard form), the standard form of conditional logit models can be inferred. The transformation is similar to the standard logit case above. Based on two axioms – the IIA and a positive probability condition for all choices – McFadden (1974, p.109–110) derives the probability of a choice j from set S with the vector of independent variables x_j as:

$$P(Y = j|S,\beta) = \frac{\exp(\beta^T x_j)}{\sum_{k\in S}\exp(\beta^T x_k)} \tag{2.12}$$

This probability is similar to the standard form in equation 2.10. However, it differs regarding the parameterization. Now, the covariates describe the actors *and* the choices instead of the actors alone. The homogeneous vector β evaluates these independent variables. In McFadden's model, actor *and* choice covariates are included. Therefore, the conditional logit model can help explain *why* certain choices are made. When analyzing a car choice, we may learn, for example, that expensive, red cars are rather bought by rich actors. The standard multinomial logit only explains *which* choice is made, depending on the actor attributes. It may explain that a *specific car model* is rather bought, if the choosing actor is rich.

In many cases, it makes sense to combine the standard multinomial logit and the conditional logit models. In McFadden (1974), also actor-related covariates were included. The general term referring to such a combination is *multinomial logit models* and will be used below. Note, that the IIA assumption of the multinomial logit model may be unrealistic, so that other models should be applied, e.g. nested logit models (see McFadden (1981)) or multinomial probit models (see Cramer (2003, p.128)).

2.2 Exponential Random Graph Models

There are many key figures that can be calculated to describe social networks. Networks (or rather their representation in the form of graphs) may be clustered, average distances between actors or degree distribution can be calculated. Networks can be decomposed and simplified or plotted using a variety of algorithms. All these techniques help to understand the characteristics of an observed social network. Good introductions to the mentioned techniques can be found in Wasserman and Faust (1994); Brandes and Erlebach (2005); Newman (2010).

The class of exponential random graph models (ERGMs, sometimes named p* models) goes beyond a purely technical description of network features. ERGMs aim at providing a family of probability distributions for social networks that are represented as graphs. ERGMs evaluate the structures within an observed graph against the expected structures of a random graph. If the network is big enough, then the available structural information is sufficient to statistically infer the importance of certain structures in the network formation process. In the following sections, the general idea of ERGMs will be presented with some basic specifications. Then, the important assumption of dependence structures will be discussed, followed by a short literature review on estimation procedures for ERGMs. This complex problem is linked to the problem of model degeneracy. It is tackled by extended model specifications that are introduced at the end of this section.

2.2.1 General Idea

Two well-known introductions to the class of ERGMs and its origins are Wasserman and Robins (2005) and Robins et al. (2007a). ERGMs provide probability distributions for graphs that are specified with counts of graph sub-structures. These variables are measured on a global level. However, they are assumed to arise in small local environments. Therefore, the interpretation of ERGMs is often related to actor decisions, although the model is not explicitly defined in this way. ERGMs are applied to static graphs that are measured at one point in time. These graphs represent an unknown history of previous actor choices about the creation of links. Hence, the data are cross-sectional.

Graphs consist of vertices and edges. In the context of social networks, the vertices represent social actors and the edges some kind of relation (e.g. friendship, collaboration, trust, advice, ...) between actors. Edges can be directed or undirected, weighted or unweighted (i.e., binary). In this overview we use the example of binary, directed graphs.

A graph X is a simple representation of a social network and defined as a tuple

$$X = (N, E) \tag{2.13}$$

where N is the set of vertices (nodes) and E is the set of edges (links). Such a graph can be described easily by an adjacency matrix which hereinafter is also denoted by X. X can be understood to be a random variable. A concrete graph x is a realization of X. In the following, we assume that all possible realizations x have a fixed set of nodes $N =$

$\{1,2,\ldots,n\}$ of size n that are connected by binary, directed edges:

$$x = (x_{ij}) = \begin{cases} 1, & \text{if there is a relation from actor } i \text{ to } j \\ 0, & \text{else} \end{cases} \tag{2.14}$$

with i and j in $\{1,\ldots,n\}$. If we talk about the graph x below, it may be considered to represent friendship nominations in a group of actors surveyed. If edges represent such a substantive relation, they may be referred to as ties (Wasserman and Faust, 1994, p.18). ERGMs express the probability of the occurrence of a specific graph given the counts of small structures in the graph. These small structures are assumed to be generated by individual choice processes. For example, a friendship nomination graph with many reciprocated and transitive tie structures is more likely than a graph in which friendship is hardly ever reciprocated and actors with common friends tend not to nominate each other as friends.

The probability of a specific graph is given by

$$P(X = x) = \frac{1}{\kappa} \exp\left(\sum_{k=1}^{P} \beta_k s_k(x) \right) \tag{2.15}$$

where s_k is one of the P sufficient statistics of the observed graph. Each statistic s_k is weighted by a real number β_k. The normalizing constant κ ensures that equation 2.15 is a proper probability distribution. As this term is a constant, the probability distribution is member of the exponential family (see Young and Smith (2005, p.81ff)). The constant is defined as an evaluation of all possible realizations x of the random variable X:

$$\kappa = \sum_{x \in X} \exp\left(\sum_{k=1}^{P} \beta_k s_k(x) \right). \tag{2.16}$$

The core of a concrete ERGM is its specification by a vector of sufficient statistics $s_k(x)$. These statistics are measured globally, but often understood to be driven by individual choices. For example, density can be calculated as the number of ties in the graph relative to all possible ties between the n nodes of the graph. Density – as a global measure – usually is low in social networks. This can be explained best by the individual actor decisions not to link with too many others (the local view). Concrete hypotheses might consider the "costs" of adding ties (Snijders et al., 2010b, p.10) or the cognitive limitations of individuals (Dunbar, 1992).

The specification of a model with concrete statistics s_k is driven by research questions ("Is there a tendency for reciprocity in friendship networks?"). Given a specified model, the next step is to estimate parameters that explain the observed graph as precisely as possible. The estimation process will be discussed later.

The density or outdegree statistic is in fact the most important ERGM specification. It is defined as

$$s_1(x) = \sum_{i,j} x_{ij} \tag{2.17}$$

which is a count of the ties in the graph over all actor tuples i,j in the set of nodes N.

The degree of reciprocity is defined as the number of dyads (Wasserman and Faust, 1994, p.18) with mutual ties. It can be measured as

$$s_2(x) = \sum_{i<j} x_{ij}x_{ji}. \tag{2.18}$$

The sum is defined over all nodes i,j in N. To prevent duplicate counts, it holds that $i < j$.

Another important statistic in ERGM is the count of transitive triads, which is defined as follows:

$$s_3(x) = \sum_{i,j,k} x_{ij}x_{jk}x_{ik}. \tag{2.19}$$

This statistic counts structures with three nodes i, j, and k in N. Other examples with three nodes are counts of two-out stars, two-in stars, mixed stars (paths of length 2), or three-circles. An overview of basic, directed structures with up to three nodes is given in Robins et al. (2007a, p.183).

A graph having a probability distribution with an empty vector of statistics is called a Bernoulli graph. The probability of each tie (and, therefore, of each graph) is equal and does not depend on other ties in the graph.

More complex structures that exceed three nodes can be part of the model specification (Snijders et al., 2006). Moreover, it is possible to take attribute values into account (Anderson et al., 1999). For example, it might be of interest whether there is a general tendency to homophily regarding a certain attribute (e.g., friendship depending on gender similarity).

The complexity of the included statistics is influenced by the local dependence assumption. Usually, the potential dependence between ties of a graph is restricted to small local environments. This assumption and its origins shall now be reviewed in more detail.

2.2.2 Dependence Structures

Assume that the graph at hand represents friendship nominations in a group of students. $X_{ij} = 1$, if actor i nominates j as a friend. Intuitively, actor i's nomination of j as a friend cannot be assumed to be independent of the realization of X_{ji}: Mutual nominations will empirically be observed more often than expected by random choice. The dependence structures are probably even more complex: If actors i and j have nominated many common actors as friends, the probability of friendship between i and j usually increases.

The problems of statistical distributions for observed graphs are the theoretically very complex dependence structures between the random variables describing the realization of one tie. Given a random graph X, the random variable X_{ij} describes the outcome of the directed tie from node i to j. In a binary network, the random variable may have two realizations: It is 1, if a tie from i to j exists. Otherwise, it is 0. Generally, it hold that the probability of each tie is theoretically conditional on the existence of all other ties:

$$P(X_{ij} = 1 | X_{12}, X_{13}, \ldots, X_{21}, X_{23}, \ldots X_{n(n-1)}). \tag{2.20}$$

The conditional dependence on other ties is restricted to local dependence structures, such as mutuality and triangles of friendship. Then, it is possible to infer the closed form of equation 2.15 as a description of the joint probability distributions of the conditional tie random variables in a graph.

In the early 1980's, research aimed at defining joint graph probability distributions with very restrictive dependency assumptions. Holland and Leinhardt (1981) introduced probability distributions for directed graphs that were (like ERGM) members of the exponential family. However, the model assumed dyadic independence. There is a dependence between ties within dyads, but not between dyads. Referring to the example above, the model may express mutuality of friendship nominations, but not triangular structures of friendship nominations. This new class of models was named p_1 models with p* (ERGM) being its later generalization. Wasserman (1980) modeled the change of a stochastic process with a Markov assumption over time. This work also assumed dyad independence.

The breakthrough towards a joint graph probability distribution with arbitrary dependence assumptions came with the application of the (until then unpublished) Hammersley-Clifford theorem to spatial systems by Besag (1974). It is proven that the most general form of a probability structure can be described as a generalized linear form that takes only cliques in an underlying and known dependence structures into account (Besag, 1974, pp.197–198). In Besag's article, these dependence structures are exemplarily related to variables of spatial sites $X_1, X_2, \ldots$ in an agricultural environment where an "infection" between two sites is only possible if the sites are neighbored. Cliques of dependence structures in this context are all single sites and all pairs of neighbored sites. The most general form that Besag derives is similar to equation 2.15 above.

The dependence structures were generalized later to so called Markov graphs by Frank and Strauss (1986).

These considerations of dependence structures were first applied to social graphs by Frank and Strauss (1986). They defined Markov graphs as graphs with a specific dependence structure: Two ties are assumed to be independent if and only if they do not share a common node. X_{ij} and X_{kl} are conditionally independent if i, j, k, l denote four different nodes in the graph. In later research this dependence structure was extended (see Snijders et al. (2006)).

In Frank (1991) and Wasserman and Pattison (1996) the basic model of Frank and Strauss (1986) was elaborated further. Anderson et al. (1999) were the first to summarize these findings and give an overview of the ERGM framework (they called it p*). They extended the framework to multiple networks, valued graphs, and statistics incorporating attributes. Further improved model specifications are discussed in section 2.2.4.

2.2.3 Parameter Estimation

A central issue regarding ERGMs is efficient and robust parameter estimation. ERGMs are usually estimated for a single cross-sectional data snapshot. The data include information about the relations of n actors at a certain point in time. The probability of an actually observed graph (see equation 2.15) with a vector of estimated parameters β can be understood as the likelihood. An estimation procedure may aim at calculating or approximating the maximum likelihood estimates. In most cases, however, a direct calculation is not possible, as the constant κ in equation 2.16 consists of a huge number of summands, namely, $2^{n(n-1)}$ in case of a directed binary graph.

Besag (1975) proposed to use pseudo likelihood estimates for the basic spatial models. Frank and Strauss (1986) commented on the estimation difficulties and discussed a simulation-based estimation method with strong constraints. Strauss and Ikeda (1990) extended Besag's ideas and provided a possibility to estimate ERGM parameters by applying a pseudo likelihood estimation. The pseudo likelihood approach was further elaborated by Frank (1991); Wasserman and Pattison (1996).

Geyer and Thompson (1992) proposed Monte Carlo Markov chains (MCMC) for the estimation of the maximum likelihood of dependent data. This idea was applied to the estimation of ERGM parameters by Corander et al. (1998). Snijders (2002) proposed an improved variant of an MCMC algorithm. Unlike pseudo likelihood methods which are related to likelihood maximization, the MCMC algorithms use the method of moments. In general, the method of moments processes less information than a maximum likelihood approach. Snijders (2002) found that simple simulation algorithms often suffer from convergence problems. For poorly specified models, it turned out that often bi-modal or multi-modal distributions were simulated. So, a simulation of concrete models would, for example, often return an empty network or a complete network. In general, many model simulations create degenerate graphs. The space of "good" models often turned out to be very small and volatile. Snijders (2002) also provides detailed analyses of the sometimes poor accuracy of pseudo likelihood estimates.

2.2.4 Extended Specifications

The basic specifications of ERGMs as introduced in Frank and Strauss (1986), extended in Anderson et al. (1999) and summarized in Robins et al. (2007a) come with one important problem. When estimated, most possible parameter combinations often return nearly degenerate distributions. This problem was discussed in the context of parameter estimation. A stepwise simulation of graph changes with such a set of parameters (as implemented in the MCMC algorithms) would then lead to graphs with unnatural structures, such as almost empty or almost complete graphs. Snijders (2002) explained the estimation challenges that are related to this problem.

Therefore, ongoing research aims at defining new specifications for ERGMs that are less prone to the problem of degeneracy. An important milestone was the specifications paper by Snijders et al. (2006). It proposes geometrically weighted degree statistics instead of linear count statistics (p.111ff). Each additional tie in a structure is then weighted less. Alternating triangles as a representation of transitive effects are introduced (p.115ff). In the standard models, transitivity is measured as shown in equation 2.19. The problem can be understood intuitively: Whenever denser regions with many nodes exist in the graph and there is a positive transitivity effect, then the inclusion of an additional tie will lead to an extremely high likelihood improvement. The (likely) inclusion makes the region even denser, such that unnaturally dense regions are observed easily in simulated graphs. A new structure is proposed that measures the existence of k triangular structures with one common tie. The Markov dependence graph of Frank and Strauss (1986) has to be extended for this purpose. A similar statistic is discussed for two-paths structures (alternating independent two-path, p.123fff). With these new specifications, the global graph is less sensitive to changes in structural parameter estimates. New specifications for ERGMs of directed graphs are discussed in Robins et al. (2009).

2.3 Stochastic Actor-oriented Models

Stochastic actor-oriented models (SAOMs) are used to model change in social networks over time. Networks are represented by binary graphs that are observed at several discrete points in time. Between two subsequent observations, many changes in the graphs are possible. Good introductions to this class of models are provided by Snijders (2005) and Snijders et al. (2010b). Each change in the graphs is understood to be driven by individual decision processes. Two decisions are embedded in a Markov process: First, there is a general activity of actors that determines the points in time at which actors change their configuration of outgoing ties. Second, an actor-oriented multinomial logit model determines the concrete outgoing tie that should be changed. Non-existing ties can be included, existing ties can be removed by the actor. The ideas are based on earlier literature in which longitudinal network models were described as Markov processes (see Holland and Leinhardt (1977a,b)). Coleman (1964, p.132ff) already proposed to model social (non-network) processes as Markov processes even if the data at hand are available at discrete points in time only. The introduction of SAOMs was partly motivated by the successful development of ERGMs for cross-sectional data and by the wide availability of panel data sets. Dynamic network models have been of increasing interest. Especially, as improvements in the estimation procedures allow to estimate complex models.

The underlying data of SAOMs are panel data. This type of data is usually collected by asking a fixed set of actors (e.g., pupils of a school class) about the relationship to all other actors in the same panel (e.g., about their friendship relations to each other). In contrast to cross-sectional data sets, actors are asked these questions repeatedly, for example, every month over a one-year period. The available data are discrete, while the underlying process that changes this network is assumed to be continuous.

A first approach to stochastic actor-oriented modeling of social networks panel data was introduced by Snijders (1996) and elaborated further in Snijders (2001, 2005); Snijders et al. (2010b). In van de Bunt et al. (1999), one of the first SAOM applications was presented, which illustrates the idea of SAOMs quite well: Friendship relations among a group of university freshmen (that did not know each other before) were observed at seven discrete time points within their first academic year. For each two sub-sequent time points, a model was defined to explain the observed relationship changes. The research questions were related to whether independent variables like gender, age, or smoking behavior had an influence on the formation of friendship ties.

In SAOMs, the dependent variables are actor choices about changes of social network relations. The independent variables can be endogenous or exogenous (see Snijders et al. (2010b, p.369)). Endogenous independent variables explain network changes with existing network structures. An endogenous structure may express whether the out-degree or the number of reciprocated relations has an effect on changes in the set of individually controlled outgoing network ties. Exogenous independent variables explain network changes with actor covariates or dyadic covariates. An exogenous actor covariate may measure whether an actor's gender has an effect on individual decisions on network changes. Dyadic, exogenous covariates may measure whether similarity in gender has an effect on individual decisions on network changes, e.g., whether female actors prefer female over male friends. Some commonly used independent variables will be presented below.

A SAOM is defined as a two-level Markov process as explained in Snijders (2005, p.224–227). In the basic form, the model depends only on the state of a graph $x(t)$ at time t, a set of actor covariates, and dyadic covariates. A rate function $\lambda_i(x)$ (see Snijders (2005, p.224)) determines the general "activity" of each actor i. It depends on the communication graph x. The time δ^i between two configuration changes of actor i is $\lambda_i(x)$-exponentially distributed:

$$P(\delta^i \geq \delta; \lambda_i(x)) = e^{-\lambda_i(x)\delta} \tag{2.21}$$

with δ being any positive time span.

An actor-driven change in the graph x (a creation or a dissolving of an outgoing tie) is called a *ministep*. As the underlying data are panel data, ministeps are not observed directly. Only the number of changed ties between two observed networks indicates the number of ministeps that probably occurred.

An objective function $f_i(x)$ (see Snijders (2005, p.225)) is an evaluation of a network x by actor i. This evaluation is non-stochastic and, hence, not a random utility function. However, a random utility function can be constructed by adding an additional random variable with mean 0 to $f_i(x)$. This random variable describes the unexplained part of the actor evaluation of x. This idea equals the utility function of multinomial logit models in equation 2.11. If an actor is assumed to change the configuration of his or her individually controlled outgoing ties, the objective function evaluates the network *after* a corresponding change. A probability distribution describes this choice among all possible changes (creations or dissolving of ties). The actor choice of the actually changed tie is modeled with a random utility model, as introduced by Marschak (1960). In Snijders (2005, p. 226, equation 11.17) the probability of actor i changing the relational tie to j is defined as

$$p_{ij}(x) = \frac{\exp(f_i(x(i \leadsto j)))}{\sum_{k=1,k \neq i}^{n} \exp(f_i(x(i \leadsto k)))}, j \neq i. \tag{2.22}$$

$x(i \leadsto j)$ is the state of graph x after the tie update of x_{ij}. This update or ministep is defined either as a removal or as an inclusion of a tie. n is the number of nodes in the graph. Multinomial logit models describe discrete choices. In this context, the choice lies within the set of possible relationship updates $i \leadsto k$. The graph resulting from each choice is described with $x(i \rightsquigarrow k)$. Note that with a linear objective function the random utility model in equation 2.22 equals a multinomial logit model as introduced by McFadden (1974). The probability of creating or dissolving a tie to another actor is evaluated by the objective function applied to the "updated" graph. Therefore, the model cannot distinguish between two choices that lead to the same objective functions, although one structure may have been established by a tie dissolving and another structure by a tie creation. To measure this difference, the multinomial model can be extended by an additional *gratification function* (see Snijders (2005, p.226, equation 11.19)) $g_i(x,j)$:

$$p_{ij}(x) = \frac{\exp(f_i(x(i \rightsquigarrow j))) + g_i(x,j)}{\sum_{k=1,k\neq i}^{n} \exp(f_i(x(i \rightsquigarrow k))) + g_i(x,k)}, j \neq i. \tag{2.23}$$

Below, the basic form from equation 2.22 will be used.

SAOMs assume independence between the two decision levels (actor activity and chosen tie). The Markov process transition rates can then be defined as

$$q_{ij}(x) = \lambda_i(x) p_{ij}(x) \tag{2.24}$$

with $p_{ij}(x)$ being defined as in equation 2.22 or 2.23.

A concrete model can be specified, which takes certain research questions into account: It may be of interest whether there are endogenous or exogenous effects that define the probability of certain tie changes in the graph. It is possible to specify the rate function $\lambda_i(x)$ as explained in Snijders (2005, p.232). However, as the number of observed actors (n) is much smaller than the number of observed network ties $(n \cdot (n-1))$, more effort is usually made to specify the objective function in practical applications. In many cases, the rate function is even simplified to a homogeneous parameter ρ that reflects the *general* level of actor activity in the observed social network:

$$\forall i : \lambda_i(x) := \rho \tag{2.25}$$

The objective function $f_i(x)$ is usually defined as a linear function that additionally depends on a parameter vector β (see Snijders (2001, p.369) and Snijders et al. (2010b, p.9, equation 1)):

$$f_i(\beta, x) = \sum_k \beta_k s_{ki}(x) = \beta^T s_i(x) \tag{2.26}$$

where i indicates the actor evaluating the set of his or her outgoing ties. $s_i(x)$ is a vector of statistics measured in x with elements $s_{ki}(x)$. Each statistic $s_{ki}(x)$ is weighted by a real value β_k in vector β. Specifying probability 2.22 with a linear objective function returns a multinomial logit model, as shown in equation 2.12 and introduced by McFadden (1974).

Concrete specifications of the linear objective function are introduced in Snijders et al. (2010b, p.10ff). The two simplest statistics measure outdegree and number of reciprocated ties after the decision. Especially the outdegree effect is fundamental. Snijders (2001, p.371) states that it is "advisable to include the [outdegree] effect" in "practically all applications". A rational and strategic actor is assumed to have a limited number of outgoing ties, because the creation and maintenance is costly. This is reflected by the fact that most social networks have a density that is significantly below 1. Snijders et al. (2010b, p.10) state that the effect can be "regarded as the balance of benefits and costs of an arbitrary tie". The outdegree effect is defined as (see Snijders (2001, p.369, equation 1)):

$$s_{1i}(x) = \sum_h x_{ih}. \tag{2.27}$$

A negative parameter β_1 would usually ensure that the overall network density is limited.

The reciprocity effect (Snijders, 2001, p.369, equation 2) counts the number of reciprocated ties in a new network configuration:

$$s_{2i}(x) = \sum_h x_{ih}x_{hi}. \tag{2.28}$$

Reciprocity (measured by mutually connected actors) is a common structure in social networks (Wasserman and Faust, 1994, p.507). Snijders (2001, p.371) phrases that the "reciprocity effect is so fundamental" that it should be included "in most applications".

Other endogenous structures include third actors (Snijders et al., 2010b, p.11–12) like the *transitivity* effect

$$s_{3i}(x) = \sum_{h,k} x_{ih}x_{ik}x_{hk} \tag{2.29}$$

that is defined in Snijders (2001, p.370, equation 5). Other endogenous variables are degree-related (Snijders et al., 2010b, p.12–13). Exemplarily, the effect of indegree popularity is

$$s_{4i}(x) = \sum_h x_{ih} \sum_k x_{kh} \tag{2.30}$$

defined in Snijders (2001, p.370, equation 3). Besides the endogenous variables that explain the choices of changed ties by structures in the graph x, different exogenous variables can be tested. These statistics can be related to attributes of the chosen actor h. The actor may have a certain attribute (e.g., gender) that is denoted by covariate v_h in $\{0,1\}$. Snijders (2001, p.371, equation 8) defines the following exogenous statistic

$$s_{5i}(x) = \sum_h x_{ih}v_h \tag{2.31}$$

that measures whether outdegree varies according to the gender of the chosen actor. Dyadic covariates (e.g. gender similarity) w_{ih} for the relation between actors i and h can be included as exogenous effects as well. Snijders (2001, p.372, equation 11) propose the following form:

$$s_{6i}(x) = \sum_h x_{ih}w_{ih} \tag{2.32}$$

An extensive list of independent variables (implemented in the software tool SIENA) can be found in Ripley et al. (2011).

The specification of basic SAOMs is straightforward. The estimation of parameters, however, is rather complicated. As the underlying data are usually panel data that do not include information on ministeps, a straightforward maximum likelihood estimation is not possible. In fact, the sequences of possible ministeps that lead to the observed networks at a certain point in time are simulated until a model is found that explains the observed networks as good as possible. The applied algorithm belongs to the class of Markov chain Monte Carlo (MCMC) algorithms and uses the method of moments instead of the maximum likelihood function. A brief introduction to the estimation algorithm is given in Snijders (2005, p.234ff). The rate functions $\lambda_i(x)$ (or usually rather the simplified rate ρ) can be estimated separately from the parameter vector β of the objective function.

In recent years, the class of SAOMs gained increasing popularity, especially in social sciences. Five exemplary studies are mentioned in the following to sketch the broad applicability of the framework. van de Bunt et al. (1999) and van Duijn et al. (2003) analyze the evolution of friendship relations between university freshmen. The second study, for example, finds that proximity and visible similarity both increase the likelihood of friendship formation in the early stages. Agneessens and Wittek (2008) model the evolution of trust relations between employees in a Dutch housing company and its relation to individual well-being. It is shown that individuals with a low level of job satisfaction are more likely to create trust ties to colleagues. Johnson et al. (2009) apply the model to the dynamics of relations between ecosystem components such as species in water layers of different depth. Exogenous factors such as seasons and human impact are considered in the dynamic model. Lubbers et al. (2010) focus on personal support networks among Argentinean migrants in Spain. One of their findings is that the density of the personal network has a positive influence on the stability of social ties within.

The basic model presented here has been extended in several publications: The analysis of bipartite networks is introduced in Koskinen and Edling (2010). Snijders et al. (2010a) discuss a new maximum likelihood estimation algorithm. There are interesting new extensions that define a combined model for selection processes (creation and dissolving of ties) and influence processes (change of individual attributes). A good introduction to combined selection and influence models is given by Steglich et al. (2010).

2.4 Event History Analysis and the Relational Event Framework

Event history modeling has become increasingly popular since the early 1980's (Blossfeld and Rohwer, 1995). Until then, data collected and analyzed by social scientists mostly were either cross-sectional data representing a state of a researched population of individuals at a particular point in time or panel data consisting of a series of cross-sectional observations at discrete points in time. Event history analysis takes advantage of the fact that many dynamic trends can be represented best by the exact points in time when some state of an object changes. These changes are named *events*. Events can be very different, for example, births and deaths in demographic analyses, changes of governments or regimes in political sciences, or changes in consumer behavior in marketing research (see Blossfeld and Rohwer (1995, pp.1–2)). In this book, we are interested in directed *relational events* between social actors. Event history analysis for relational event data has been used in the political sciences for many years to understand changes in state relationships (see, e.g., Box-Steffensmeier and Jones (1997)). An introduction to event history modeling in general and its enhancements over time was written by Blossfeld and Rohwer (1995).

The occurrence of events over time in event history models is described by two competing functions, a *transition rate* (sometimes called hazard function) that expresses the tendency of an event to occur and a *survival function* that models the length of an interval without changes. Events change the state of individuals in the population. A simple way to model these competing functions is by describing the occurrence of events as an *exponential transition rate model* (see Blossfeld and Rohwer (1995, p.80ff)). This class is closely related to Markov processes as applied in stochastic actor-oriented models.

In exponential transition rate models the time until an event ω_i occurs is exponentially distributed with a parameter λ_i. For each possible event ω_i, λ_i is the *transition rate* of the event history model.

The *survival function* is defined as follows. Within a certain time span, many events may possibly occur. The set $A(t)$ includes all possible next events $\{\omega_1', \omega_2', \dots\}$ at a time t. The corresponding transition rates are $\lambda_1, \lambda_2, \dots$. The time until the next event occurs can be described by an exponential function with a parameter $\lambda = \lambda_1 + \lambda_2 + \dots$. The exponentially distributed probability of occurrence of any event within a time span δ is

$$P(t \leq \delta) = 1 - \exp(-\delta\lambda) \tag{2.33}$$

where t is the time of the “early event”. The probability of no event taking place is then

$$1 - P(t \leq \delta) = P(t > \delta) = \exp(-\delta\lambda). \tag{2.34}$$

This probability defines the *survival function* of the exponential transition rate model.

The likelihood of one particular event can be described by a multiplication of transition rate and survival function. This follows from the fact that – given the time span until the next event is at least of length δ – the probability of observing the particular event ω_i is given by $\frac{\lambda_i}{\lambda}$ (as long as the transition rates are independent, see Waldmann and Stocker (2004, p.92)):

$$\begin{aligned} L(\omega_i; \delta) =& \lambda \exp(-\lambda\delta) \cdot \frac{\lambda_i}{\lambda} \\ =& \underbrace{(\lambda_i)}_{\textit{transition rate}} \cdot \underbrace{\exp(-\lambda\delta)}_{\textit{survival function}} \end{aligned} \tag{2.35}$$

The transition rates for each possible event can be parameterized to take covariates into account. Covariates can, for example, be measured on the current set of states in the population. Another interpretation of this process is that of a Markov process where the λ_i are Poisson transition rates. As soon as variables are considered when determining the transition rates, these variables can be interpreted as defining the state space of the Markov process (e.g., see Waldmann and Stocker (2004)). In the case of formulating a Markov process, it has to be made sure that all relevant information is part of the process state. The process state may also include the complete event history.

Based on the exponential transition rate model, Butts (2008) introduced a *relational event framework* that allows to model sequences of relational events between the roles of senders and receivers. A directional action may be sent from a sender to a receiver. Here, an event models the change of a relational state. Senders and receivers can be individuals but also "collective entities, sets of individuals, or even inanimate objects" (Butts, 2008, p.159). The relational event framework and one extension will be presented below. Compared to the original papers a simplified notation is used, which is closer to the notation introduced later in this book.

An observed event stream $\Omega = \{\omega_1, \omega_2, \ldots\}$ is defined as an ordered set of events. Each event ω_v is defined by a sender i_v, a receiver j_v, and a timestamp t_v:

$$\omega_v = (i_v, j_v, t_v) \tag{2.36}$$

Additionally, Butts allows to distinguish between events of different type k_v. The time span between two consecutive events ω_v and ω_{v-1} is defined as

$$\delta_v = t_v - t_{v-1} \tag{2.37}$$

If within a time interval $\tau = [t_0, t_{max}]$ an event stream Ω_τ is observed with $t_0 \leq t_1, \ldots, t_{|\Omega_\tau|} \leq t_{max}$, then all time intervals before the first event, between the events, and between the last event and the end of the observed time span have to be considered, when the stream Ω_τ is modeled as an event history model. The ordered set Δ_τ represents these intervals:

$$\Delta_\tau = \{t_1 - t_0, \delta_2, \ldots, \delta_{|\Omega_\tau|}, t_{max} - t_{|\Omega|}\}. \tag{2.38}$$

The time-independent (but parameterizable) transition rate for a relational event $(i \rightarrow j)$ to occur (for simplicity, we do not distinguish different event types) is defined by a Poisson rate λ_{ij}. This rate may depend on a set of covariates. For each point in time, the set of potentially occurring events can be restricted. Butts argues that, "for instance, certain actions may remove a potential target from the receiver set, or enable new types of actions to be taken" (Butts, 2008, p.161). Let $A(t)$ be this set of potential events for a point in time t. It includes ordered tuples $(i \rightarrow j)$ that indicate potential senders and receivers. For reasons of simplicity, we only show the case of $A(t)$ being independent of the current time: $A(t) = A$.

The likelihood $L(\Omega_\tau)$ of the observed sub-event stream Ω_τ introduced in Butts (2008, equ. (2), p.163) is defined similarly to equation 2.35 by separating transition rates and survival functions:

$$L(\Omega_\tau) = \overbrace{\prod_{\omega_v \in \Omega_\tau} \lambda_{i_v j_v}}^{\textit{transition rates}} \overbrace{\prod_{\delta \in \Delta_\tau} \prod_{(k \rightarrow l) \in A} \exp\left(-\delta \lambda_{kl}\right)}^{\textit{survival functions}} \tag{2.39}$$

The left part represents the actually observed transition rates $\lambda_{i_v j_v}$ with sender i_v and receiver j_v. The right part represents the potential, but not observed transitions in the intervals between two events for any possible sender-receiver combination.

From this model which assumes that the underlying data includes information about exact timestamps, a sub-model is derived. It only takes the order of events into account (ordinal data), but not information about timestamps. This is similar to the idea of simplifying a (continuous-time) Markov process to a Markov chain with transition rates being independent from the timestamp. Given several independent Poisson transition rates $\lambda_1, \lambda_2, \ldots$ from one state in a Markov process, the probability of a specific transition with rate λ_k equals $\frac{\lambda_k}{\lambda_1+\lambda_2+\ldots}$ as long as the rate parameters are independent (see proofs in Waldmann and Stocker (2004, p.92) and Butts (2008, p.164)).

As shown in Butts (2008, equation 3, p.165), the likelihood of the parameterized time-discrete sub-model is

$$L(\Omega_\tau) = \prod_{\omega_v \in \Omega_\tau} \left(\frac{\lambda_{i_v j_v}}{\sum_{(k \to l) \in A} \lambda_{kl}} \right) \tag{2.40}$$

Butts proposes to parameterize the λ parameters with an exponentially transformed linear function that can take into account attributes of event ω_v (sender i_v, receiver j_v, or type k_v), covariates (in a set X_v) and the complete event history $(\Omega_{\tau'}, \tau' = v - 1)$:

$$\lambda_{i_v, j_v} = \exp\left(\beta_0 + \beta_1 s_1(i_v, j_v, k_v, X_v, \Omega_{v-1}) + \ldots\right) \tag{2.41}$$

As in the stochastic actor-oriented models, it is proposed to test, for example, the effect of endogenous network structures on tie emergence. Equation 2.40 can then be transformed using these parameterized exponential transition rates. As proposed in the "receiver choice sub-model" of Snijders (2005), the determination of concrete events is therefore modeled as a multinomial probability. However, it is not formulated as an actor-oriented model like the multinomial logit model introduced by McFadden (1974) (see equation 2.12). The multinomial probability instead expresses a tendency that a relational event occurs. The denominator does not represent a set of individual decisions, but the set of possible events, including all sender-receiver-type combinations allowed in the current state. Butts names the model "behavior-oriented" (Butts, 2008, p.167) in contrast to actor- and tie-oriented models.

Brandes et al. (2009) introduce an application of the relational event model, in which the original ideas of Butts (2008) are extended by modeling event weights. In their analysis of political events between states, the authors incorporate the level of hostility or cooperativeness on a continuous scale. This approach also was not considered an actor-oriented model driven by individual choices. Together with the basic relational event model, the work provides a highly flexible framework to evaluate event history models in which the changing state is a relation between two entities. In section 3.4.3 these relational event models are compared with the event framework introduced in this book.

2.5 Summary

In this chapter, we introduced three different stochastic frameworks for the analysis of structural effects in social networks.

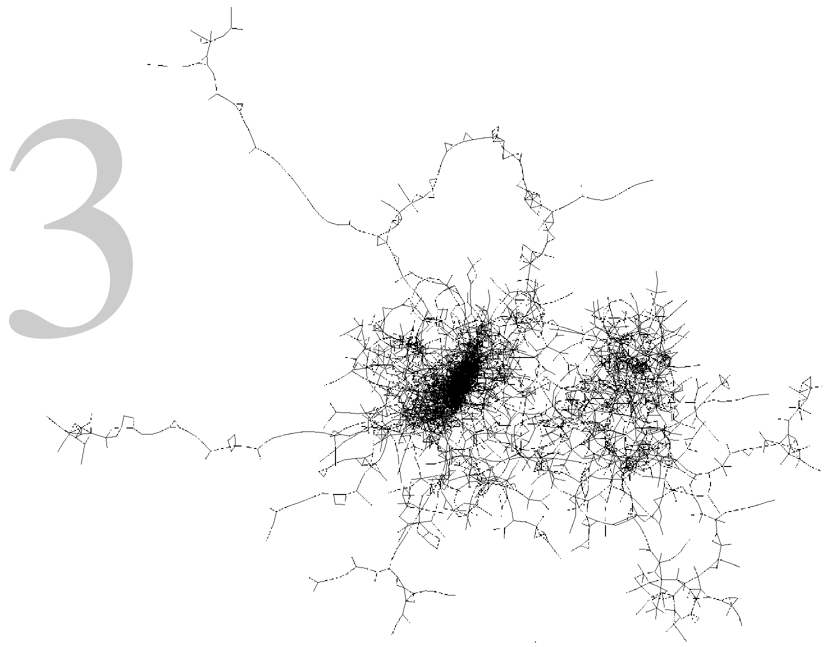
Exponential random graph models (ERGMs) are defined for cross-sectional network data, as a result of which they cannot be applied to event stream data.

Stochastic actor-oriented models (SAOMs) are a dynamic advancement of ERGMs for panel data. The core of this model is a multinomial choice model. SAOMs incorporate the assumption of an underlying, unobserved ministep process, which is similar to the idea of event stream data. However, SAOMs cannot be applied directly to event stream data, as the ministep process only models a concrete creation and dissolving of ties, which is different from the kind of event streams usually observed.

The relational event framework, an extension of event history modeling, can be specified for any type of event stream. It is, however, not explicitly actor-oriented, it does not define update rules of events and does not allow external processes to change the process state.

In chapter 3, we will introduce a new basic event framework that is based on the actor-oriented ideas of SAOMs. It will be compared to the structural network models presented in this chapter. In chapter 4, the basic model will be extended by introducing advanced specifications.

3 Basic Event Framework



An earlier version of this chapter is published in Stadtfeld et al. (2010). New parts of this chapter have been written during a research visit at the University of Melbourne (funded by Karlsruhe House of Young Scientists, KHYS) in collaboration with Garry Robins and Philippa Pattison.

In this chapter, a basic event model framework is presented. The framework aims at providing a new tool for the analysis of actor-oriented decisions in networks that are represented by dynamic event streams. This framework can be specified flexibly, as it offers many possibilities for extensions. Possible extensions are shown in chapter 4. However, in this chapter we consider the basic case of communication events between actors (e.g., phone calls) that are described by individualized activity rates and stochastic, multinomial decisions depending on dyadic relations between the communication partners in the communication network. The number of dyadic events between two actors is a good indicator of communication intensity. The level of communication intensity can be encoded in a communication graph which is updated every time when an event takes place. The model of the graph can be used as a predictor for future events.

The underlying data, the representation in a graph and the updates of the graph are explained in section 3.1. The basic actor-oriented decision process is explained in section 3.2. The process consists of two decision levels, which can be analyzed separately. The first level describes the points in time when actors communicate, the second level describes the choice of communication partners. In section 3.3, the sub-model explaining the choice of communication partners is discussed in detail. The estimation process is explained and information about the interpretation of estimates is given. After this generic introduction, in section 3.4 the model is compared with the network models introduced in chapter 2 of this book.

3.1 Data, Graph and Updates

Information about social network dynamics can often be understood by evaluating *event data* that represents dyadic, time-stamped interactions of different type. The example we consider in this chapter is communication data, where each communication choice (e.g., a phone call from one person to another) can potentially be logged for later analyses of individual communication choices. We assume the existence of a communication network that is represented by a communication graph.

3.1.1 Event Stream

An event stream Ω is a chronologically ordered, finite list of events

$$\Omega = (\omega_1, \omega_2, \dots, \omega_v, \dots, \omega_{|\Omega|}). \tag{3.1}$$

Events ω_v are defined as triplets

$$\begin{aligned} \omega_v &= (\omega_v.\mathit{sender}, \omega_v.\mathit{receiver}, \omega_v.\mathit{time\text{-}stamp}) \\ &= (i_v, j_v, t_v), \end{aligned} \tag{3.2}$$

where the first element indicates the sender of the event (the person who starts a phone call), the second element indicates the receiver (the person answering the call), and the third element indicates the time-stamp when the event takes place. For each pair of events ω_ν (ν is the index $\in \{1,\dots,|\Omega|\}$) and ω_ϕ holds

$$\nu < \phi \Leftrightarrow t_\nu \le t_\phi. \tag{3.3}$$

The elements i_ν and j_ν are indices of two actors in the set of actors

$$A = \{a_1,\dots,a_{|A|}\}. \tag{3.4}$$

The sending actor is $a_{i_\nu} \in A$, the receiving actor is $a_{j_\nu} \in A$. Whenever actor a_{i_ν} starts an event ω_ν as sender, he can choose any actor a_k in $A \backslash \{a_{i_\nu}\}$ as a receiver. The index set is defined as

$$R_\nu = \{1,\dots,i_\nu - 1, i_\nu + 1,\dots,|A|\} \tag{3.5}$$

The time-stamps t_ν have values in $\mathbb{R}^+$. The time between two subsequent events is named

$$\delta_\nu = t_\nu - t_{\nu-1} \tag{3.6}$$

Let us consider the exemplary phone call event stream in table 3.1. Events in this stream are observed between five actors $A = \{a_1,\dots,a_5\}$. The stream includes nine events $\Omega = \{\omega_1,\dots\omega_9\}$ in the continuous time span $t_\nu \in [0,50]$. In the first event ω_1, for example, actor a_1 is observed to call actor a_2 at time $t_1 = 2$. The next event ω_2 includes sender a_2 and receiver a_3 and takes place at time $t_2 = 4$. The time span δ_2 between the first two events is t_2 minus t_1 and equals 2.

ν	i_ν	j_ν	t_ν	δ_ν
1	1	2	2	-
2	2	1	4	2
3	2	5	6	2
4	5	3	8	2
5	5	4	12	4
6	1	4	24	12
7	4	5	40	16
8	5	2	48	8
9	1	4	50	2

Table 3.1: Exemplary phone call event stream Ω with five actors $a_1,\dots,a_5 \in A$ and nine events $\omega_1,\dots,\omega_9$ with index ν (one event per row) measured in the time span $t \in [0,50]$

This exemplary event stream will be used to illustrate the basic event framework in this chapter.

3.1.2 Communication Graph

The set of actors A and the communication intensities between them are reflected by the stream of events Ω. The "real" communication intensities of two actors in a communication network are unknown but can be inferred from the observations in the event stream. For each point in time

$$t \in [t_0, t_{max}] \text{ with } t_0 \leq t_1, t_{max} \geq t_{|\Omega|},$$

communication intensities can be represented in a directed, weighted *communication graph* $X(t)$. A graph consists of vertices and edges. Vertices of the communication graph represent the actors occurring in the event stream, i.e. senders and receivers. Directed edges of the communication graph represent communication intensity.

$$X(t) = (x_{kl}(t)), k,l \in \{1, \ldots, |A|\} \tag{3.7}$$

is the adjacency matrix of the communication graph at time t. The elements are in $\mathbb{R}_0^+$ and are defined for all ordered actor combinations $(a_k, a_l) \in A \times A$. Only reflexive edges $x_{kk}(t)$ are not defined. This excludes communication to oneself which is usually not measurable in computer-mediated communication data. Two rules change the communication graph:

1. When an event ω_ν takes place, the value of the directed edge from actor a_{i_ν} to a_{j_ν} is increased in the graph.
2. During the time span between two events, edge values decay.

Those two rules are the basic graph update rules. They are introduced in detail in section 3.1.3. As edges represent communication relations, we may also refer to them as *relational ties* or *ties* (see Wasserman and Faust (1994, p.18)).

Of specific interest are two graph states, one right before an event ω_ν takes place ($X(t_\nu)$), and one right after the event ($X(t_\nu - \varepsilon)$), hence after the event triggered changes have been realized. The process is assumed to be right-continuous: This means that at the time of an event we assume the graph to be already updated. These two relevant graphs are abbreviated in the following way:

$$X_\nu = (x_{\nu;kl}) := X(t_\nu) \qquad \text{(after } \omega_\nu\text{)} \tag{3.8}$$

$$X_{\tilde{\nu}} = (x_{\tilde{\nu};kl}) := X(t_\nu - \varepsilon) \qquad \text{(before } \omega_\nu\text{)} \tag{3.9}$$

The elements of these graphs are named $x_{\nu;kl}$, where k and l in $\{1, \ldots, |A|\}$ are line and column index of X_ν, and ν stands for the corresponding event index. ε is a very short time span so that the following approximate relation generally holds

$$X(t) \approx X(t + \varepsilon) \tag{3.10}$$

if no events take place in the time span $\in [t, t+\varepsilon]$. During very short time spans, the network is approximately stable, although tie values decay slowly. Thus, the network states $X_{\tilde{\nu}}$ and X_ν are assumed to be almost equal, except for an event-triggered update of tie $x_{i_\nu j_\nu}$. In case of simultaneous events, several ties may be updated simultaneously. For events at the same time holds that

$$t_\nu = t_{\nu+1} \Rightarrow X(t_\nu) = X(t_{\nu+1}). \tag{3.11}$$

Later in this chapter, we model the time spans between two events as realizations of a continuous exponential probability distribution. Note, that this allows the assumption that two events never take place at exactly the same time. However, because in practical analyses the accuracy of measurement is limited, the case that two events take place at the same time is possible and hence considered here. Moreover, some means of communications (e.g. e-mail) allow sending several events at the same time.

Figure 3.1 shows the nine graphs that are generated by the exemplary event stream in table 3.1. With each event a new tie is added to the graph as long as it does not already exists. The decay of tie values is not depicted. The events ω_6 and ω_9 have the same sender and receiver (a_1 and a_4). Therefore, ω_9 does not change sub-figure (i) in figure 3.1.

3.1.3 Graph Update Rules

People do not communicate randomly with one another. They would usually, for example, prefer those as communication partners who they have communicated with before, or have common contacts with. Therefore, the occurrence of events should be influenced by structures in the communication graph X.

The communication graph is changed either if events take place, or due to a decay of ties over time. A communication event ω_ν leads to the update of the network X. This is modeled by increasing the edge weight between actors a_{i_ν} and a_{j_ν} by 1 at time t_ν. Since X is a continuous-time network, inactivity leads to an exponential decay of all edge weights with a defined half-life. It is reasonable to assume that the communication intensity between two actors decreases over time if no further communication takes place. Otherwise high communication levels would remain stable and the assumed underlying communication network could only grow.

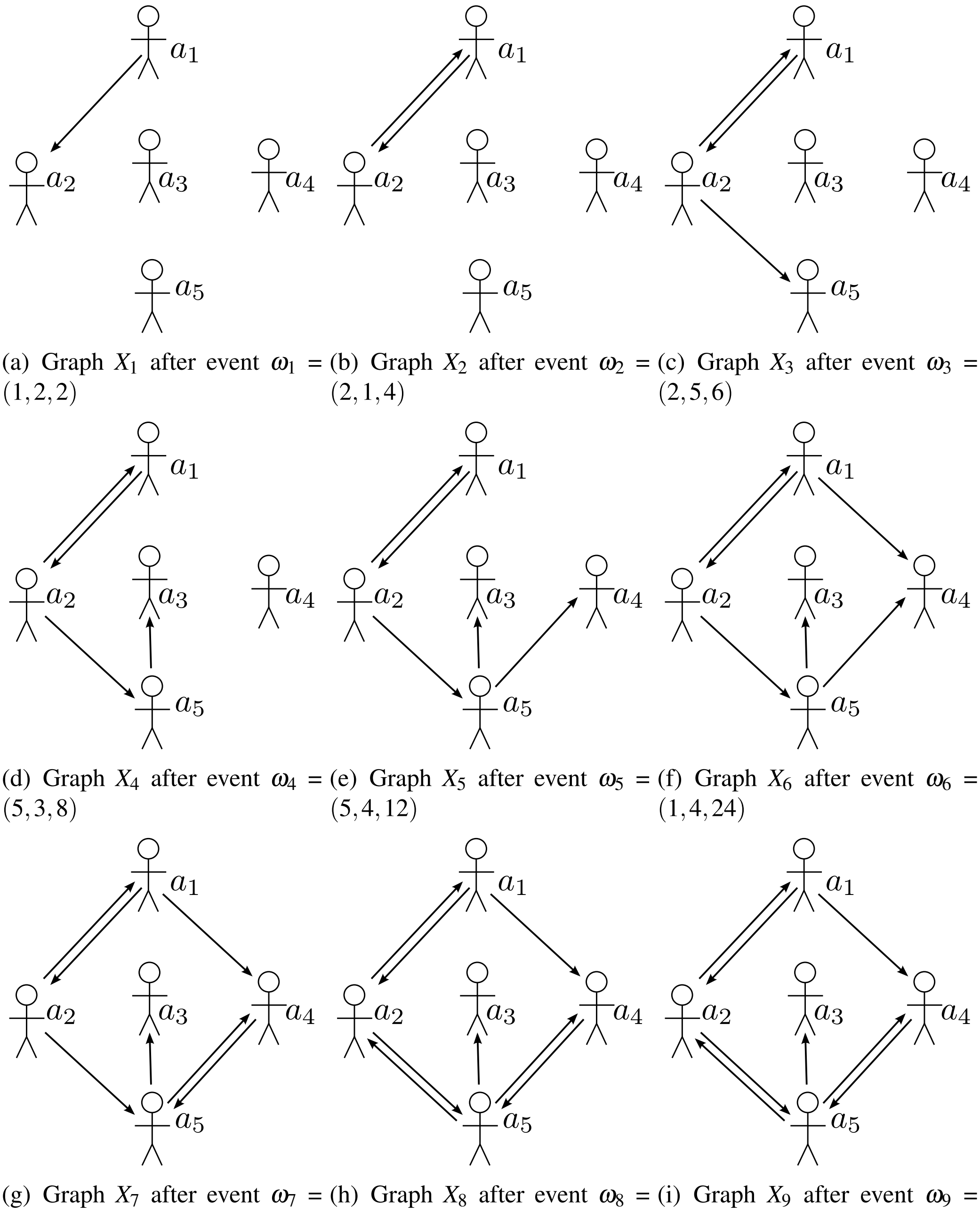
An update triggered by events is defined as follows: If an event ω_ν takes place, the tie value from actor i_ν to j_ν is increased by 1. This transformation is described by a function $r^{(1)} : X \times \Omega \mapsto X$:

$$r^{(1)}(X_{\tilde{\nu}}; i_\nu, j_\nu) = X_\nu = (x_{\nu;kl}) = \begin{cases} x_{\tilde{\nu};kl} + 1 & \text{, if } k = i_\nu, l = j_\nu \\ x_{\tilde{\nu};kl} + 1 & \text{, if } \exists \omega_\phi : k = i_\phi, l = j_\phi, t_\phi = t_\nu \\ x_{\tilde{\nu};kl} & \text{, else} \end{cases} \tag{3.12}$$

In between two events, tie values decrease with an exponential decay function $r^{(2)} : X \times \mathbb{R}^+ \mapsto X$. The decrease depends on the time span δ_ν between the current and the previous event.

$$r^{(2)}(X_{\nu-1}, \delta_\nu; \tau_{\frac{1}{2}}) = X_{\tilde{\nu}} = (x_{\tilde{\nu};kl}) = (x_{\nu-1;kl} \cdot e^{-\theta\delta_\nu}), \theta = \frac{\ln 2}{\tau_{\frac{1}{2}}} \tag{3.13}$$

$$\delta_\nu = t_\nu - t_{\nu-1}$$



(a) Graph X_1 after event $\omega_1 = (1,2,2)$

(b) Graph X_2 after event $\omega_2 = (2,1,4)$

(c) Graph X_3 after event $\omega_3 = (2,5,6)$

(d) Graph X_4 after event $\omega_4 = (5,3,8)$

(e) Graph X_5 after event $\omega_5 = (5,4,12)$

(f) Graph X_6 after event $\omega_6 = (1,4,24)$

(g) Graph X_7 after event $\omega_7 = (4,5,40)$

(h) Graph X_8 after event $\omega_8 = (5,2,48)$

(i) Graph X_9 after event $\omega_9 = (1,4,50)$

Figure 3.1: The nine graphs $X_\nu = X(t_\nu)$ $(\nu \in 1,\ldots,9)$ of the example event stream from table 3.1. Decay of ties is not considered here.

Parameter $\tau_{\frac{1}{2}}$ is the half-life of all tie values, which is defined as the time span after which a tie has halved its weight. It is independent from the absolute tie value.

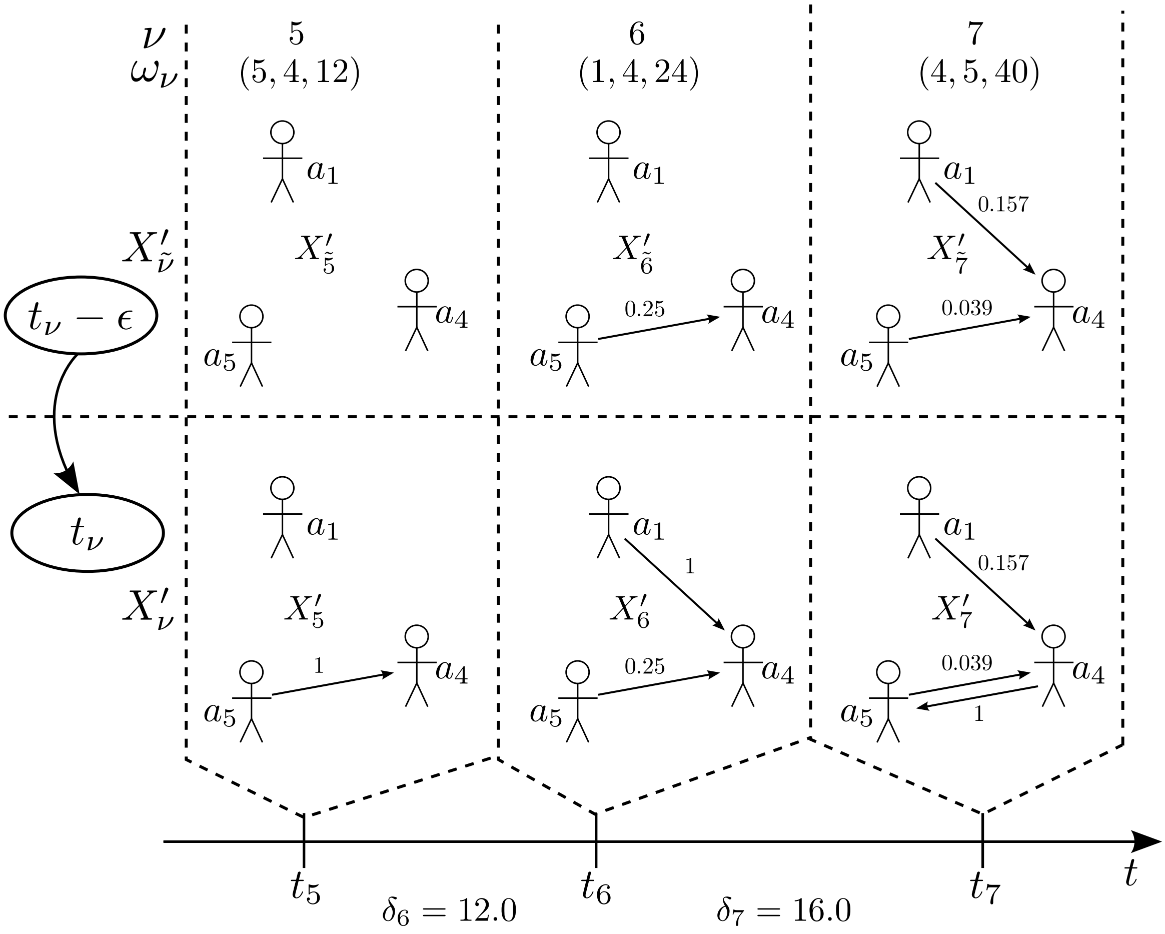


Figure 3.2: Two types of update rules are defined for the exemplary event stream from table 3.1: Ties are increased by 1 when events take place. In the time spans between events, ties decay exponentially with a half-life of $\tau_{1/2} = 6$.

In figure 3.2, this process is exemplarily sketched for the three events $\Omega' = \{\omega_5, \omega_6, \omega_7\}$ of the event stream in table 3.1. Only a sub-graph of the communication graph X is illustrated. It includes the three actors a_5, a_6 and a_7 ($\in A' \subset A$). The corresponding graph is named $X' \subset X$. As no events took place within X' before time t_5, the graph $X'_{\tilde{5}}$ just before event ω_5 is empty.

$X'_{\tilde{5}}$, $X'_{\tilde{6}}$ and $X'_{\tilde{7}}$ in the lower part of the figure represent the graphs immediately before the events ω_5, ω_6 and ω_7 take place. X'_5, X'_6 and X'_7 in the upper part of figure 3.2 represent the graphs at the time of the events ω_5, ω_6 and ω_7, respectively. It can be seen that two update rules are applied: With each event the corresponding tie from sender to receiver is increased by 1, as shown in equation 3.12. Between events, tie values decay with an exponential decay function, as in equation 3.13. The half-life was set to six time units ($\tau_{1/2} = 6$). For example, in time span δ_6 between events ω_5 and ω_6 tie x'_{54} decays by 75% as the time span is twice as long as the half-life ($\delta_6 = 12.0$). Note that although all time stamps are natural numbers in the example data stream of figure 3.1, the time scale is continuous. The indices v of events ω_v are discrete.

3.2 Basic Actor-oriented Decision Process

The emergence of events in an event stream is driven by actor choices. In the continuing example of table 3.1, people decide whether to make phone calls at all, and whom they want to call. The actor-oriented decision in an event stream can be embedded in a stochastic process. In its basic form, these choices can be expressed by two decision levels:

1. Every actor has a personal activity rate that determines whether and at what time an event is sent.

2. If an actor starts an event, the choice of the actual event receiver is influenced by network structures embedding sender and potential communication partners at the event time.

The process is modeled as a Markov process, which means that at each point in time all relevant information for the process transition is encoded in the state of the process.

3.2.1 Actor Activity Rate

Informally, the probability of the first decision of an actor a_i (to send an event) in a time span $[t, t+\varepsilon]$ is

$$P(a_i \text{ active in } [t, t+\varepsilon]) \tag{3.14}$$

where ε is a very short time span, so that a_i is active once or never. The points in time when actors start events are determined by a Poisson process with parameters ρ_i for actor $a_i \in A$. This also means that the time δ^i between two subsequent events started by the same sender a_i is ρ_i-exponentially distributed with an individual parameter ρ_i (see Waldmann and Stocker (2004, p.64)):

$$P(\delta^i < \delta; \rho_i) = 1 - e^{-\rho_i \delta} \tag{3.15}$$

is the probability that actor a_i starts at least one event in a time span δ. It is assumed to be independent from the decisions of any other actor to start events. Given a data stream, this actor activity can easily be estimated with a maximum likelihood (ML) estimation (see Davison (2003)). Actor activity rates can also be parameterized. However, in this basic model it is only assumed that different actors have individual activity rates, without aiming at further explaining existing variances with additional parameters. The model can be simplified by assuming homogeneous activity rates for all actors in an observed event stream.

3.2.2 Choice of Receivers

After the decision of sending an event, the second decision of actor a_i is choosing a communication partner (event receiver). This decision depends on the state of the communication graph $X(t+\delta^i)$ at time $(t+\delta^i) \in [t, t+\varepsilon]$. From equation 3.10 follows that $X(t+\delta^i) \approx X(t)$ (if no other actors have started an event in this time span). Informally, the probability that a_i would choose a_j as the event receiver (given a_i made a choice at all) at a time in $[t, t+\varepsilon]$ is then

$$P(a_i \text{ chooses } a_j | X(t)). \tag{3.16}$$

Since this decision is independent from the first decision as long as ε is very small and the network has not been changed by other actors, the probability of a new communication event between a_i to a_j in time span $[t, t+\varepsilon]$ is then

$$P(a_i \text{ active in } [t, t+\varepsilon]) P(a_i \text{ chooses } a_j | X(t)). \tag{3.17}$$

It is assumed that the second decision on receivers depends on structures in subgraphs of $X(t+\delta^i)$, but not on individual activity rates.

The choice of a certain event receiver out of a set of possible event receivers can now be modeled as a multinomial logit model (originally named *conditional logit model*). This model was introduced by McFadden (1974) and is shortly explained in Cramer (2003, p.130ff). This class of probability models for qualitative choice behavior was already described in section 2.1.

$$P(a_i \text{ chooses } a_j | X(t)) = \frac{\exp\left(\beta^T s(i, j, X(t))\right)}{\sum_{k=0, k\neq i}^{|A|} \exp\left(\beta^T s(i, k, X(t))\right)} \tag{3.18}$$

The choice of the event sender a_i to select actor a_j as the event receiver over any other actor in the set of potential receivers $A \backslash \{a_i\}$ depends on a vector of *choice statistics* $s(i, j, X(t))$. This vector of statistics is weighted by a vector β of the same dimension. The choice statistics can be understood as covariates in a regression model that predicts the multinomial choice of event receivers. The covariates could, for example, describe whether the sender and the receiver have communicated before. The general form of the choice statistics vector $s(i, j, X(t))$ is described in the next section. A concrete, dyadic example that explains receiver choices based on previous communication between sender and receiver is introduced in section 3.3.2.

Given a particular event ω_ν with a sending actor a_{i_ν}, a graph state $X_{\tilde{\nu}}$ and a parameter vector β, the multinomial decision probability can be phrased as

$$p(j; i_\nu, X_{\tilde{\nu}}, \beta) = \frac{\exp\left(\beta^T s(i_\nu, j, X_{\tilde{\nu}})\right)}{\sum_{k \in R_\nu} exp\left(\beta^T s(i_\nu, k, X_{\tilde{\nu}})\right)} \tag{3.19}$$

giving the probability that a_j is chosen as the receiver. The receiver index j is in R_ν, the set of indexes of potential receivers of event ω_ν as defined in equation 3.5. The probability $p(j_\nu; i_\nu, X_{\tilde{\nu}}, \beta)$ of the observed receiver choice a_{j_ν} in event ω_ν can be understood as an event likelihood, given that the time stamp t_ν and the sender a_{i_ν} have already been determined.

3.2.3 Network Statistics

The choice of the event receiver is based on graph structures. For each possible event receiver a_k with $k \in R_\nu$ of an event ω_ν, statistics are measured in the sub-graphs that surround sender a_{i_ν} and a_k. The resulting vector of statistics is evaluated in a multinomial choice model as explained above. Each statistic measures a structure that expresses, for example, the tendency to communicate with the same actors repeatedly or to reciprocate previously incoming communication events. The set of these structures is named Q.

For each event ω_ν exists a set of statistic vectors in which each element corresponds to one receiver choice.

$$S_\nu = \{s(i_\nu, 1, X_{\tilde{\nu}}), \ldots, s(i_\nu, j_\nu, X_{\tilde{\nu}}), \ldots, s(i_\nu, |A|, X_{\tilde{\nu}})\} \tag{3.20}$$

whereas $s(i_\nu, i_\nu, X_{\tilde{\nu}})$ is not evaluated as it does not represent a valid receiver choice. Each statistics vector $s(i_\nu, k, X_{\tilde{\nu}})$ in S_ν has $|Q|$ entries. The q-th element of the statistics vector $s(i_\nu, k, X_{\tilde{\nu}})$ is denoted by

$$s_q(i_\nu, k, X_{\tilde{\nu}}) \tag{3.21}$$

with q in $\{1, \ldots, |Q|\}$.

In the following sections we may also use the following abbreviations:

$$s_{\nu;\cdot k} = s(i_\nu, k, X_{\tilde{\nu}}) \tag{3.22}$$

$$s_{\nu;qk} = s_q(i_\nu, k, X_{\tilde{\nu}}) \tag{3.23}$$

The vector of statistics of a receiver choice a_k of event ω_ν is denoted by $s_{\nu;\cdot k}$, the q-th element of this vector by $s_{\nu;qk}$.

3.2.4 Markov Process

After having sketched the idea of the two-level decision process and having briefly explained the two probability functions of actor activity and multinomial receiver choices, we now integrate both parts into the framework of a Markov process (or continuous-time Markov chain). It models the occurrence of events with each possible sender-receiver combination (i, j) at arbitrary points in time t. The formulation is similar to the Markov process of stochastic actor-oriented models for network panel data introduced by Snijders (2005).

Let $\{X(t)|t \geq 0\}$ with state space $\mathscr{X}$ be a Markov process with right-continuous realizations. The state space $\mathscr{X}$ is defined by all possible states of the graph X. The Markov process describes updates of ties in X due to the occurrence of events in the event stream. Each event is understood as a change of the process state. Formally, it holds that the state space is

$$\mathscr{X} = \left\{X \in (\mathbb{R}_0^+)^{|A| \times |A|}\right\}. \tag{3.24}$$

A Markov process is a process "without memory", which means that all relevant information for the next process change is represented by the current state. Therefore, the emergence of new ties in the event stream at time t is assumed to depend only on current network structures in the communication graph $X(t)$. In the basic model that we present here, it does not matter through which sequence of events the graphs actually evolved. However, the state space can be extended to model this event history. This idea will be discussed in chapter 4.

For two subsequent private message events

$$\omega_\nu = (i_\nu, j_\nu, t_\nu),$$
$$\omega_{\nu+1} = (i_{\nu+1}, j_{\nu+1}, t_{\nu+1})$$

the Markov property holds:

$$\begin{aligned} &P\big(X(t_{\nu+1}) = x_{\nu+1} | X(t_\nu) = x_\nu, X(t_{\nu-1}) = x_{\nu-1}, \dots, X(t_0) = x_0\big) \\ =&P\big(X(t_{\nu+1}) = x_{\nu+1} | X(t_\nu) = x_\nu\big) \end{aligned} \tag{3.25}$$

For each message event from a sender $a_i \in A$ to a receiver $a_j \in A$ a "tendency" for its occurrence is defined as a Poisson process with a rate λ_{ij} (explained below). This definition is similar to the statement that time between two consecutive messages from a_i to a_j is λ_{ij}-exponentially distributed.

Rates vary across sending and receiving actors. λ_{ij} can be understood as the propensity of actor a_i to write a private message to a_j. It depends, first, on the general activity of a_i and, second, on the local graph structures that a_i, a_j and all other potential event receivers are embedded in. Based on these local structures, a_i is assumed to make a choice with regard to the receiver. We understand that both decision levels of the Markov process are driven by individual choices of the sending actor.

The event rate λ_{ij} (the transition rate of the Markov process) can then be defined as a combination of the individual activity rates (denoted by ρ_i, see equation 3.15) and the probability of actor a_i to choose a_j as event receiver over any other actor. This probability is defined by $p(j;i,X(t),\beta)$ and was introduced in equation 3.19.

$$\lambda_{ij}(X(t);\rho_i,\beta) \approx \rho_i p(j;i,X(t),\beta) \tag{3.26}$$

The transition rate is only defined as an approximation, because the process state is only stable for short time spans, as it changes with the external decay of ties. However, the effect of the decay process is the smaller, the more events per time are observed. For event streams with many active actors, the decay effect on the transition rates is rather irrelevant.

We defined the Poisson rate λ_{ij} by multiplying the Poisson rate ρ_i with a discrete probability distribution $p(j;i,X(t),\beta)$. Thereby, we implicitly assume independence between both decisions (see Waldmann and Stocker (2004, p.159)). This may be reasonable for the basic model we present here. However, this assumptions gets more critical, as soon as the individual activity rate is further parameterized or additional actor decision levels are included. We discuss these issues together with other extended specifications in chapter 4.

Applications of such extended specifications of this basic Markov process are presented in chapters 5 and 6.

The following section focuses on the receiver choice decision of the actor-oriented two-level decision process. The multinomial probability $p(j;i,X,\beta)$ defines the choice between potential event receivers.

3.3 Receiver Choice Sub-Model

3.3.1 Multinomial Choices of Event Receivers

In section 2.1, the origins of multinomial choice models in econometrics were reviewed. As proposed by Snijders (2005), we apply this framework on choices in social networks. We understand an event sender as a rational actor who maximizes his or her individual "communication utility" function when choosing event receivers. Maximizing an individual utility is partly understood as minimizing communication costs. Communicating with others is costly as it requires investment of time, social effort, and sometimes even monetary communication costs, like in case of cell phone communication. The initial creation of new ties is assumed to be costly because it is easier to maintain existing ties. Therefore, we expect a rational actor to choose event receivers not randomly, but to direct communication investments strategically. Benefits of communicating with others include the availability of social capital, status acquisition, information exchange, retrieval of advice and feelings of embeddedness, which all potentially promote subjective well-being. A detailed overview of social-psychological theories of communication networks was written by Monge and Contractor (2003). Examples are

- Theory of Social Capital (Monge and Contractor, 2003, p.142ff)
- Transaction Cost Economic Theory (Monge and Contractor, 2003, p.150ff)
- Collective Action Theory (Monge and Contractor, 2003, p.168ff)
- Contagion Theories (Monge and Contractor, 2003, p.173ff)
- Social Support Theories (Monge and Contractor, 2003, p.235ff)
- Evolutionary Theories (Monge and Contractor, 2003, p.241ff)

It is hard to separate out specific motivations of communication decisions, but the assumption that there is rational optimization, allows us to apply multinomial choice models.

Network structures are a good predictor for communication choices in networks. In communication data sets, we would usually expect to observe repeated communication with the same receivers, because the creation and maintenance of communication relations is costly. Repeated communication can be easily measured as a dyadic structure in the communication graph. This is exemplarily shown in section 3.3.2.

Structures can also incorporate third (and more) actors. For example, it can be tested whether communication within dense groups is more likely than between ”isolated“ dyads.

This observation would usually be made in multinomial communication choice models. Network structures can incorporate attributes (e.g., to test homophily effects), or the state of other networks. In the following, we present a simple dyadic model. Other specifications will be discussed in section 4.4 (p.71ff), and will be applied on real data streams in chapters 5 and 6.

3.3.2 A Dyadic Choice Example

Let us consider a concrete example using the last two events ω_8 and ω_9 of the phone call event stream in table 3.1. We might be interested in whether the actor choices are influenced by existing communication ties between event sender and receiver. Two structures are considered as independent variables of this choice:

- Actors may tend to call others whom they have called before (Repeated communication).
- Actors may tend to call others who have called them before (Reciprocity).

The set Q of included effects is then defined as

$$Q = \{\mathit{Repeated\ communication}, \mathit{Reciprocity}\} \tag{3.27}$$

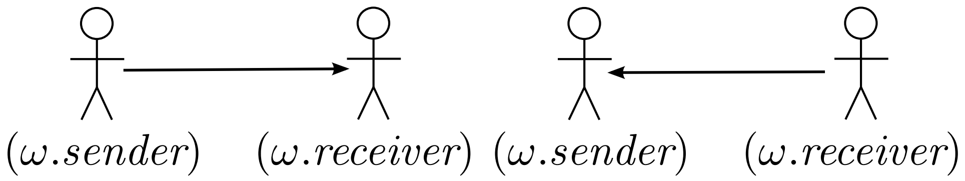
and the independent variables are measured with the following statistics:

$$s_1(i,k,X_{\tilde{v}}) = \begin{cases} 1, \text{ if } x_{\tilde{v};ik} > 0 \\ 0, \text{ else} \end{cases} \tag{3.28}$$

measures *Repeated communication* and

$$s_1(i,k,X_{\tilde{v}}) = \begin{cases} 1, \text{ if } x_{\tilde{v};ki} > 0 \\ 0, \text{ else} \end{cases} \tag{3.29}$$

measures *Reciprocity*. The two structures are shown in figure 3.3.



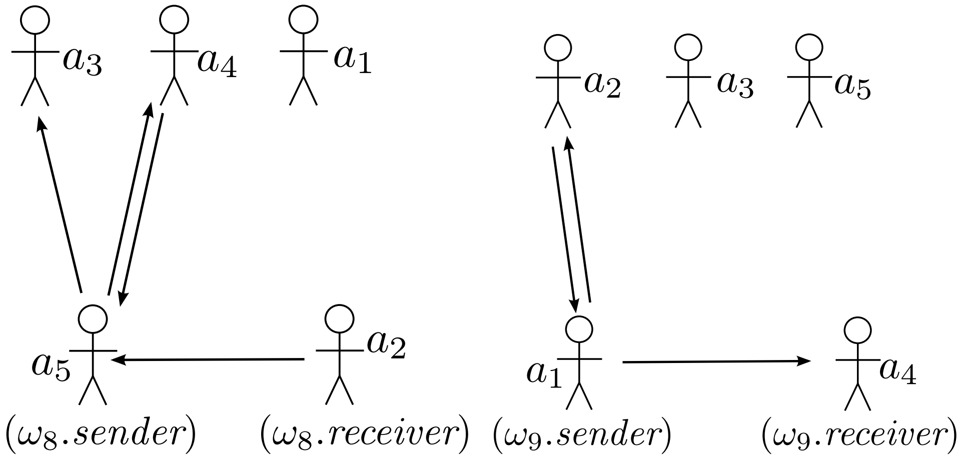
(a) Repeated communication

(b) Reciprocity

Figure 3.3: Two dyadic structures influencing communication choices (dyadic choice example).

Figure 3.4 shows the local choice environments (that are measured with the dyadic structures between event sender and potential receivers) of the two events ω_8 and ω_9. The senders are actors a_5 (in event ω_8) and a_1 (in event ω_9). Tie values are not depicted, as they are irrelevant for the proposed statistics. The figures can be understood as follows: In both sub-figures 3.4(a) and 3.4(b) the event sender is shown on the lower left, the event receiver on the lower right. All other potential receivers are depicted on the top. Arrows indicate previously created ties in the communication graph. The figures show only those ties in the communication graph to which the event sender is connected as only those ties are evaluated by the dyadic choice example. The complete graphs can be found in figure 3.1.

For example, event ω_8 is shown in sub-figure 3.4(a). Here, the sender is actor a_5, the observed receiver is actor a_2. Actors a_1, a_3 and a_4 are potential receivers that were not chosen as receivers by the event sender.



(a) Receiver Choice of ω_8

(b) Receiver Choice of ω_9

Figure 3.4: Events ω_8 and ω_9 from table 3.1 on page 29.

The statistics from equations 3.28 and 3.29 are calculated for the structures of figure 3.4, and according to the abbreviated notation in equation 3.22. The statistics vectors of the receiver choice of event ω_8 are

$$s_{8;.2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, s_{8;.1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, s_{8;.3} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, s_{8;.4} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.30}$$

and of event ω_9 are

$$s_{9;.4} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, s_{9;.2} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, s_{9;.3} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, s_{9;.5} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.31}$$

The first digit in the index denotes the event index; the last digit indicates the receiver index. The vector of the chosen receiver is shown left. In the following sections, we use this concrete dyadic model to explore the likelihood and the estimation process of receiver choice sub models in general.

3.3.3 Likelihood

For a single event observation, the receiver choice probability from equation 3.19 can – for the choice of the actually observed receiver a_{j_v} – be understood as the receiver choice likelihood:

$$L(\omega_v.j;\beta) = p_{j_v}(i_v, S_v, \beta). \tag{3.32}$$

Within an event stream, all receiver choices are assumed to be conditionally independent given the state of the process, namely the communication graph X. Therefore, the event stream likelihood can be expressed as the product of the event likelihoods:

$$\begin{aligned} L(\Omega;\beta) &= \prod_{v=1}^{|\Omega|} p_{j_v}(i_v, S_v, \beta) \\ &= \prod_{v=1}^{|\Omega|} \frac{\exp\left(\beta^T s(i_v, j_v, X_{\tilde{v}})\right)}{\sum_{k\in R_v} exp\left(\beta^T s(i_v, k, X_{\tilde{v}})\right)} \end{aligned} \tag{3.33}$$

Note that this likelihood is only related to receiver choices. For reasons of simplicity, we denote it by $L(\Omega;\beta)$, although the complete likelihood of Ω also includes the determination of event time and event senders. The determination of time and the event sender are modeled separately by the actor activity rates ρ_i, which were introduced as the first decision level of the Markov process.

The parameter vector β is unknown. It is estimated, so that it maximizes the event stream likelihood. To apply a maximum likelihood (ML) estimation efficiently, certain properties of the event stream likelihood have to be shown. First, the ML function should be concave. It will be shown that this feature holds for the logarithmized likelihood function. Therefore, the log-likelihood function is maximized:

$$\begin{aligned} \log L(\Omega;\beta) &= \sum_{v=1}^{|\Omega|} \log p_{j_v}(i_v, S_v, \beta) \\ &= \sum_{v=1}^{|\Omega|} \left(\beta^T s(i_v, j_v, X_{\tilde{v}}) - \log \sum_{k\in R_v} exp\left(\beta^T s(i_v, k, X_{\tilde{v}})\right)\right) \end{aligned} \tag{3.34}$$

Both receiver choices of the example in section 3.3.2 can be formulated that way. We now continue with this example based on the event stream in table 3.1 on page 29. Let $\Omega' = \{\omega_8, \omega_9\}$ be an event stream that is defined on the initial graph $X_{\tilde{8}}$. Next, the event stream likelihood can be described by the likelihood

$$\begin{aligned} L(\Omega',\beta) =& \frac{\exp(\beta_2)}{\exp(\beta_1+\beta_2)+\exp(\beta_1)+\exp(\beta_2)+1} \\ &\cdot \frac{\exp(\beta_1)}{\exp(\beta_1+\beta_2)+\exp(\beta_1)+2} \end{aligned} \tag{3.35}$$

and the log-likelihood is defined by

$$\begin{aligned} \log L(\Omega',\beta) =& \beta_2 - \log\left(\exp(\beta_1+\beta_2)+\exp(\beta_1)+\exp(\beta_2)+1\right) \\ &+\beta_1 - \log\left(\exp(\beta_1+\beta_2)+\exp(\beta_1)+2\right). \end{aligned} \tag{3.36}$$

Plots of the likelihood function and the log-likelihood function over a discrete grid are shown in figures 3.5 and 3.6. The x- and y-axis represent the two parameters β_1 (weighting *Repeated communication*) and β_2 (weighting *Reciprocity*). The z-axis gives the likelihood and the log-likelihood, respectively. Both functions have a clear maximum, hence there is one parameter combination that best explains the observed decisions in the two events (the maximum likelihood estimate). The likelihood function clearly is not concave.

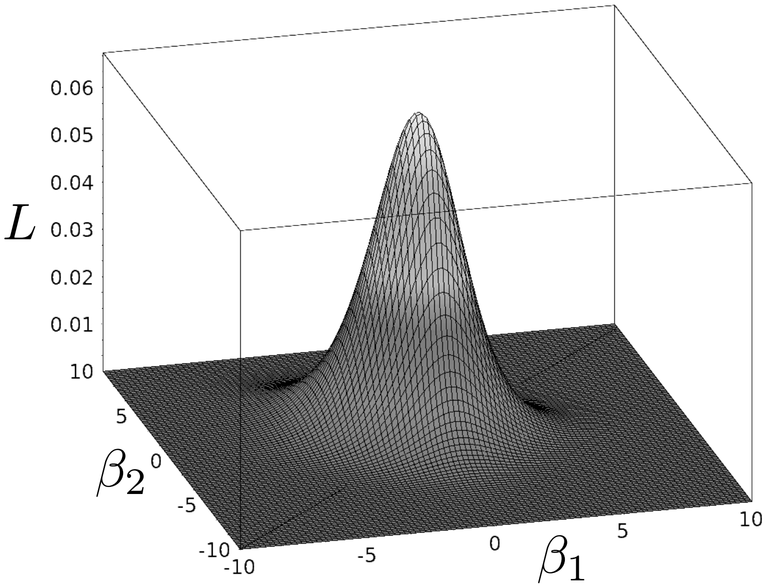


Figure 3.5: Likelihood L of the dyadic example from equation 3.35

The vector of the maximum likelihood parameters is denoted by $\hat{\beta}$. The interpretation of parameters is discussed in section 3.3.6. In section 3.3.4 we discuss the fact that the log-likelihood function is a concave function and even strictly concave in most cases. Therefore, a Newton Raphson algorithm can be applied to retrieve the maximum likelihood estimates – this algorithm is discussed in section 3.3.5. Both the concavity proof and the optimization algorithm use the *Jacobian vector* $J(\Omega;\beta)$, the vector of first derivatives of the log-likelihood with respect to β_1 to $\beta_{|Q|}$, and the *Hessian matrix* $H(\Omega;\beta)$, the matrix of second derivatives with respect to two arbitrary elements of β, β_q and β_r.

In the following, we again use the abbreviated form of the statistics vector from equation 3.23 and denote its entries according to equation 3.23. The q-th element of the Jacobian vector J is defined as

$$\begin{aligned} J_q(\Omega;\beta) &= \sum_{v=1}^{|\Omega|} \frac{\partial \log p_{j_v}(i_v, S_v, \beta)}{\partial \beta_q} \\ &= \sum_{v=1}^{|\Omega|} \left[s_{v;qj_v} - \frac{\sum_{k\in R_v} s_{v;qk} \exp\left(\beta^T s_{v;.k}\right)}{\sum_{k\in R_v} \exp\left(\beta^T s_{v;.k}\right)} \right] \end{aligned} \tag{3.37}$$

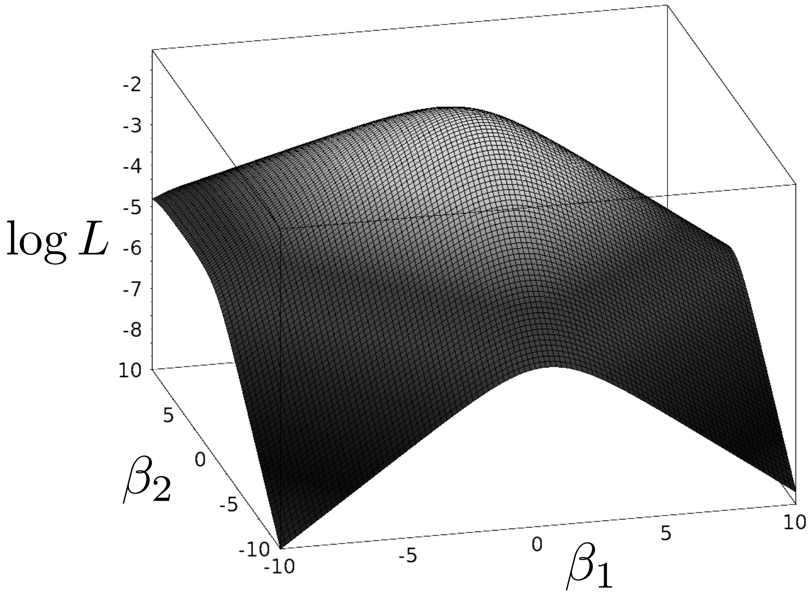


Figure 3.6: Log likelihood $\log L$ of the dyadic example from equation 3.36

Elements $h_{qr}(\Omega;\beta)$ of the Hessian are defined as

$$\begin{aligned} h_{qr}(\Omega;\beta) &= \sum_{v=1}^{|\Omega|} \frac{\partial \log p_{j_v}(i_v, S_v, \beta)}{\partial \beta_q \partial \beta_r} \\ &= \sum_{v=1}^{|\Omega|} \left[\frac{\Big(\sum_{k\in R_v} s_{v;qk} \exp\left(\beta^T s_{v;.k}\right)\Big)\Big(\sum_{k\in R_v} s_{v;rk} \exp\left(\beta^T s_{v;.k})\right)\Big)}{\Big(\sum_{k\in R_v} \exp\left(\beta^T s_{v;.k}\right)\Big)^2} \right. \\ &\qquad \left. - \frac{\Big(\sum_{k\in R_v} s_{v;qk} s_{v;rk} \exp\left(\beta^T s_{v;.k}\right)\Big)\Big(\sum_{k\in R_v} \exp\left(\beta^T s_{v;.k}\right)\Big)}{\Big(\sum_{k\in R_v} \exp\left(\beta^T s_{v;.k}\right)\Big)^2} \right] \end{aligned} \tag{3.38}$$

which can be further simplified to

$$\begin{aligned}
&= \sum_{\nu=1}^{|\Omega|} \left[\frac{\sum_{k\in R_\nu} \sum_{l\in R_\nu} \left(s_{\nu;qk} s_{\nu;rl} \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right. \\
&\quad \left. - \frac{\sum_{k\in R_\nu} \sum_{l\in R_\nu} \left(s_{\nu;qk} s_{\nu;rk} \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right] \\
&= \sum_{\nu=1}^{|\Omega|} \left[\frac{\sum_{k\in R_\nu} \sum_{l\in R_\nu} \left(\left(s_{\nu;qk} s_{\nu;rl} - s_{\nu;qk} s_{\nu;rk}\right) \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right] \\
&= \frac{1}{2} \sum_{\nu=1}^{|\Omega|} \left[\frac{\sum\limits_{k\in R_\nu} \sum\limits_{l\in R_\nu} \left(\left(s_{\nu;qk} s_{\nu;rl} - s_{\nu;qk} s_{\nu;rk} + s_{\nu;qk} s_{\nu;rl} - s_{\nu;qk} s_{\nu;rk}\right) \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right].
\end{aligned} \tag{3.39}$$

A re-ordering of summands finally leads to a shorter form:

$$\begin{aligned}
&= \frac{1}{2} \sum_{\nu=1}^{|\Omega|} \left[\frac{\sum\limits_{k\in R_\nu} \sum\limits_{l\in R_\nu} \left(\left(s_{\nu;qk} s_{\nu;rl} - s_{\nu;qk} s_{\nu;rk} + s_{\nu;ql} s_{\nu;rk} - s_{\nu;ql} s_{\nu;rl}\right) \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right] \\
&= -\frac{1}{2} \sum_{\nu=1}^{|\Omega|} \left[\frac{\sum\limits_{k\in R_\nu} \sum\limits_{l\in R_\nu} \left(\left(s_{\nu;qk} - s_{\nu;ql}\right) \left(s_{\nu;rk} - s_{\nu;ql}\right) \exp\left(\beta^T (s_{\nu;.k} + s_{\nu;.l})\right)\right)}{\left(\sum_{k\in R_\nu} \exp\left(\beta^T s_{\nu;.k}\right)\right)^2} \right]
\end{aligned} \tag{3.40}$$

3.3.4 Concavity of the Log Likelihood

There are many possible proofs to show the concavity of the log-likelihood function 3.34. McFadden argues that the concavity follows from the fact that the Hessian matrix is a "negative of a weighted moment matrix of the independent variables" (see McFadden (1974, p.115)). Under the assumption that an observed event stream is not a sample but a rather complete observation, the concavity of the event stream log-likelihood follows from the fact that the Hessian (see equation 3.40) is negative semi-definite (Bronstein et al., 2001, p.291). If so, for each real vector z with $|z| = |\beta|$ and where z is not the null vector, holds

$$0 \geq z^T H(\Omega;\beta). \tag{3.41}$$

which equals

$$
\begin{aligned}
&=\sum_{q=1}^{|Q|}\sum_{r=1}^{|Q|} h_{qr}(\Omega;\beta) z_q z_r \\
&=-\frac{1}{2}\sum_{q=1}^{|Q|}\sum_{r=1}^{|Q|}\sum_{v=1}^{|\Omega|}\left[\frac{\sum_{k\in R_v}\sum_{l\in R_v}\left(\left(s_{v;qk}-s_{v;ql}\right)\left(s_{v;rk}-s_{v;ql}\right)\exp\left(\beta^T\left(s_{v;.k}+s_{v;.l}\right)\right)\right)}{\left(\sum_{k\in R_v}\exp\left(\beta^T s_{v;.k}\right)\right)^2}\right] z_q z_r \\
&=-\frac{1}{2}\sum_{v=1}^{|\Omega|}\sum_{q=1}^{|Q|}\sum_{r=1}^{|Q|}\left[\frac{\sum_{k\in R_v}\sum_{l\in R_v}\left(\left(s_{v;qk}-s_{v;ql}\right)\left(s_{v;rk}-s_{v;ql}\right)\exp\left(\beta^T\left(s_{v;.k}+s_{v;.l}\right)\right)\right)}{\left(\sum_{k\in R_v}\exp\left(\beta^T s_{v;.k}\right)\right)^2}\right] z_q z_r \\
&=-\frac{1}{2}\sum_{v=1}^{|\Omega|}\left[\frac{\sum_{k\in R_v}\sum_{l\in R_v}\exp\left(\beta^T\left(s_{v;.k}+s_{v;.l}\right)\right)\sum_{q=1}^{|Q|}\sum_{r=1}^{|Q|}\left(\left(s_{v;qk}-s_{v;ql}\right)\left(s_{v;rk}-s_{v;ql}\right)z_q z_r\right)}{\left(\sum_{k\in R_v}\exp\left(\beta^T s_{v;.k}\right)\right)^2}\right] \\
&=-\frac{1}{2}\sum_{v=1}^{|\Omega|}\left[\frac{\sum_{k\in R_v}\sum_{l\in R_v}\exp\left(\beta^T\left(s_{v;.k}+s_{v;.l}\right)\right)\sum_{q=1}^{|Q|}\left(s_{v;qk}-s_{v;ql}\right)z_q\sum_{r=1}^{|Q|}\left(s_{v;rk}-s_{v;ql}\right)z_r}{\left(\sum_{k\in R_v}\exp\left(\beta^T s_{v;.k}\right)\right)^2}\right] \\
&=-\frac{1}{2}\sum_{v=1}^{|\Omega|}\left[\frac{\sum_{k\in R_v}\sum_{l\in R_v}\overbrace{\left(\exp\left(\beta^T\left(s_{v;.k}+s_{v;.l}\right)\right)\right)}^{>0}\overbrace{\left(\sum_{q=1}^{|Q|}\left(s_{v;qk}-s_{v;ql}\right)z_q\right)^2}^{\geq 0}}{\underbrace{\left(\sum_{k\in R_v}\exp\left(\beta^T s_{v;.k}\right)\right)^2}_{>0}}\right] \qquad (3.42)
\end{aligned}
$$

The last equation is less than or equal to zero. Therefore, condition 3.41 is fulfilled. The Hessian is negative semi-definite, which proves that the log-likelihood function is concave. If the quadratic form in condition 3.41 is smaller than zero for any vector z, the log-likelihood function is strictly concave and has a unique maximum (as long as a real maximum exists). McFadden (1974, p.116) discusses this in more detail. However, when observing long event streams this distinction is not very relevant: If only one of the summands in the log-likelihood function (denoting the likelihood of a single receiver choice) is strictly concave, then the whole log-likelihood is strictly concave.

In the following section, the concavity feature is used to apply a Newton Raphson maximization algorithm.

3.3.5 Estimation Procedure

The maximum of a concave and in many cases strictly concave log-likelihood function can be estimated straightforwardly with a Newton Raphson method (Deuflhard, 2004). The receiver choice log-likelihood function is concave. We present the optimization method in section 3.3.5. In section 3.3.5 statistical tests are investigated. Section 3.3.5 discusses the computational complexity of the estimation.

Newton Raphson Estimation

The maximum of a concave function, like the log-likelihood in equation 3.34, can be estimated using Newton methods (Deuflhard, 2004, p.7ff). Generally, these methods allow finding the null of a function iteratively by assuming its linearity in each iteration step. Finding the optimum of a strictly concave function is similar to finding the null of the first derivative. In case of a function $f(x)$ with one parameter, the idea is as follows: To find the null $\hat{x}$, $f(x)$ at a certain point x^0 can be approximated by the first two terms of its Taylor expansion:

$$0 = f(\hat{x}) = f(x^0 + \Delta x) \approx f(x^0) + f'(x^0)(\hat{x} - x^0) \tag{3.43}$$

which is equivalent to

$$\hat{x} - x^0 \approx \Delta x^0 = -\frac{f(x^0)}{f'(x^0)}. \tag{3.44}$$

If the function is linear, $\hat{x}$ equals $x^0 - \frac{f(x^0)}{f'(x^0)}$. If the function is not linear, further iteration steps are needed to approach the null. Thus, the change rule for parameter x^k after k iteration steps can be generalized to

$$\Delta x^{k+1} = -\frac{f(x^k)}{f'(x^k)}. \tag{3.45}$$

In case of a multidimensional function $\log L(\Omega;\beta)$ (Deuflhard, 2004, pp.16,22), the null has to be found for the Jacobian $J(\Omega;\beta)$ (the vector of first derivatives shown in equation 3.37) with the equivalent of f' being the Hessian matrix $H(\Omega;\beta)$. $H(\Omega;\beta)$ is shown in equation 3.40. Generally, $\Delta\beta^{k+1}$ for the parameter adaption of β^k in the $(k+1)$-th iteration step is defined as

$$\Delta\beta^{k+1} = -H(\Omega;\beta^k)^{-1} J(\Omega;\beta^k) \tag{3.46}$$

To compute the next iteration, the inverse $H(\Omega;\beta^k)^{-1}$ of the Hessian $H(\Omega;\beta^k)$ has to be found. Each (negative) definite matrix has an inverse.

A simple stop criterion in the iterative optimization process can be that all derivatives in the Jacobian have to be below a certain threshold. However, this may lead to imprecise results if the maximum likelihood estimator lies in an area with very flat slopes. To avoid this, the algorithm can be started from different (randomized) starting points β^s (seeds). This way the optimum is approached from different directions. The new stop criterion could be defined as follows: Stop the optimization process, if all entries in the Jacobian are below a certain threshold *and* the distances between the results obtained from different seeds are below another threshold.

It may happen that in early iteration steps parameter vectors β^k have to be evaluated that include high absolute values. The precision of the used data type (for example: *double*) may then not be sufficient to represent the exponential transformation of $\beta^T s(i_v, k, X_{\tilde{v}})$ in the log-likelihood function (see equation 3.34). A solution is to reduce the changes of the first iteration steps by including a damping factor α with

$$\Delta'\beta^{k+1} = \alpha\Delta\beta^{k+1}, \alpha \in (0,1). \tag{3.47}$$

The damping can be reduced in subsequent iteration steps with $\alpha \to 1$.

Statistical Tests

In most cases, the log-likelihood function has a global optimum that can be found in a straightforward manner with the Newton algorithm. However, it is not yet clear whether the parameter estimates at this optimal point are statistically significant. For example, the model with two events from section 3.3.2 allows the estimation of a global optimizing parameter $\hat{\beta}$, but we should be careful with making substantive inference from the two observed choices. Assume that there are many different, contradictory actor strategies that cannot be disentangled due to a poor model specification. In this case, $\hat{\beta}$, the best estimator found, might lie somewhere on a "surface" with a small negative slope in each direction. A large standard error would indicate a lack of statistical significance. On a more formal level, a t-test can help to decide whether the found optimum is significant.

The negative of the Hessian $-H(\Omega;\hat{\beta})$ of the maximum likelihood solution can be understood as the observed Fisher information matrix $F(\hat{\beta})$ (see Efron and Hinkley (1978, p.458) and Young and Smith (2005, p.123)):

$$F(\hat{\beta}) = -H(\Omega;\hat{\beta}) \tag{3.48}$$

Based on the Fisher information matrix, a covariance matrix $G(\hat{\beta})$ can be approximated by calculating the inverse of the observed Fisher information matrix:

$$G(\hat{\beta}) = F^{-1}(\hat{\beta}), \tag{3.49}$$

The standard deviation of a parameter estimate is also known as a parameter's standard error $\text{se}(\hat{\beta})$. It can be derived from the estimated covariance matrix as

$$\text{se}(\hat{\beta}_k) = \sqrt{G(\hat{\beta})_{kk}} = \sqrt{\big(-H(\Omega;\hat{\beta})\big)^{-1}_{kk}} \tag{3.50}$$

Using the parameter estimate and its corresponding standard error, a t-ratio

$$t_k = \frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)} \tag{3.51}$$

can be calculated in a statistical test. Assuming that parameters in given a sample of stochastic decisions are normally distributed, the t-ratios indicate the probability of a wrong conclusion regarding the sign of these parameters. Thresholds of the t-ratio can be defined for different levels of significance. Significance tests of this type are used to evaluate the results of the application chapters 5 and 6.

Standard errors can alternatively be calculated by means of bootstrap methods (Efron and Tibshirani, 1986). This approach is computationally harder. In large samples the results of both approaches – direct standard error calculation based on the estimated covariance matrix and bootstrapping – are similar. In the first application (in chapter 5) we apply a bootstrap method to estimate standard errors. In the second application in chapter 6 we derive the standard errors directly from the estimated covariance matrix, as in equation 3.50.

Computational Complexity

The computational complexity of the parameter estimation and of the statistical tests depends on several parameters:

$$n := |A| - 1$$

is the size of the set of possible receivers per event. Furthermore, the set of tested statistics

$$m := |Q|$$

and the length of the event stream

$$o := |\Omega|$$

are important. For the identification of the preprocessing complexity we also assume a maximum number of neighbors per actor. This constant is denoted by

$$\overline{k}.$$

The parameter $\overline{k}$ is independent from the size of the network as the amount of communication partners is naturally limited, for example due to cognitive limitations. Research has shoen that human beings have a natural limit with respect to the number of social contacts they can handle (see, e.g., Dunbar (1992)). In the following, we use the O-notation for denoting computational complexity (Goos, 2000, p.313).

The complexity of one iteration step of the Newton Raphson algorithm is in the complexity class of

$$O(om^3 n). \tag{3.52}$$

For each entry of the Hessian (m^2) each event's derivatives (o) have to be summed. An event derivative includes the sum over all statistics for each possible receiver (mn). The subsequent calculation of the inverses $H(\Omega,\beta^k)^{-1}$ and $(-H(\Omega,\beta^k))^{-1}$ has a complexity of less than m^3, so that does not increase the total complexity. The number of iteration steps depends on the similarity of the log-likelihood function to a quadratic function and the desired precision of the results. Our tests with large data sets show that the speed of convergence is robust to an increasing number of parameters, and the necessary time for estimation increases sub-linearly.

Before the estimation process can be started, the statistics vectors need to be preprocessed. This is done only once per Newton Raphson optimization. The preprocessing has a complexity of

$$O(omn\overline{k}) = O(omn). \tag{3.53}$$

For each event (o) and each potential receiver (n) m statistics have to be calculated. Usually, each statistic is defined as a count process. For example, the researcher could be interested in the number of circle structures in which both sender and receivers are embedded. The count of these structures is relatively simple if every node (i.e., actor) in the graph has a list of neighbors attached and can answer the request whether or not he is neighbor of a specific node in $O(1)$. The neighbor lists can be updated whenever a graph update is made, for example, due to occurring events or decay of ties. When structures with only three nodes are counted, $\overline{k}$ additions have to be made. Imagine the circle-structure example: All $\overline{k}$ neighbors of a receiver are considered potential third actors. For each neighbor of the receiver, it is checked whether this node is also a neighbor of the sending actor ($= O(1)$), i.e., whether it is a mutual neighbor of sender and receiver. As we assume a maximum number of neighbors per node, the constant $\overline{k}$ can be removed from the complexity notation. In big networks holds that $\overline{k} << n$.

3.3.6 Interpretation of Estimates

Interpreting maximum likelihood estimates is rather complex. Many factors influence these estimates: the distributions of local environments in which structures are measured, the number of observed structures, the existence of sub- or super-structures in the same model, and the size of the set of potential receivers. In this section, we calculate the log-likelihood for two basic models and event streams analytically.

The first model only includes the statistic *Repeated communication* from equation 3.28 (p.39). Similar to equation 3.27 the set of included effects is $Q = \{Repeated\ communication\}$. *Repeated communication* expresses whether a directed tie from sender to receiver exists. In this case, the statistic returns value 1, else it returns 0. Therefore, in each event ω_v the sender a_{i_v} can either choose a receiver a_{j_v} he is already connected to by an outgoing tie ($a_{j_v} \in A_v^+$), or a receiver he is not yet connected to ($a_{j_v} \in A_v^-$). Figure 3.7 shows this dyadic, binary choice environment. Ties that are not measured by the statistic are left out. For simplification, we set the following variables:

- n_v^+ is the size of set A_v^+.
- n_v^- is the size of set A_v^-.
- m^+ is the number of events ω_v in Ω where a receiver a_{j_v} in A_v^+ is chosen ($m^+ = \sum_{v=1}^{|\Omega|} s_{v;1j_v}$).
- m^- is the number of events where a receiver a_{j_v} in A_v^- is chosen: $m^- = |\Omega| - m^+$.
- β_1 is the weighting parameter of the dyadic statistic.

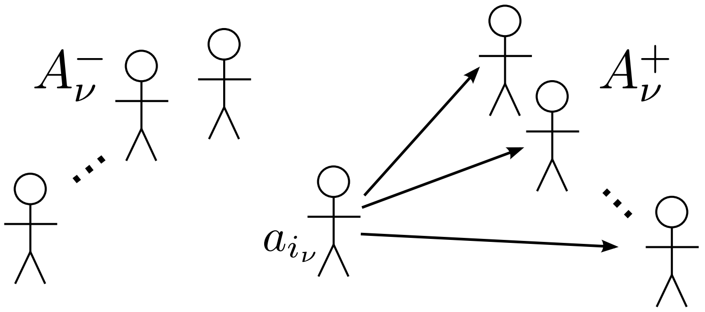


Figure 3.7: Receiver choice of an event ω_ν in an environment with one measured dyadic structure that either exists (receiver in A^+_ν) or not (receiver in A^-_ν).

In case of this dyadic, binary, one-parameter model the log-likelihood from equation 3.34 (p.41) can be simplified to the following form:

$$\log L(\Omega,\beta) = \beta_1 m^+ - \sum_{\nu=1}^{|\Omega|} \log\left(n^+_\nu exp(\beta_1) + n^-_\nu\right). \tag{3.54}$$

The simplified first derivative can be set to zero to calculate an optimal solution analytically.

$$J_1(\Omega;\beta) = \frac{\partial \log L(\Omega,;\beta)}{\partial \beta_1} = m^+ - \sum_{\nu=1}^{|\Omega|} \frac{n^+_\nu exp(\beta_1)}{n^+_\nu exp(\beta_1) + n^-_\nu} \overset{!}{=} 0 \tag{3.55}$$

The solution of this model is still not straightforward. Therefore, we now further assume that the local environments of the sending actors are stable for each observed decision. This *constant neighborhood assumption* is a strong constraint based on the idea that local environments often follow a stationary distribution. Then, for all ω_ν it holds that n^+_ν equals the constant parameter n^+ (also, $n^-_\nu =: n^-$). An optimal solution $\hat{\beta}_1$ for this simplified case can be found by setting the first derivative to zero.

$$
\begin{aligned}
m^+ - |\Omega| \frac{n^+ exp\big(\hat{\beta}_1\big)}{n^+ exp\big(\hat{\beta}_1\big) + n^-} &= 0 \\
\frac{m^+}{|\Omega|} - \frac{n^+ exp\big(\hat{\beta}_1\big)}{n^+ exp\big(\hat{\beta}_1\big) + n^-} &= 0 \\
\frac{m^+}{|\Omega|}\big(n^+ exp(\hat{\beta}_1) + n^-\big) - n^+ exp\big(\hat{\beta}_1\big) &= 0 \\
\frac{m^+}{|\Omega|} n^+ exp(\hat{\beta}_1) + \frac{m^+}{|\Omega|} n^- - n^+ exp\big(\hat{\beta}_1\big) &= 0 \\
exp\big(\hat{\beta}_1\big)\left(\frac{m^+}{|\Omega|} n^+ - n^+\right) &= -\frac{m^+}{|\Omega|} n^- \\
exp\big(\hat{\beta}_1\big) &= \frac{-\frac{m^+}{|\Omega|} n^-}{n^+\big(\frac{m^+}{|\Omega|} - 1\big)} \\
exp\big(\hat{\beta}_1\big) &= \frac{\frac{m^+}{|\Omega|}}{1 - \frac{m^+}{|\Omega|}} \times \frac{n^-}{n^+} \\
\hat{\beta}_1 &= \log\left(\frac{m^+}{m^-} \times \frac{n^-}{n^+}\right) \qquad (3.56)
\end{aligned}
$$

Under the *constant neighborhood assumption* an estimate $\hat{\beta}_1$ can therefore be interpreted as the log of the ratio of "repeatedly used" (m^+) and newly established communication ties (m^-) times the ratio of not connected neighbors (n^-) and connected neighbors (n^+) in the constant environment. In case the ratio of repeatedly used ties equals the ratio of neighbors with an outgoing tie, $\hat{\beta}_1 = 0$. If more outgoing ties are chosen than predicted by the neighborhood ratio, the parameter estimate gets positive as the inner function of the log is bigger than 1. If less outgoing ties are chosen than predicted, the estimate becomes negative.

Let us consider a second example using the same model with one parameter *Repeated communication*. Now $A^+ \cup A^-$ (the neighborhood of the sender) is not assumed to be constant, but to vary across events. If only two events are observed and if only in one case an element in A^+ is chosen ($m^+ = m^- = 1$), then we can set the simplified first derivative of the log-likelihood in equation 3.55 to zero and solve the equation:

$$
\begin{aligned}
0 \overset{!}{=} & 1 - \sum_{v=1}^{2} \frac{n_v^+ exp\big(\hat{\beta}_1\big)}{n_v^+ exp\big(\hat{\beta}_1\big) + n_v^-} \\
= & \big(n_1^+ \exp(\hat{\beta}_1) + n_1^-\big)\big(n_2^+ \exp(\hat{\beta}_1) + n_2^-\big) \\
& - \big(n_1^+ \exp(\hat{\beta}_1)\big)\big(n_2^+ \exp(\hat{\beta}_1) + n_2^-\big) \\
& - \big(n_2^+ \exp(\hat{\beta}_1)\big)\big(n_1^+ \exp(\hat{\beta}_1) + n_1^-\big) \qquad (3.57)
\end{aligned}
$$

After expanding and reducing it, the form follows:

$$\begin{aligned} 0 =&(-n_1^+n_2^+)\exp(2\hat{\beta}_1)+n_1^-n_2^- \\ \Leftrightarrow \hat{\beta}_1 =&\frac{1}{2}\log\left(\frac{n_1^+n_2^+}{n_1^-n_2^-}\right) \end{aligned} \tag{3.58}$$

Once again, the maximum likelihood parameter can be expressed as a "ratio" between the number of actors connected with and without outgoing ties. As m^+ and m^- were set to 1, they are not part of the equation.

The general form for any m^+ of the first line in equation 3.58 is

$$\begin{aligned} 0 =&m^-(-n_1^+n_2^+)\exp(2\hat{\beta}) \\ &+(m^+-1)(n_1^+n_2^-+n_1^-n_2^+)\exp(\hat{\beta})+m^+n_1^-n_2^-. \end{aligned} \tag{3.59}$$

The analytical solutions we discussed here give an intuitive idea about how the absolute values of simple parameter estimates can be interpreted. However, the solutions in equations 3.56 and 3.58 are only applicable in very simple cases. The results of the exemplary dyadic model from 3.3.2 can in parts be interpreted this way. We discuss this in the next section.

However, in most cases, an analytical solution of the log-likelihood maximization cannot be found. Instead, the estimated parameters are interpreted based on their effect on the choice probabilities. Let the optimizing parameter of the structure *Repeated communication* be $\hat{\beta}_1$. Now imagine an event sender a_i choosing between two receivers a_k and a_l, where a_k is connected with the sender by a directed tie $a_i \rightarrow a_k$ ($a_k \in A^+$) and a_l is not ($a_l \in A^-$). Then the choice probabilities from equation 3.19 (p.35) have the following relation:

$$\begin{aligned} p(k;i,X,\hat{\beta}_1) =&\theta \cdot p(k;i,X,\hat{\beta}_1) \\ \Leftrightarrow \exp(\beta_1) =& \theta exp(0) = \theta. \end{aligned}$$

The probability of choosing a connected receiver is $\theta = \exp(\beta_1)$ times higher than choosing a disconnected receiver. The same interpretation can be applied to interpreting parameters of count statistics (e.g. β_q weighting the number of mutual contacts of sender and receiver). Statistics of this kind are introduced later. We can then infer that with each additional count (each additional mutual contact) the probability of a corresponding choice is increased by $exp(\hat{\beta}_q)$. For statistics that are not measured binary (either 1 or 0) generally holds that estimates are interpreted by the change of the choice probabilities with each "change of one unit". One unit could be one additional contact. Non-binary receiver choice statistics are discussed in section 4.4. Interpretations of non-binary statistics can be found in the application chapters in sections 5.5 and 6.5.

Interpretation of a density parameter of stochastic actor-oriented models for panel data, which is closest to the *Repeated communication* parameter in the proposed event models, is discussed in Snijders (2005, p.241ff).

	Repeated communiation		*Reciprocity*			
Model	$\hat{\beta}_1$	(s.e.)	$\hat{\beta}_2$	(s.e.)	L	$\log L$
M_0					0.0625	-2.7726
M_1	0.000	(1.414)			0.0625	-2.7726
M_2			0.549	(1.468)	0.0670	-2.7032
M_3	-0.175	(1.508)	0.609	(1.573)	0.0674	-2.6965

Table 3.2: Estimation results of the exemplary dyadic specification in section 3.3.2. Four models with different parameter sets are tested.

3.3.7 Results of the Dyadic Example

At the beginning of this chapter, an exemplary event stream was introduced (see table 3.1 on page 29). The last two events were defined as a separate event stream, and the dyadic parameters *Repeated communication* and *Reciprocity* were introduced to describe the receiver choices (section 3.3.2, p.39). Here, we present the maximum likelihood estimates that were derived using a Newton Raphson procedure, and discuss them based on the analytical observations from the previous section. The Newton Raphson estimation of the exemplary event stream with two events needs four iteration steps until the maximum absolute value of the derivatives is below 10^{-4}. As the log-likelihood function is very "flat" (standard errors are high), several additional iteration steps were applied that lead to the results in table 3.2.

Four different models were estimated as shown by four rows: First, an empty model without any parameters (this is similar to a random choice model), second and third, models with one parameter only, and fourth, the model with both parameters included. The models are named M_0 to M_3.

As can be seen in figures 3.5 and 3.6 (p.42), the maximum likelihood estimates are close to zero. This is what we also find in the results table 3.2. For explanation purposes we used a very simple model: The event stream incorporates only two choices with a set of four possible event receivers. This small sample is used to estimate two parameters. Such a small event stream cannot be used to answer general questions about choice behavior in dynamic environments. Due to the small sample size the estimates have a high standard error and are insignificant. But we can use this example to learn more about the interpretation of parameter values.

The signs of the estimates indicate (according to the discussion in section 3.3.6) whether a structure was chosen above, below or according to expectations: The structure *Repeated communication* was chosen as often as expected ($\hat{\beta}_1 = 0.000$ in model M_1) and structure *Reciprocity* was chosen above expectations ($\hat{\beta}_2 = 0.549$ in model M_2).

If we look at the local choice environments of the example (see figure 3.4, p.40), we can see that in two choices with four alternatives, the sending actor could choose an outgoing communication tie in 50% of all cases. This equals his actual choice of one repeated communication choice given two observed choices. Because the only parameter of model M_1 is zero, the model is equivalent to the random choice model M_0. The likelihood is the same and equals an equally distributed choice likelihood with four potential receivers: $\frac{1}{4} \cdot \frac{1}{4} = 0.625$.

This is different from the parameter *Reciprocity*: The sending actor could reciprocate a tie in three out of eight choices (37.5%) but chose to reciprocate in 50% of all cases, which is more than expected. According to the equation 3.58 on page 52 the estimate can be derived as $\hat{\beta}_2 = \frac{1}{2} \cdot \log\left(\frac{2}{2} \cdot \frac{3}{1}\right) = 0.549$. If the sender in this model had to decide whether to choose a receiver who previously called the sender (a chance to reciprocate) or a receiver who did not call the sender before, the (sparse) data would suggest that the probability of the reciprocating choice is $\exp(0.549) = 1.7315$ times higher than for non-reciprocating. Hence, reciprocity of ties is 73.15% more likely than non-reciprocity. The likelihood can be calculated as $\frac{\exp(0.549)}{2\exp(0.459)+2} \cdot \frac{1}{\exp(0.459)+3} = 0.0670$

The combined interpretation of both parameters in model M_3 is more complicated. The interpretation of each estimate has to take into account the value of the other estimates. In figure 3.4 on page 40 we observe that the two parameters are never chosen together although a random choice would expect the choice of a bi-directionally connected receiver – given the data – in every fourth choice (25%). Also, an empty choice (calling a completely disconnected receiver) is never made although randomly expected in 37.5% of all choices. In total, there are four potential structures that can be chosen by the sender, as shown in figure 3.8.

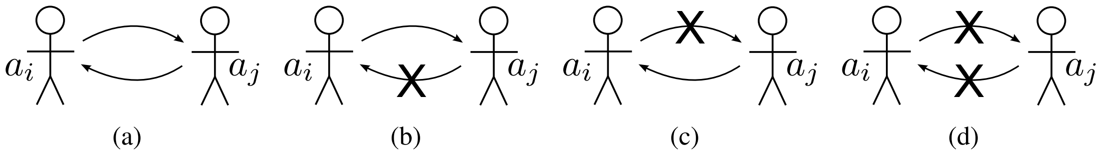


Figure 3.8: Senders (a_i) and receivers (a_j) can be connected by four different dyadic structures. They consist of all possible combinations of the basic structures *Repeated communication* and *Reciprocity*

The sender can choose a “complete dyad”, a dyad with only one tie in either direction or an empty dyad without communication ties between sender and receiver. In the combined model M_3, the probability of choosing a reciprocated tie over a similar structure *without* the reciprocated tie (Structure in figure 3.8(b)) is $\exp(0.609) = 1.8386$ times higher (83.86%). The negative estimate of the *Repeated communication* structure makes the complete dyad less likely: a corresponding receiver choice probability is only $\exp(0.609 - 0.175) = 1.5434$ times higher (54.34%) than the probability to choose a disconnected receiver.

Detailed discussions of receiver choice parameters that were estimated on non-artificial event streams can be found in the application chapters 5 and 6.

3.4 Comparison with Related Models

The introduced basic event framework has similarities to other dynamic network models. In the following, it will be compared to exponential random graph models (ERGMs), to stochastic actor-oriented models (SAOMs) for panel network data and to the relational event framework, a relational extension of event history modeling. The three frameworks were introduced in the literature review sections 2.2, 2.3 and 2.4.

3.4.1 Comparison with Exponential Random Graph Models

The probability in equation 3.19 (p.35) is similar to the graph probability distribution of exponential random graph models (ERGMs, section 2.2, p.13) in equation 2.15 (p.14). A non-parameterized probability function can in both cases be understood as a base line model. The base line model of ERGM is a random graph, while in case of the proposed event framework, the base line model is a random decision over all potential event receivers.

The specification of the probabilities is similar: Both ERGM and the probability of the proposed basic event framework depend on the existence of network structures. The interpretation of estimated effects is similar and is in both cases related to individually driven processes. However, ERGMs are no explicit actor-oriented models. This class provides probability distributions for graphs with structures measured on a global level. However, the set of sufficient statistics is restricted to small, local structures that are assumed to be the result of actor decision processes. This is reflected in the estimation algorithm. Usually, a ministep process similar to the estimation of stochastic actor-oriented models is applied, although the ERGM view is not based on rational choice utility functions. Importantly, no statements on dynamic effects (like: in which order do transitive triangles arise) can be made. This is possible in dynamic event models.

In ERGMs, the denominator is a (often not computable) *constant* over all possible outcomes of a graph. Therefore, ERGMs can be interpreted as exponential family models. With a homogeneity assumption in the proposed Markov process, the expected outcome of the denominator of the multinomial probability in equation 3.19 can be interpreted as a stationary distribution of local environments with different realizations in the denominator of the likelihood function. This is similar to the idea of a normalizing constant.

3.4.2 Comparison with Stochastic Actor-oriented Models

The proposed basic event framework is very similar to the stochastic actor oriented model (SAOM, see chapter 2.3 on page 18). It can be understood as an adaption and extension of the ideas by SAOM. There are, however, some differences that we discuss in the following. Both models describe changes in social network structures as an actor-oriented Markov process. This process is described in two different steps: Both models distinguish between actor activities (modeled as Poisson rates, see equations 2.21, p.19, and equation 3.15, p.34), and multinomial, discrete choices between other actors in the network (see equations 2.22, p.19, and equation 3.19, p.35).

The first noticeable difference is the type of underlying data: SAOMs were developed for network panel data. In contrast, the proposed event framework models the occurrence of relational events. Still, the underlying idea of SAOM is that changes occur in “ministeps”, which can be understood as a specific type of events.

Event data streams usually incorporate a lot more information than network panel data. Between two panel observations, only estimations on the actual number of ministeps can be made. Some changes will not be observed at all, especially when ties change their values more than once. Also, the order of events occurring between two static graph snapshots is unknown. All this information is available in event streams and can thus be exploited.

Second, The actor activity rates of SAOMs (see equation 2.21 on page 19) are usually estimated on a global level. For each two subsequent network observations a general tendency of the actors to undergo change are estimated. The reason for this generalization is the lack of information about individual activity rates. Event analyses allow more detailed analyses of activity rates, beginning with the proposed individual rates. Activity rates can be determined for arbitrary time spans and be parameterized (e.g., with dyadic covariates) in a flexible way. This extension was discussed in chapter 4.

Third, unlike in SAOMs, the proposed event framework does not model *creation* or *dissolving* of binary network ties, but *updating* of ties in a weighted graph. Actors in SAOMs choose between different configurations of their outgoing ties that equal the current configuration, except for one tie. If a tie exists, it can be removed; if no tie exists, it can be created. The decision on which receiver to choose is interpreted as an optimization process of the “structural configuration” of a sender. In the proposed event framework, actors choose event receivers based on the existing structures between sender and potential receivers and on a uniform random process. If there already exists a directed tie between sender and receiver, it is updated – usually the value of the tie would be increased by a certain amount. Dissolving ties is not explicitly modeled as a ministep, but realized by an external decay process. In many weighted event networks this assumption is reasonable. It expresses that ties, which are not updated regularly, have a decreasing importance over time. An intuitive example is communication – if no communication events took place between two actors in a long period, it seems reasonable to assume a lowered communication level.

However, tie creation and dissolving might still be measured on an event level and could be expressed with the proposed model framework: In this case, tie creation and dissolving could be defined as different event types. Receivers (of a creation or dissolving event) could then be chosen based on existing local environments. This would be a special case of the tie update argument. Binary creation and dissolving event streams could, for example, be observed in friendship relations on social network sites. This event type and other event types are discussed in chapter 4.

Fourth, another difference is the specification of the multinomial logit sub models (equations 2.22 and 3.19) that both approaches use to model the “second decision” in the Markov process. In SAOMs, the choice regards graphs that are created by an actor choice. In the proposed event framework the choice regards event receivers; the updates are applied after the choice. The research question in SAOMs is therefore, "which local structures do actors want to create by their choices?", while in the proposed event framework the question is "which actors are chosen as receivers depending on the structures sender and potential receivers are embedded in?". The interpretation of statistics is therefore slightly different. The event-based model, for example, measures the tendency of actors to repeat communication with the same actors, whereas SAOMs measure the outdegree of the sender after the choice. The effects are opposed: If an actor decides to communicate repeatedly with the same receiver, he implicitly decides *not* to increase his or her outdegree. Therefore, a positive parameter estimate of the *Repeated communication* statistic expresses a similar choice behavior than a negative estimate of the SAOM *Outdegree* statistic. Both models, however, evaluate local structures. A practical implication in the proposed event framework is that no gratification function – as in equation 2.23 on page 20 – is necessary.

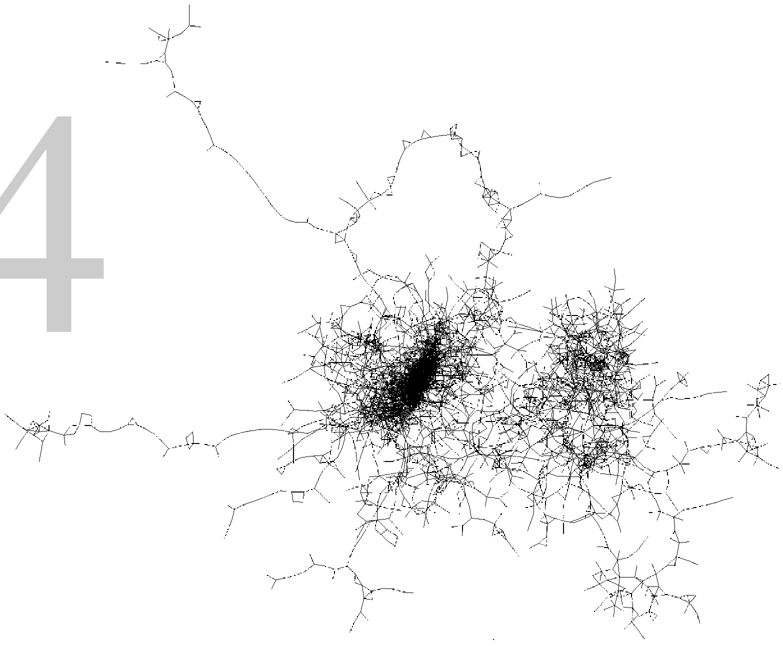
Fifth, in event streams, every single "ministep" of the Markov process is observed. Therefore, it is not necessary to use the method of moments to estimate the different parameters. Instead, both the Poisson rates and the independent variables in the multinomial logit model can be estimated with a maximum likelihood algorithm, as introduced before. This means that the estimation is less complex. It makes simulations redundant that generate possible sequences of ministeps in the MCMC estimation used in SAOMs. As a result, the event based model can be estimated for larger networks and a higher number of events or (estimated) ministeps.

Finally, the implemented features of the two frameworks are different. Estimation of SAOMs is implemented in the software-tool *SIENA* (Ripley et al., 2011). The software developed to estimate the proposed event model framework is called *ESNA* (Stadtfeld, 2011b). In the latest version, SIENA allows the estimation of processes with several dependent variables. Thereby researchers can, for example, compare the relevance of selection and influence effects of actor covariates in longitudinal networks. In the current version ESNA does not support this. However, the set of supported covariates is more elaborate: It can be tested for weighted structures or multi-network structures; we discuss these effects in chapter 4.

3.4.3 Comparison with Event History Modeling

The relational event framework is based on classical event history modeling. Unlike the stochastic actor-oriented model, the relational event framework (see chapter 2.4 on page 22) is not explicitly actor-driven, i.e. not based on the assumption of rational actor choices. It is rather a dyad-oriented model, or as (Butts, 2008, p.167) states, "behavior-oriented". There are two similarities with the framework proposed in this book. First, the underlying data is event data. Second, the occurrence of events in the relational event framework is described by Poisson rates that can be parameterized in a very flexible way. Therefore, the time intervals between events are exponentially distributed. However, it is not distinguished between different decision levels, like actor activities and receiver choices. In an exemplary application, Butts uses a time-discrete sub-model, which only takes the order of events over time into account but not the Poisson process leading to particular time spans. The multinomial probabilities defined for each event observation in this discrete sub-model in equation 2.40 on page 25 look similar to equation 3.19 on page 35. They determine the probabilities among all possible events at that time instead of possible choices of an actor. The relational event model does not allow external processes to change the process states like the external decay function in our model.

4 Extended Specifications of the Event Framework



Parts of this chapter are published in Stadtfeld (2010).

In chapter 3, a new event framework was introduced. This framework allows to model the dynamic occurrence of (communication) events as a stochastic actor-oriented process. Individual activity and multinomial choices of event receivers receiver are understood to be different sub-processes that can be estimated separately. Both sub-processes are assumed to be driven by individual decisions. To keep the introduction comprehensible, only simple specifications were applied in this basic event framework.

The basic framework in chapter 3 is limited in several aspects:

- Available information:
  - Events were described by three elements only (sender, receiver, time).
  - No other graphs or actor attributes were assumed to be available.
- Definition of the process state space:
  - The process state was defined as the state of one graph only.
  - The process state was altered by two update rules only, an increase of ties and a decay function.
- Individual actor decisions:
  - Only two decision levels were modeled (actor activity and receiver choice).
  - The actor activity was assumed to be independent from individual attributes or other parameters.
  - The choice of receiver was defined as a global, homogeneous model.
  - Actors were assumed to consider all actors in the data set as potential event receivers.
- Only two parameters were considered as potentially influencing the choice of event receivers (*Repeated communication* and *Reciprocity*).
- The receiver-choice sub-model was fitted by comparing all possible models. This is not efficient when the number of potential parameters increases.

The basic event model helps to answer a limited set of research questions about individual communication behavior in dyadic network structures: What is the general tendency of actors to communicate in the dataset? Does previous communication increase the probability of future communication? Is there a tendency of actors to reciprocate incoming communication events? However, researchers may want to exploit additionally available information. Or they may have more complex research questions in mind that create the need for additional data collection and new specifications.

In this chapter, we propose a number of extended specifications. If researchers have event data available and want to model individual decision behavior, we propose the following approach. Researchers applying the new stochastic actor-oriented event framework to empirical data sets should ask themselves the following questions when specifying a model:

1. Which information is available in the dataset?
2. How should the state space of the dynamic process be defined?
3. Which individual actor decisions should be modeled?
4. Which structures do *potentially* influence the choice of event receivers?
5. Which of these potential structures are actually good *explanatory variables* of the model?

In this chapter, we discuss extended specifications of the basic event framework that allow to investigate more detailed event streams and more sophisticated research questions. The structure of this chapter follows the five questions above.

Section 4.1 discusses additional information that may have been collected: new event types, the state of other graphs or actor attributes.

Section 4.2 extends the state space of the dynamic model. It is shown how additionally available information can be transformed into the process state.

Section 4.3 proposes extended specifications of individual decision sub processes. These processes are influenced by the state space of the process. Parameterized activity rates, additional independent decision levels, additional process transition rates and individualized receiver choice models are discussed.

Section 4.4 focuses on the specification of the most relevant individual sub-process which is the multinomial choice of event receivers. The new parameters may potentially have an influence receiver choices. The new specifications take the state of different graphs and actor attributes into account.

Section 4.5 presents ideas on how to identify those receiver choice parameters that are actually good explanatory variables. A systematic model-fitting of these parameters is discussed.

Several of the extended specifications introduced in this chapter are applied in the case studies in chapters 5 and 6.

4.1 Additional Information

In the basic event framework, only limited availability of information was assumed: All information was extracted from an event stream of phone call events with known senders, receivers, and time-stamp. Here, additional information that is often observed in event stream data collections is presented. First, events may incorporate additional information. Second, graphs other than the communication graph may be known or be inferred from an event stream. Third, information about individual attributes may be available. Changes of attributes may be expressed in the event stream. Fourth, there may be information about the joining and leaving of actors in the community. If it is not explicitly represented in the event stream, heuristics can be used to differentiate active and inactive actors.

4.1.1 Information in Events

In equation 3.2 (page 28), events $\omega_v \in \Omega$ were introduced as triplets, including the sender index i_v, the receiver index j_v and a time-stamp t_v:

$$\omega_v = (i_v, j_v, t_v).$$

Communication events, however, may include additional information like, for example,

- τ_v: The type of the event (phone call, text message, face-to-face talk)
- η_v: The intensity of the event (length of a phone call)
- γ_v: Information with regard to the contents, like business-related vs. private, or friendly vs. hostile

The event might then be defined as

$$\omega_v = (i_v, j_v, t_v, \tau_v, \eta_v, \gamma_v). \tag{4.1}$$

If such rich information is available in the event stream, it may be useful to exploit it in the actor-oriented decision model. The individual activity rates may be significantly different for events of different types (e.g., text messages and phone calls). Then, separate Markov transition rates can be defined (see section 4.3.3). An actor might also choose the type of the event depending on the kind of relation (friend, colleague, family, ...) to the event receiver. The same holds for event intensities: Friendship might be a good predictor for the length of phone calls. A decision on call lengths could be included in the Markov process transition rates (see section 4.3.2). The decision on the content of an event does probably depends on relations in a similar way.

Events may indicate types of interaction other than (written or spoken) communication events. On social network sites, for example, the creation and dissolving of formalized friendship ties can be expressed as events. The form is equivalent to the general event form in equation 4.1, whereas in case of the friendship example, intensity and information are not necessary. The event type η_v may, for example, be *friendship creation* or *friendship request*. Such events can be directly observed on many social network sites.

Event receivers do not necessarily have to be human actors. The affiliation of an actor with an entity (e.g. a user on a social network site posting a picture) can be expressed as an event.

All these new types of non-communication events are related to graphs (e.g. a formalized friendship graph, picture affiliation graph) other than the communication graph (see section 4.1.2). Information in other graphs can be exploited to predict individual actor communication decisions (see section 4.4.3).

Attributes such as gender, sex, age, or location can be assigned to actors (see section 4.1.3). Some attributes are constant (gender, sex) but others (age, location) change over time. These changes can be expressed as non-dyadic events. The general form in equation 4.1 can be reduced to indicate attribute changes

$$\omega_v = (i_v, t_v, \tau_v, \eta_v)$$

with i_v being the corresponding actor, t_v the time-stamp, τ_v the attribute type (e.g. *cell tower the actor is connected to* or *age*), and η_v the corresponding value (the concrete cell tower id, the new age). If the location of an actor is measured by the next celltower he is connected to, a change of the physical location could be indicated by such an event. Individual attribute values can be used for a better description of receiver choices in the event stream (see section 4.4.4).

The joining and leaving of actors in an observed community might be measured as events (see also section 4.2.2). Similar to the attribute change example, a simplified form of the general event type in equation 4.1 can be used:

$$\omega_v = (i_v, t_v, \tau_v)$$

where τ_v is *joining the community* or *leaving the community*. Detailed information about the current actors in the community can be exploited for a better description of the receiver choice probabilities (see also section 4.2.2).

4.1.2 Other Graphs

In addition to the communication graph, it is possible to evaluate the state of other graphs in the Markov model. These graphs can be derived from the event stream similar to the communication graph in the basic event framework. In section 4.1.1, the example of "formalized" friendship relations on a social network site was discussed. These events may be translated directly into a *formalized friendship graph*. Other graphs may be assumed to be stable over time – either because the data were logged once only (e.g. when collecting friendship, trust or advice relations by a questionnaire once) or because the relation is hardly dynamic by definition (graphs representing hierarchy, family relations, marriage). Other graphs may be derived from other information: A graph representing dyadic covariates like attribute similarity can be inferred from the current actor attributes, graphs that represent time spans between events can be inferred from the event time spans, graphs representing a spatial distance may be inferred from attributes indicating the current location of actors.

The dynamic choices in an event model may depend on any of these graphs. In the basic event framework, we only allowed (dyadic) structures in the communication graph to be independent variables of individual choice processes. The following types of graphs are now defined:

- Weighted, directed actor graphs with vertices being actors (e.g. the communication graph),
- Binary, directed actor graphs with vertices being actors (e.g. friendship nominations measured with a questionnaire),
- Binary, undirected graphs with vertices being actors (e.g. formal friendship on a social networks site),

- Weighted, undirected graphs with vertices being actors (e.g. last communication time, spatial distance),
- Two-mode graphs with actors connected to other entities (e.g. actors "liking" media items, actors being affiliated with a project).

In the basic event framework, $X(t)$ denoted the communication graph X at time t. In the following, different graphs are distinguished by a unique number

$$X^{(1)}(t), X^{(2)}(t), \ldots.$$

Edges (directed or undirected) of the graphs on the set of actors (for example, $X^{(1)}$) will in the following be denoted by $x_{ij}^{(a)}$ with i and j being actor indices in $\{1,\ldots,n\}$ and a being the number indicating the graph. Edges of the entity affiliation graphs will be denoted similarly by $x_{ie}^{(a)}$ with e indicating a concrete entity index with e in $\{1,\ldots,q\}$. The communication from section 3.1.2 (page 30ff) will in the following be denoted by $X^{(1)}$.

4.1.3 Actor Attributes

Individual actor attributes may be defined as changeable (location, age) or as constant (gender, date of birth). Some attributes may be assumed to be constant, if the observed time span is rather short (expertise, formal role in a group, position within a hierarchy). Changeable attributes may be updated by attribute events as discussed in section 4.1.1. Formally, to each actor a_i with index i an attribute vector $y_i(t)$ is assigned at time t:

$$y_i(t) = (y_{i;1}(t), y_{i;2}(t), \ldots) \tag{4.2}$$

Attributes can be measured metrically (e.g. distance), categorically (e.g. gender) or ordinally (e.g. level of expertise).

4.2 Defining the Process State

The additionally available information from section 4.1 can be defined as part of the state of the Markov process. Thereby, individual decisions can be modeled as depending on this additional information. In this section, we explain how available information in the dataset can be transformed into an extended state space.

4.2.1 Update Rules

According to section 4.1.1, extended events may change the communication graph, other graphs, actor attributes, and the set of active actors (actors after joining and before leaving the community). In section 3.1.3 (page 31), two update rules of the graph were introduced: First, whenever an event takes place, the corresponding (directed) tie in the communication graph is increased by one. Second, in the time between events, all ties in the graph decay with an exponential function. Similar to these rules we now, exemplarily, define some additional update rules.

Events may indicate a creation or dissolving of ties. This idea is different from the update idea of the communication graph and close to the idea of *ministeps* in stochastic actor-oriented models for panel data (see Snijders (2005, p.224) and section 2.3). On social networks sites, for example, the creation and dissolving of formalized friendship relations can be observed as events. In organizations, a starting time and an end time of a collaboration between two actors may be expressed as events. In both cases, the proposed update and decay rules should be replaced by creation and dissolving rules. These binary update rules can be formulated similarly to the update rule in equation 3.12 (page 31):

$$r^{(3)}(X^{(1)}_{\tilde{\nu}}; i_\nu, j_\nu) = X^{(1)}_\nu = (x^{(1)}_{\nu;kl}) = \begin{cases} v & \text{, if } k = i_\nu, l = j_\nu \\ v & \text{, if } \exists \omega_\phi : k = i_\phi, l = j_\phi, t_\phi = t_\nu \\ x_{\tilde{\nu};kl} & \text{, else} \end{cases} \tag{4.3}$$

with $v = 1$ for the creation of ties and $v = 0$ for dissolving.

Events ω_ν are sometimes weighted with an intensity η_ν. For example, the length of a phone call may be included in the data set. It may make sense to increase directed tie values in the communication network even more, if a call was longer. An update function of communication graph $X^{(1)}$ could be defined as follows:

$$\begin{aligned} r^{(4)}(X^{(1)}_{\tilde{\nu}}; i_\nu, j_\nu) = X^{(1)}_\nu &= (x^{(1)}_{\nu;kl}) \\ &= \begin{cases} x^{(1)}_{\tilde{\nu};kl} + f(\eta_\nu) & \text{, if } k = i_\nu, l = j_\nu \\ x^{(1)}_{\tilde{\nu};kl} + f(\eta_\nu) & \text{, if } \exists \omega_\phi : k = i_\phi, l = j_\phi, t_\phi = t_\nu \\ x^{(1)}_{\tilde{\nu};kl} & \text{, else} \end{cases} \end{aligned} \tag{4.4}$$

with $f(\eta_\nu)$ being a real function, for example, a linear function.

In some cases, the entering and leaving of actors in a community is measured with time stamps. These (non-dyadic) events can trigger an update rule that adds a node to the communication graph or removes it. If such information is available, it should be used in this way, because the multinomial choice model of event receivers then can only evaluate those actors as potential event receivers, who can technically receive events. The set of *active actors* is named $A^+(t)$ at time t and is updated with each event of the type *joining* or *leaving*:

$$A^+(t_\nu) = \begin{cases} A(t_{\tilde{\nu}}) \cup a_{i_\nu}, \text{ if } \tau_\nu = \textit{joining} \\ A(t_{\tilde{\nu}}) \backslash a_{i_\nu}, \text{ if } \tau_\nu = \textit{leaving} \end{cases} \tag{4.5}$$

In case joining and leaving are not reflected by events, heuristics can be applied to determine the set of (potentially) active actors. This idea is presented in section 4.2.2 and applied in the application chapter 5.

If the time since the last event is supposed to be tested as an independent variable (see the extended specifications in section 4.4), time intervals without communication can be encoded as a *time count graph*. For time spans between two events all tie values would increase linearly similar to the exponential decay of the rule in equation 3.13 on page 31:

$$r^{(5)}(X^{(g)}_{v-1}, \delta_v) = X^{(g)}_{\tilde{v}} = (x^{(g)}_{\tilde{v};kl}) = x^{(g)}_{v-1;kl} + \delta_v \tag{4.6}$$

Graph $X^{(g)}$ is the time count graph. Use of this information in individual actor decision models is explained in section 4.4.3. If an event ω_v takes place, the tie $x^{(g)}_{i_v j_v}$) in the time count graph is reset to zero (see update rule 4.3 with $v = 0$).

Many more update rules are possible. The application in chapter 5, for example, uses different rules to update or delete ties in graphs, express a decay over time, and update the set of active actors.

4.2.2 Heuristics for Potential Receivers

There are several reasons why the set of potential event receivers might need to be reduced. Especially in very big data sets, it is sometimes not reasonable to assume that an individual's choice about event receivers considers all actors in a network. Imagine a network with millions of nodes. If the constant in equation 3.19 (page 35) evaluates millions of potential choices, then this may lead to numerical problems. Therefore, it may be reasonable to restrict the actually made choice to a certain neighborhood of the sender, for example, all receivers within the distance of three steps (the shortest path in the graph is less or equal to 3). All other observed choices that do not take place within this environment can then be estimated separately. This idea is discussed in the following section.

Sometimes, it is the research questions that suggest a limitation of the set of potential receivers. In Zenk et al. (2010), an event-based analysis was applied (using the framework of this book) that only estimated receiver choice effects within formal groups that were found to be sub-sets of the overall communication graph. Any other communication was ignored. In the denominator of equation 3.19, only members of the sender's group were evaluated.

Sometimes, information about the joining and leaving of groups may not be available in the data set, but actors actually join and leave the observed community. This is the case with the data used in the application chapter 5. In such a case, a heuristic can be applied based on the connectivity of actors, which distinguishes active from inactive actors in a community. Once again, the denominator of equation 3.19 can be adapted to evaluate active actors only. Section 4.3.2 explains how the rarely observed choice of inactive actors can be expressed by additional decision levels in the Markov process Poisson rates.

A^+ indicates a reduced set of actors ($A^+ \subseteq A$) with $A^+ \backslash \{a_{i_v}\}$ being the set of potential receivers of event ω_v.

4.2.3 Extending the State Space of the Process

In equation 3.24 on page 36, the state space of the Markov process was defined as the set of possible communication graph realizations. The Markov process models the occurrence of events in the event stream and assumes that the future development from a point in time t

depends on the current process state only. In the last sections, however, we discussed that additional information may be relevant when individual event decisions are made.

Therefore, the process state (which is still called $X(t)$ at time t) is extended to incorporate the state of different graphs (see section 4.1.2), actor attributes (see section 4.1.3), the set of "active actors" A^+ (4.2.2):

$$X(t) = (X^{(1)}(t), X^{(2)}(t), \ldots, y_1(t), \ldots, y_n(t), A^+(t)) \tag{4.7}$$

The process state time t_ν (when event ω_ν takes place) is denoted by

$$X_\nu := X(t_\nu) \tag{4.8}$$

with elements $X_\nu^{(1)}, X_\nu^{(2)}, \ldots, y_1(t_\nu), \ldots, y_n(t_\nu), A_\nu^+$. The process state at time $t_\nu - \varepsilon$ (directly before the event-triggered updates of event ω_ν are applied) is denoted by

$$X_{\tilde{\nu}} := X(t_\nu - \varepsilon). \tag{4.9}$$

The state space could even be extended further. It might be of interest to differentiate global process regimes. A community, for example, may be in a general growth phase. This might influence the individual decisions as well. We discuss this issue in chapter 5. It may also be of interest to investigate global "flows of information" that affect individual actor decisions. The global distribution of information may also extend the state space. A simple start is to observe sequences of events. Partly, this idea can be expressed by including additional graphs that represent information flows.

Given such an extended state space, some exemplary research questions (that are partly asked in the application chapters 5 and 6) are:

- Does the existence of common contacts of sender and receiver increase the choice probability? See chapters 5, 6.
- Does affiliation with the same media items increase the probability of communication? See chapter 5.
- Can communication choices be predicted better, if friendship ties are considered?
- Is communication more likely, if the same actors communicated within the last day? See chapter6.
- Does the spatial distance between actors have an effect on communication choices? See chapter 6.
- Do actors with similar interests have a higher probability of choosing each other as event receiver?

We discuss the necessary extended specifications in section 4.3. Structural receiver choice parameters are discussed separately in section 4.4.

4.3 Specifying Individual Decisions in the Model

In section 4.2 we explained how the state of the process can be extended. The new state space can be understood as to be changed by a number of different individual decisions. Therefore, we present extended activity rates, additional decision levels, additional transition rates, and the specification of individual receiver choice models.

4.3.1 Parameterized Activity Rates

In the basic model, the first individual decision was defined to be the general actor activity of sending events (see section 3.2.1). It is modeled as a Poisson process with parameters ρ_i. This means that the time spans between two events of actor a_i are ρ_i-exponentially distributed. The basic idea can be extended by parameterizing the Poisson rates.

One interesting question is whether there is a significant difference in communication activity between males and females or times of the day. Actors in a community might have different formal or informal roles that potentially influence the activity rates. Similar to the Poisson rates proposed by Butts (2008), the individual rates can then be parameterized in the following way

$$\rho_i(t) = \exp(\theta^T \sigma(i,t)) \tag{4.10}$$

where $\sigma(i)$ is a vector of actor- or time-covariates. For example, $\sigma_1(i,t) = 1$ may be measured if the time t is daytime (0 at night). $\sigma_2(i,t) = 1$ may be measured, if a_i is female (0 if male). Vector θ^T is similar to vector β^T in equation 3.19 and represents a weight of the attributes. It can be estimated with a maximum likelihood estimation.

Alternatively, several activity rates for each actor can be estimated. In chapter 6 the rates for each hour of the day are estimated separately on an individual level.

The communication activity may also depend on dyadic covariates – friends might have a higher communication frequency – or on structures in the communication graph. Such a parameterization is also possible. However, it is associated with the problem of the independence assumption between the two basic decision levels being violated easily. If the researcher is, for example, interested in the effect of friendship on the multinomial choice of receivers, this effect should not be tested as a parameter of the activity rates in the same model.

Sometimes, the proposed individual parameter rates might already be too specific. Then, parameter rates can be defined to be homogeneous for all actors in the dataset. This estimate then equals the activity rates of stochastic actor-oriented models for panel data which are usually estimated on a global level (see Snijders (2005, p.224)).

4.3.2 Additional Decision Levels

Section 4.3 introduced how to model a range of new individual decisions processes. In this section, we now focus on the core decision process that describes the choice of event receivers.

In the basic event framework presented in section 3.2, actor choices consisted of two decision levels:

- First, each actor has an individual activity that determines the event starting times.
- Second, given the determination of sender and event time, the sender chooses a receiver based on graph structures.

Both decisions are assumed to be independent. This means that we assume that the activity level differs among actors and is independent of the receiver choices. Although this assumption may be critical in some contexts, it allows a very straightforward estimation of the process parameters. As long as there are good reasons for assuming independence from the two initial decisions, further decision levels can be modeled, thus expressing the transition rates as a product of the independent decision probabilities.

The update of message ties in the communication graph can be modeled as depending on the intensity of a phone call or the type of the event. We might assume that in a phone call more information is transmitted than in case of a text message and that, therefore, a communication tie should be increased more. The same could be assumed for long phone calls compared to very short phone calls. The Markov process transition rates of equation 3.26 (page 37) can then, for example, be extended to

$$\lambda_{ijmd}(X(t);\rho_i,\beta) \approx \rho_i p_1(j;i,X,\beta) p_2(\omega.\tau = m) p_3(\omega.\eta = d) \tag{4.11}$$

with the new probabilities p_2 and p_3 being the probabilities of choosing event type $\omega.\tau$ equals m (e.g. *phone call*) and intensity $\omega.\eta$ equals d. The probability p_1 (formerly denoted by p) is the probability of choosing receiver a_j. The probability p_2 might be modeled as a multinomial logit model, p_3 might be modeled as a continuous probability distribution. Then, λ_{ijmd} is the Poisson rate defining the propensity of an event from a_i to a_j with the means of communication m and an intensity of d.

Furthermore, it is possible to use additional decision levels to limit the set of potential receivers as discussed in section 4.2.2. A^+ indicates a subset of all actors. Imagine that, for example, only "close" actors are considered as potential receivers of an event to reduce the computational complexity in very big networks. Close actors could be those that are only up to three hops away from the event sender. Then, the process transition rate λ_{ij} can be specified as:

$$\lambda_{ij}(x;\rho_i,\beta,p^+) \approx \begin{cases} \rho_i p_1(j;i,X,\beta) p^+ & ,\text{if } a_j \in A^+ \textbf{ (i)} \\ \rho_i \frac{1}{|A^-|}(1-p^+) & ,\text{if } a_j \in A^- \textbf{ (ii)} \end{cases}. \tag{4.12}$$

A heuristic determines whether the sender "chooses" to send the event to a close receiver in A^+ with Bernoulli probability p^+. The probability of choosing an actor that is far away in the communication graph (the receiver is in $A^-;A^- = A - A^+$) is $(1-p^+)$. If p^+ is high and A^+ is small compared to A^-, this heuristic leads to a significant reduction of the computation time. Only if the event is sent to an active actor, is the receiver choice determined by a multinomial choice model with probability p_1. In the denominator of p_1 (see equation 3.19) only actors in A^+ are then evaluated. Otherwise, the choice likelihood equals a random choice over all actors in A^-. p^+ can – as p_2 and p_3 in equation 4.11 – be estimated independently of the other parameters.

Reductions of the set of potential receivers due to other reasons can be modeled similarly. In chapter 5, the choice is restricted to potential receivers that were assumed to be "active". In that case, A^+ only includes actors that are not isolated in at least one of several graphs.

4.3.3 Additional Transition Rates

The assumption of different levels of individual decisions being independent may not be reasonable in some cases. For example, a pre-analysis of the data set at hand could reveal that some actors in a mobile communication data set only write text messages and other actors only make phone calls. Then, it would be critical to assume that the individual activity rates and the choice of the means of communication are independent. Also, text messages could be written at a different rate than phone calls are made. The choice of event receivers may, however, still be independent of the activity and a specific means of communication.

In this case, it is advisable to define separate and competing individual Markov process transition rates for text messages and phone calls.

The text message (*tm*) transition rate can (similar to equation 3.26) be defined as

$$\lambda_{ij}^{tm}(X(t);\rho_i^{tm},\beta) \approx \rho_i^{tm} p(j;i,X(t),\beta) \tag{4.13}$$

and the phone call (*pc*) transition rates as

$$\lambda_{ij}^{pc}(X(t);\rho_i^{pc},\beta) \approx \rho_i^{pc} p(j;i,X(t),\beta). \tag{4.14}$$

For each means of communication, the activity rates then can be estimated separately. Each actor a_i then has two activity rates. ρ_i^{tm} describes the individual propensity of a_i to write text messages, ρ_i^{pc} the propensity to make phone calls. The multinomial receiver choice can be estimated independently of the other decision levels – no matter which means of communication was determined by the activity rates.

These transition rates define competing Poisson processes. The probability of actor a_i to choosing a text message as the next event type is $\frac{\rho_i^{tm}}{\rho_i^{pc}+\rho_i^{tm}}$ (Waldmann and Stocker, 2004, p.92).

Before analyzing complex data sets with a high number of potential decision levels, it is always advisable to test the independence assumptions before deciding on the definition of decision levels and competing transition rates.

4.3.4 Individual Receiver Choice Models

In section 3.2.2 the choice of event receivers was defined as a multinomial logit model. The maximum likelihood estimates are calculated on a global level (see equation 3.34, page

41). The results, therefore, reflect general and homogeneous choice patterns of the whole population. This is very helpful when it is aimed at explaining general behavior. However, *individual* choice patterns may be of interest in some cases. A typical example is marketing research: Research questions in this area are often directed to understanding *differences* among individuals to identify – in a second step – groups, segments or distributions of consumers.

To estimate the individual behavior of actor a_i, the event stream Ω has to be reduced to incorporate only those events in which a_i is the sender:

$$\Omega^i = \{\omega_v : \omega_v.sender = i\}. \tag{4.15}$$

The individual parameter set $\hat{\beta}^i$ can then be calculated as the maximum likelihood estimate of the likelihood $L(\Omega^i;\beta^i)$ as explained in section 3.3.5 (page 45ff). However, it has to be ensured that the analyzed part of the event stream includes enough individual choices so that significant estimates for individuals can be retrieved.

4.4 Specification of Structural Receiver Choice Variables

Event receivers in the context of social networks are not determined randomly. These choices can often be described well by taking network structures and actor attributes into account. For example, people often tend to communicate with people they have communicated before or with whom they have friends in common.

The multinomial receiver choice model was introduced in equation 3.19 on page 35. In section 3.3.2 (page 39) it was specified using simple, dyadic communication structures as independent variables. Binary, dyadic communication structures are important predictors for communication choices, but communication structures with more than two participants are also very important (see chapters 5 and 6). The introduced estimation procedure can be applied, if more complex graph statistics and attribute statistics are defined as independent variables of the multinomial choice model. According to section 4.2.3, we assume that the state $X(t)$ of a Markov process at time t is defined by a set of different graphs and individual vectors of attributes:

$$X(t) = (X^{(1)}(t), X^{(2)}(t), \dots, y_1(t), \dots, y_n(t))$$

where $X^{(1)}$ denotes the communication graph as defined in section 3.1.2.

In the following sections, different types of independent variables of the receiver choice sub-model will be discussed separately. Endogenous communication statistics (explaining communication by previous communication) are discussed in section 4.4.1, entity affiliation statistics in section 4.4.2, other network statistics in section 4.4.3. Statistics measuring actor attributes are explained in section 4.4.4, the influence of edge weights is discussed in section 4.4.5. The possibility of combining basic statistics to complex statistics is introduced briefly in section 4.4.6. Many of the structures proposed for event decisions are used in a similar form in stochastic actor-oriented models for network panel data and are implemented in the software SIENA (see Snijders (2001)). An overview of structures used in exponential random graph models can be found in Snijders et al. (2006). A discussion of two-mode structures was introduced by Wang et al. (2009)). An overview of possible (undirected) three-node structures in graphs is given in Wasserman and Faust (1994, p.566). The specification of event-based models is a bit different from these examples, as only local structures are evaluated and the underlying graphs are often weighted.

4.4.1 Endogenous Communication Statistics

Choices of event receivers can partly be explained *endogenously* by previous communication choices. Figure 4.1 shows six different structures that are measured in the local environment of event sender and receiver in the communication graph before an event takes place. The structures shown are not a complete overview, but only an exemplary selection. More structures can, for example, be derived by permuting ties in the triadic structures or by combining the basic structures in figure 4.1. This idea is discussed further in section 4.4.6. In each structure, the actor on the lower left is the event sender, the actor on the lower right is the event receiver. Arrows indicate positive, directed communication ties in the communication graph that represent recent event sending. The communication graph is weighted but weights are not depicted as all statistics introduced in this section are measured as binary variables. Weighted extensions are discussed in section 4.4.5.

The first two structures in figures 4.1(a) and 4.1(b) are *Repeated communication* and *Reciprocity*. They were introduced in the dyadic example of the basic event framework in section 3.3.2. The statistics were introduced in equations 3.28 and 3.29 (page 39). We assume an extended state space with $X^{(1)}(t)$ being the communication graph at time t. The first two statistics can be defined as

$$s_1(i,k,X_{\tilde{v}}) = \begin{cases} 1, \text{ if } x^{(1)}_{\tilde{v};ik} > 0 \\ 0, \text{ else} \end{cases} \tag{4.16}$$

$$s_1(i,k,X_{\tilde{v}}) = \begin{cases} 1, \text{ if } x^{(1)}_{\tilde{v};ki} > 0 \\ 0, \text{ else} \end{cases} \quad . \tag{4.17}$$

The statistics measure whether a positive communication tie from actor a_i to a potential receiver a_k or from a_k to a_i exists. If so, the function is 1, otherwise it is 0. The sending actor has the index i, the receiver has the index k. $X^{(1)}_{\tilde{v}}$ is the communication graph directly before event ω_v. The first statistic is very important: It expresses the tendency not to initiate new, costly communication ties, but to repeatedly use existing communication paths. The importance of this behavior is discussed in the results sections of chapters 5 and 6. The effect is a local, actor-oriented equivalent to the *density* effects in stochastic actor-oriented models and exponential random graph models that control the global network density.

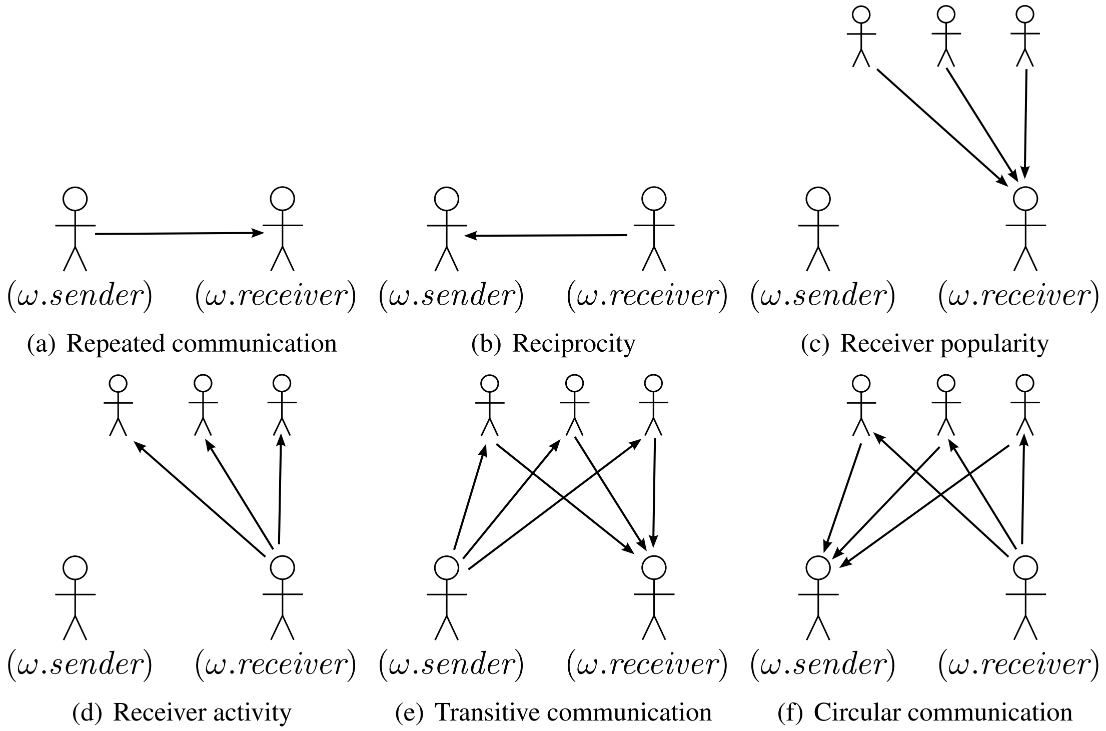


Figure 4.1: Endogenous communication structures that explain communication choices by previous communication of the actors in the graph.

The structure in figure 4.1(c) is measured by the number of incoming ties of the receiver. The statistic expresses, whether there is a higher likelihood of communication with "popular" actors. As popularity in this context is only related to the number of incoming communication ties, it cannot be interpreted as social status. The structure in figure 4.1(d) is similar. It expresses the tendency of choosing receivers with a high outdegree or "activity". Activity in this context is not understood to be the activity rate that determines the number of events started in a time interval (see section 3.2). It is rather defined as the *number of communication partners* the receiver communicates with. The statistics measuring these structures are (using the notation above and from section 3.3.2) formally defined as follows:

$$s_3(i,k,X_{\tilde{v}}) = \sum_{l \in R_v \setminus \{k\}} s_1(k,l,X_{\tilde{v}}) \tag{4.18}$$

$$s_4(i,k,X_{\tilde{v}}) = \sum_{l \in R_v \setminus \{k\}} s_2(k,l,X_{\tilde{v}}) \tag{4.19}$$

R_v is the index set of potential receivers in $A \setminus \{a_{i_v}\}$ (see equation 3.5, page 29). The last two structures in figures 4.1(e) and 4.1(f) also evaluate ties between the sender and a *third* actor. The structures are measured to test a tendency to communicate in transitive relations or circles. *Transitivity* in figure 4.1(e) expresses whether there is a higher likelihood for direct communication, if two-step communication paths already exist between the sender and a potential receiver. Directed paths of a certain length can be interpreted as information flow. The same holds for the *Circle* structure in figure 4.1(f). It can be measured to test whether communication is more likely, if previous two-step communication existed from the receiver to the sender. By choosing a specific receiver, the event sender closes a potentially unclosed circular structure including three nodes. The structure could be extended to measure only circular structures that are already closed at event time. The statistics measuring the structures *Transitivity* and *Circle* are defined as

$$s_5(i,k,X_{\tilde{v}}) = \sum_{l \in R_v \setminus \{k\}} s_2(k,l,X_{\tilde{v}}) s_2(l,i,X_{\tilde{v}}) \tag{4.20}$$

$$s_6(i,k,X_{\tilde{v}}) = \sum_{l \in R_v \setminus \{k\}} s_1(k,l,X_{\tilde{v}}) s_1(l,i,X_{\tilde{v}}). \tag{4.21}$$

Two similar (and not depicted) structures count the number of third actors who previously sent a message to both sender and receiver or received a message from both. These statistics can be defined similarly to equations 4.20 and 4.21.

The structures in figures 4.1(a) and 4.1(b) and a triadic structure similar to those in figures 4.1(e) and 4.1(f) are tested in both application chapters 5 and 6. Chapter 6 also tests *Popularity* and *Activity* (named *Receiver Indegree* and *Receiver Outdegree*). In the application chapters most of the endogenous structures turn out to increase the probability of communication choices. Detailed discussions can be found in sections 5.5 (page 101ff) and 6.5 (page 121ff).

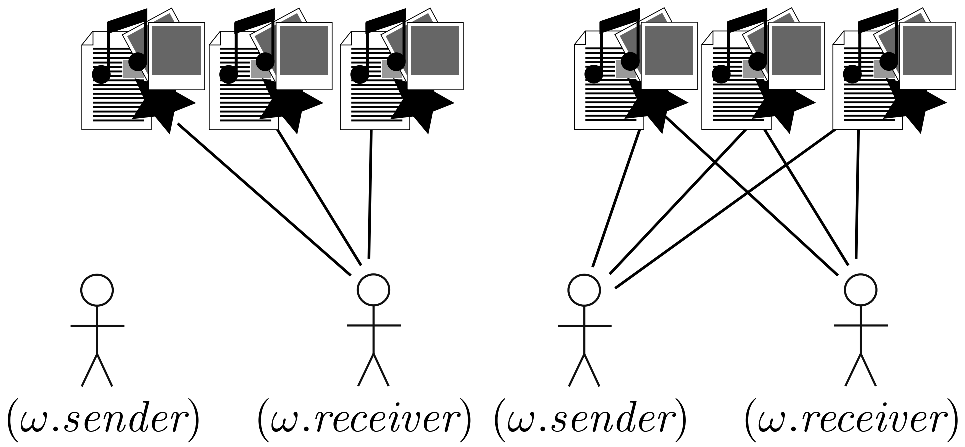
4.4.2 Entity Affiliation Statistics

Communication choices may depend on the entity affiliation of the event receiver or on the *common* entity affiliation of sender and receiver. In social media environments, for example, communication may depend on media items actors are connected with: People may be chosen as event receivers, because they are connected to videos, images or texts – sometimes without having a common communication history or common contacts with the event sender. In organizations, the affiliation with a common project may influence communication choices. Effects of other entity affiliations are possible. This actor-entity affiliation can be represented by two-mode graphs. If such data are available and there are reasons to suggest that an influence is possible, it is proposed to test the effect of entity affiliations on communication choices. Figure 4.2 shows two different structures that could be tested for entities of different types. In both cases, the event sender is shown on the left and the event receiver on the right. The icons at the top depict entities to which the actors are affiliated (e.g. media items).

In figure 4.2(a) only the receiver is connected to entities, in figure 4.2(b) both sender and receiver are connected to the same entities. If these structures are measured on a two-mode graph $X_2(t)$ at time t, the corresponding statistics simply count the affiliation structures:

$$s_7(i,k,X_{\tilde{v}}) = \sum_{e \in E} x^{(2)}_{\tilde{v};ke} \tag{4.22}$$

$$s_8(i,k,X_{\tilde{v}}) = \sum_{e \in E} x^{(2)}_{\tilde{v};ie} x^{(2)}_{\tilde{v};ke} \tag{4.23}$$



(a) Receiver connected to entities (b) Both connected to the same entities

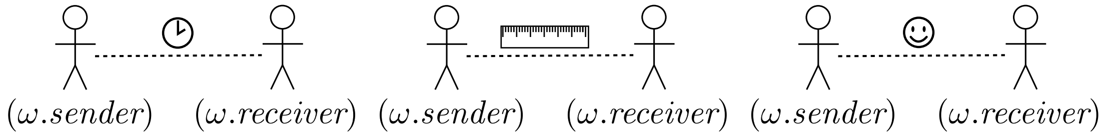
Figure 4.2: Communication influenced by entity affiliation

with e being an entity in the entity set E and the two-mode ties of $x^{(2)}_{\tilde{v};ke}$ being 1, if an affiliation between actor a_k and entity e exists and else 0. The entity affiliation graph $X^{(2)}(t)$ is part of the process state at time t.

The influence of entity affiliation statistics on communication choices is tested in the applied chapter 5. In this chapter we investigate whether the affiliation of actors in a question and answer web community is influenced by the affiliation of actors to recently and publicly posed questions.

4.4.3 Other Actor Graph Statistics

Structures in *actor graphs* other than the communication graph may influence communication choices. Three examples are shown in figure 4.3. Sender and receiver are shown on the left and right. The type of the graph is denoted by an icon on the tie between the actors.



(a) Time since the last dyadic event (b) Spatial distance between sender and receiver (c) Friendship between sender and receiver

Figure 4.3: Communication influenced by actor graphs other than the communication graphs

The time since the last event from one actor to another may be encoded in a graph. Then, it is possible – without violating the Markov property of the event framework – to test whether the time since the last event influences future communication. The corresponding structure is depicted in figure 4.3(a). In chapter 6 this independent variable is tested. The influence of spatial distance on event receiver choices can be measured similarly. It is also estimated in the application chapter 6. Figure 4.3(b) presents this structure. The existence of friendship relations may also have an influence on communication choices. This structure is shown in figure 4.3(c). Friendship in this context may be "real" friendship relations that are measured with a questionnaire or formalized friendship relations as those on social networks sites. If the "other graph" (time, distance, or friendship) is undirected and denoted by $X^{(3)}$, a simple statistic can be defined as

$$s_9(i,k,X_{\tilde{v}}) = x^{(3)}_{\tilde{v};ik}. \tag{4.24}$$

where the independent variable is the value of the tie. In case of the time and distance graph, it is a positive, real value. In case of the friendship graph, it is either 1 or 0. The statistic can be extended to directed graphs and represent structures with more than one tie, similar to the endogenous statistics discussed in section 4.4.1. For example, the friendship popularity might be of interest or triangular effects regarding the spatial distance. Contrary to the endogenous examples, spatial distance and the time since the last event are not binary statistics, but weighted. Estimates of the statistic in figure 4.24 can then be interpreted in terms of "unit of change". A maximum likelihood estimate $\hat{\beta}_9$ expresses that with each additional unit of time or distance, respectively, the probability of a choice is changed by $e^{\hat{\beta}_9}$. In case of distance, this unit could be kilometers. In case of time, the unit could be hours. Weighted network statistics are further discussed in section 4.4.5.

The influence of other graphs connecting the actors on communication choices is tested in chapter 6. In this chapter we investigate whether the time since the last interaction and spatial distances have an effect on the choices of communication partners.

4.4.4 Actor Attribute Statistics

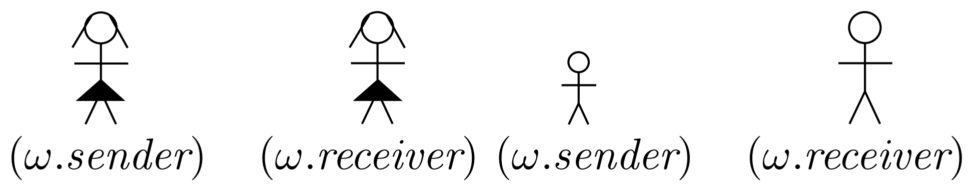
Actor attributes can be important predictors for social network relations. Steglich et al. (2010), for example, discuss a model for social selection and influence processes. Monge and Contractor (2003, p.223 ff) discuss the concept of homophily (selection of similar others). It could be of interest, whether actors with similar attributes are more likely to communicate.

In this section, it is not distinguished between unchangeable or rarely changing attributes (e.g., gender, year of birth, nationality, race) and changeable attributes (e.g., location, age, beliefs, interests), although such a distinction is important when analyzing dynamics of attributes. Figure 4.4 exemplarily illustrates a gender-homogeneous selection, figure 4.4(b) an age-heterogeneous selection.

Regarding a binary attribute (like gender), homophily can be tested as

$$s_{10}(i,k,X_{\tilde{v}}) = \begin{cases} 1, \text{ if } y_{i;1}(t_v) = y_{k;1}(t_v) \\ 0, \text{ else} \end{cases} \tag{4.25}$$

with $y_{i;1}(t_v)$ being the first attribute in attribute vector $y_i(t_v)$ of actor a_i. It includes, for example, the gender of a_i (see equation 4.2). The individual attribute vectors $y_i(t)$ at time t



(a) Gender homophily influencing communication choices

(b) Age heterophily influencing communication choices

Figure 4.4: Communication choices influenced by gender and age

are part of the process state, which is $X_{\tilde{v}}$ instantly before the event takes place. Homophily of an attribute with continuous range (like age) may be tested with a statistic measuring the distance:

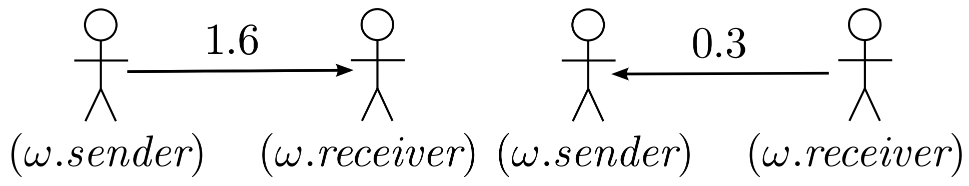
$$s_{11}(i,k,X_{\tilde{v}}) = |y_{i;2}(t_v) - y_{k;2}(t_v)| \tag{4.26}$$

Here, $y_{i;2}$ is the age of actor a_i. The interpretation of the parameter is then related to an increase or decrease of communication probabilities "with each additional time unit" (year/month/...) of age difference. Instead, a metric attribute can be categorized and only sender-attribute combinations in the same category can be evaluated as a homogeneous selection. Given vectors of attributes $y_i(t_v)$ at time t_v it is also possible to define a more general measure of distance (or closeness) between actors based on multiple vector entries.

4.4.5 Weighted Statistics

The communication graph $X^{(1)}$ has weighted edges. Whenever an event takes place between actors, the value of the corresponding, directed tie increases. In the time between events, all tie values decay with an exponential function. These basic update rules were introduced in section 3.1.3. So far, the weight of the ties has not been considered in the endogenous graph statistics. The reason is that the interpretation of estimated *weighted* statistics is less straightforward. Parameter estimates are usually interpreted "per unit of change" (see section 3.3.6, page 49ff). Given the popularity structure from figure 4.1(c), we can interpret a corresponding maximum likelihood parameter $\hat{\beta}_3$ as follows: With each additional incoming communication tie, the probability of choosing actor a_j increases by $e^{\hat{\beta}_3}$.

This is less straightforward with weighted structures, as a measured statistic has no natural unit in this context. The distribution of tie values depends on the predefined update rules. If a tie has twice the value of another, then this does not necessarily mean that the communication intensity is "twice as high". Still, weight differences can be an important predictor for communication choices, so the inclusion of weighted parameters should be considered. In figures 4.5(a) and 4.5(b) a weighted outgoing tie to the sender and a weighted incoming tie from the receiver are illustrated.



(a) Weighted repeated communication

(b) Weighted reciprocity

Figure 4.5: The weight of ties can be taken into account.

The statistics can be defined as the value of the measured dyadic tie:

$$s_{12}(i,k,X_{\tilde{v}}) = x^{(1)}_{\tilde{v};ik}, \tag{4.27}$$

$$s_{13}(i,k,X_{\tilde{v}}) = x^{(1)}_{\tilde{v};ki}. \tag{4.28}$$

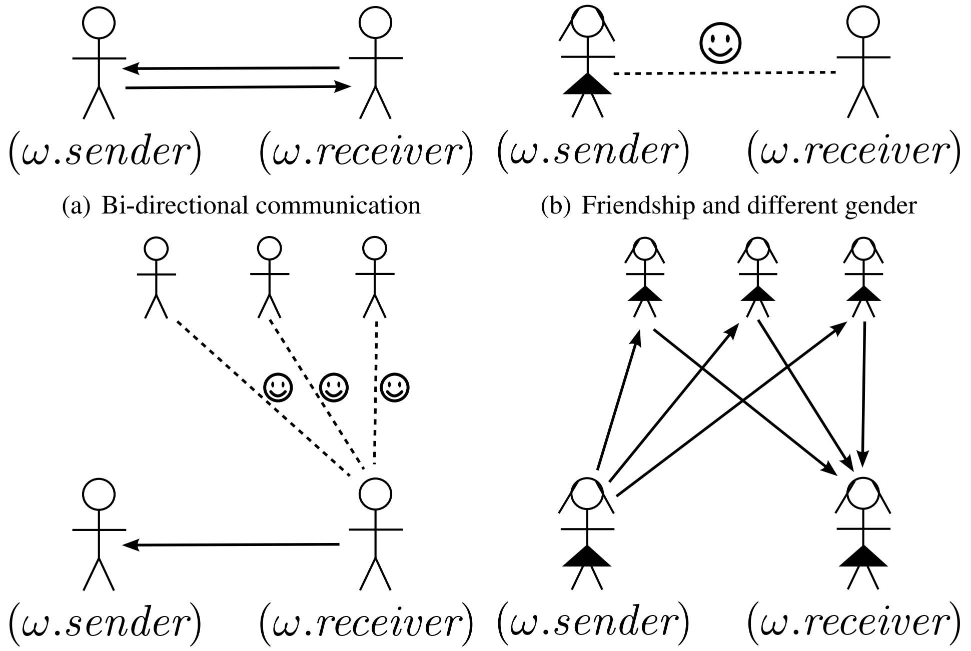
The interpretation of parameters is as follows. Imagine a sender a_i having to choose between two actors a_k and a_l with whom he has communicated before but with a different communication intensity: $x^{(1)}_{\tilde{v};ik} = x^{(1)}_{\tilde{v};il} + d$. Then, the probability of choosing a_k as receiver is $e^{\hat{\beta}_{12} \cdot d}$ times higher (lower). The absolute value of $\hat{\beta}_{12}$ is hard to interpret, but a positive value indicates that stronger ties are preferred over weak ties. The absolute value depends on the observed tie values and, hence, on the concrete formulation of the update rules. A normalization of tie values in the statistic can be considered.

Weighted statistics may incorporate the value of several ties. A combined binary structure with two ties (see section 4.4.6), for example, can be measured by the minimum of the in- and outgoing ties between sender and receiver. Weighted statistics can also be measured on structures in other networks. We briefly discussed the case of time and distance graphs in section 4.4.3. Here, interpretability is a minor issue, as tie values have a unit (kilometers, hours) assigned. However, it may make sense to categorize the tie weights as several dummy variables, if the relation between statistic values and the choice probability is not assumed to be linear. In the application chapter 6 this is done for the statistics of the time and the distance graph.

4.4.6 Combined Statistics

In the last sections we discussed network statistics in different graphs, actor attribute statistics, and weighted graph statistics. In each class we introduced at least the most simple statistics that are measured on a dyadic level between sender and receiver. In some cases we presented extended structures that incorporated other nodes like the endogenous transitivity effect in figure 4.1(e), for example.

Similar to the ideas of dependence graphs in exponential random graph models (see Besag (1974); Frank and Strauss (1986); Robins et al. (2007a)), choice structures are defined for small local environments first. A dyad independence assumption would, for example, lead to a simple model as defined in chapter 3. Simple structures can be extended to more complex models with combined, more complicated structural variables. Some basic model fitting ideas are presented in section 4.5. Figure 4.6 shows four new structures which are combined of basic structures in the previous sections.



(a) Bi-directional communication

(b) Friendship and different gender

(c) Reciprocating communication from actors with many friends

(d) Transitivity in communication of female actors

Figure 4.6: Communication influenced by combined structures

The structures *Repeated communication* and *Reciprocity* in figures 4.1(a) and 4.1(b) can be combined to a new effect that measures the existence of bi-directional communication structures. The corresponding statistic is 1, if and only if both substructures equal 1:

$$s_{14}(i,k,X_{\tilde{v}}) = s_1(i,k,X_{\tilde{v}}) \cdot s_2(i,k,X_{\tilde{v}}). \tag{4.29}$$

The structure in figure 4.6(b) measures whether differences in gender combined with friendship enforce communication event receiver choices. The corresponding statistic is measured as

$$s_{15}(i,k,X_{\tilde{v}}) = x^{(3)}_{\tilde{v};ik}(-s_{10}(i,k,X_{\tilde{v}})+1). \tag{4.30}$$

where $X^{(3)}(t)$ is the friendship graph at time t. The two combined statistics are derived from underlying statistics by multiplying the effect values. Therefore, these statistics can be understood to be *interaction effects* (see, for example, Agresti (2007, p.119–120)). The combined statistic estimate indicates whether the two sub-variables are independent or (often in case of bi-directional communication) whether the co-existence has an additional effect on communication choices.

The structure in figure 4.6(c) can be measured to identify the tendency to reciprocate events of receivers that are popular in the (here undirected) friendship network. The effect of the structure in figure 4.6(d) can be estimated to learn whether transitivity among female actors explains parts of the event receiver choices. The statistics in both cases can be defined similarly to equations 4.29 and 4.30. There are many more possible combinations of the proposed receiver-independent choice variables. These structures can be specified based on theory and the research questions of interest. Another approach is to systematically explore the space of combined statistics. This idea is briefly discussed in section 4.5.

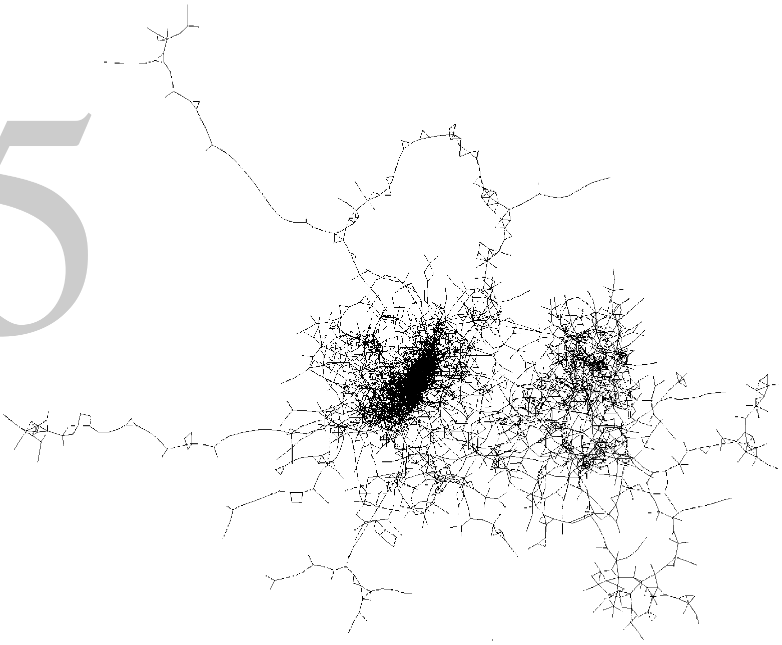
4.5 Exploring the Space of Combined Statistics

In section 4.4 we introduced structures that may potentially influence the choice of event receivers. We now briefly discuss how the *most relevant* of these structures can be identified.

A concrete receiver choice model is defined by a set of independent variables. If a complex variable (like one of those in section 4.4.6) is part of a model, it is advisable to test sub-structures as well. The structure *Bi-directional communication* in figure 4.6(a), for example, is an interaction of the basic structures in figures 4.1(a) and 4.1(b). When fitting a model, the basic structures should be tested separately first. The combined structure should be included later. The complex structure may not be needed for a good model, because the underlying structures already explain enough. Basic structures are usually easier to interpret.

In the application chapter 6 a simple model fitting algorithm is applied: In a first phase a set of *basic* structures is defined (*Repeated communication*, *Reciprocity*, *Time since the last communication*, *Spatial distance*, ...). Then, the statistics are stepwise included into the model by the additional improvement of the log-likelihood function (forward selection) as explained in Agresti (2007, p.139ff). The algorithm starts with a null model without any parameters. In each iteration step, also all possible interaction effects of the already included parameters are evaluated. If the best interaction effect explains more than the best basic effect, it is included next. Multiple interactions are possible. If the removal of a structure leads to a better log-likelihood improvement, this backward elimination is applied instead of an inclusion. Akaike's information criterion (AIC) (Akaike, 1974) can be used as s stop criterion. In some cases, it is better to use a corrected AIC (AIC$_C$) as introduced in Hurvich and Tsai (1989, p.300). In the applied chapter 6 (see section 6.4.2) these first model fitting ideas are applied and discussed in more detail.

5 Application I: Private Communication in a Question and Answer Community



This chapter is published in the journal *Social Networks* (Stadtfeld and Geyer-Schulz, 2011). An earlier version is published online (Stadtfeld and Geyer-Schulz, 2010).

5.1 Introduction

Question and answer (Q&A) communities (like *Yahoo! Answers*, *CosmIQ*, *Answers.com*) have become very popular in the web. People can easily (often even without registration) pose arbitrary questions. Members of these communities try to answer these questions quickly. Often, the only obvious incentives to answer questions are virtual points given to people who answer many questions. The more points someone has, the higher is his/her virtual ranking (e.g. ranging from *Newbie* to *Albert Einstein*). But are there any other effects that make people stay in these groups? Are there, for example, community structures that can be revealed when looking at how actors write private messages to others? Or is most of this private communication just functional and related to questions, like to provide further explanations or to say thank you if someone answered a question?

Actor oriented models are a good way to investigate tie changes in social networks dependent on structures in networks. Snijders (2005) introduced a class of models that is usually applied with panel data of binary network snapshots. The emergence of network structures can also be assessed on cross-sectional network data using the class of exponential random graph models (ERGM, see Wasserman and Pattison (1996); Snijders et al. (2006); Robins et al. (2007a)). Here, the view is not actor oriented, but rather a global view on the network data. The standard class of ERGM has been extended so that it can handle multi-mode networks (see Wang et al. (2009)). Beside models for cross-sectional data and panel data there is new research about the analysis of event stream data with dyad oriented models (Butts, 2008; Brandes et al., 2009; Stadtfeld et al., 2010). The increasing availability of event stream data allows to estimate structural models on this type of data as well. Event stream data incorporates a high amount of information that can be exploited. Algorithmic improvements in preprocessing and the estimation of local models make the application of such models feasible for long data streams and big networks.

In this paper, we present and apply a Markov process model framework to understand what drives the dynamics of private messages sent between actors in a Q&A community. Actor decisions about private communication tie formation and updates are conceptually described as a two-level decision process (for technical reasons, a third level is later added to the model). First, actors have a personal activity rate that influences the decision when to write a message at all. In case they decide to write a message, second, actors have to choose a receiver of the private message. This second decision about private message receivers is modeled as a multinomial logit model. This model expresses whether endogenous one-mode communication structures and two-mode affiliation structures have an influence on the choice of receivers in the community dataset at hand. A new java software package called *ESNA* (event based social network analysis, Stadtfeld (2011b)) was developed to estimate the parameters of this model framework.

Section 5.2 introduces the case study, a dataset of a big German speaking Q&A community, and identifies four phases of the community development. The event stream of the case study is explained. Section 5.3 introduces the Markov process model that is used to analyze the data stream of the case study. It is introduced as a a composition of actor-driven choice models. Network update rules and network structures that potentially influence actor decisions are shown. The model is compared to other related models. In section 5.4 parameter estimates for the last three sub-phases of the community development are given as well as information about the estimation algorithms. Section 5.5 discusses the parameter estimates of the multinomial receiver choice sub-model, as those results are especially interesting when trying to understand to whom people write private messages within the analyzed Q&A community. Finally, section 5.6 summarizes and gives an outlook on further research.

5.2 Case Study

A dataset of a big German Q&A community is analyzed to demonstrate the potential of an exploratory analysis with the Markov process framework introduced in section 5.3. Some characteristics of the dataset at hand will be presented in section 5.2.1. The event stream will be explained in section 5.2.2.

5.2.1 The Dataset

The dataset describes a time span from December 2005 to June 2008. Regarding their communication behavior, people in the Q&A community behave very different from other online social communities. First, the total number of members is big, but only a small subset of all actors is "active" at the same time, because a lot of people only pose one question and leave quickly. There are 416,879 activated user accounts, but 329,055 (79%) of them are "light accounts" that are just used to pose questions, but cannot be used to write or receive private messages. This implies that 87,824 (21%) actors are considered as the set of actors in the dataset. Second, the communication within the community is assumed to be influenced by questions. A virtual rank in the community is only based on how often and how good a member answers questions.

There are 946,603 questions in the dataset with 2,996,446 answers. Although the dataset starts in December 2005, private messages are only logged since August 2006. Figure 5.1 shows how the activity concerning questions, answers and private messages changes over time.

The x-axis shows the different months, beginning with December 2005. The dotted line represents the number of answers, the dashed line the number of questions and the solid line the number of private messages sent. The y-axis gives the number of events (1K = 1,000 events). Generally, the activity in the Q&A community increased. From this first visualization, four different phases were identified as shown in figure 5.1.

In phase I there is only little activity in the dataset with a rather low growth rate (from December 2005 to the end of January 2007). The number of private messages is low: In the first months, the number of messages does not exceed a few hundred messages. This first phase of the Q&A-Community is called *Initialization*.

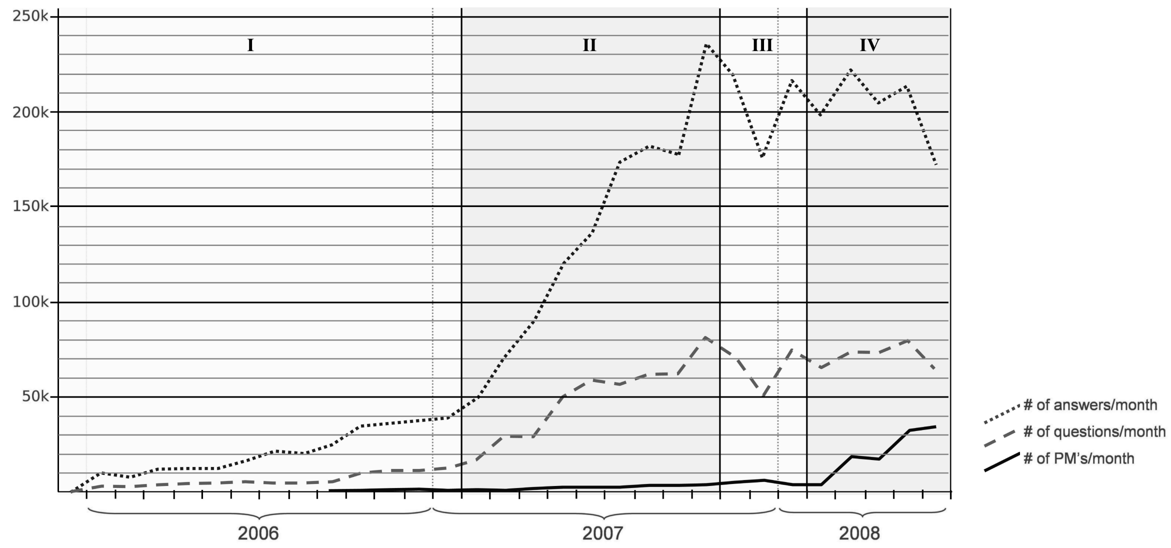


Figure 5.1: Number of three different event types (questions, answers and private messages (=PM's) per month over the whole observed period (1K = 1,000 events). Four phases were identified: I) *Initialization*, II) *Growth*, III) *Stabilization* and IV) *Community Growth*.

Phase II is identified between February 2007 and October 2007 and is characterized by a rapidly increasing amount of questions, answers and a slow increase of the number of private messages. Therefore, it is called *Growth*.

In phase III, the numbers of questions, answers and private messages seem to have reached a more or less stable level. Although there is a lot of variance between the months (probably caused by Christmas time), the total number is always about 65,000 for questions and 210,000 for answers. The number of private messages is stable at a level of about 5,000 per month. Phase III ranges from November 2007 to February 2008 and is called *Stabilization*.

Phase IV is probably the most interesting one regarding the dynamics of private communication, because the number of private messages rapidly increases, while the number of questions and answers is relatively stable. Phase IV ends – like the whole dataset – at the beginning of June 2008. The values for this last month are extrapolated as it was not completely logged. It will be tested, whether community effects are the reason for this sudden and significant increase of messages from an average of about 5,000 per months to more than 30,000 messages in the last completely observed months. This phase is therefore called *Community Growth*. Whether this name is suitable (because the increasing number of messages is based on ”community structures”) will be tested in this paper.

5.2.2 Event Stream

Events are any kind of directed, dyadic interaction between two nodes in a network for which at least a time-stamp is defined. Events may include more information, like an event

time	sender	receiver	type
2007-07-07 14:10:47	Anke	283613	question_opened
2007-07-07 14:10:51	mov_81	283604	answer
2007-07-07 14:11:00	doc-LE	283604	answer
2007-07-07 14:11:16	Snooker01	283600	answer
2007-07-07 14:11:19	mrs_incredible	270053	question_closed
2007-07-07 14:11:31	Nekoy	doc-LE	message
2007-07-07 14:11:42	Anke	283614	question_opened

Table 5.1: Part of the analyzed event stream (with fictitious names)

type or an event intensity. The dataset includes events of different types, that can be used to describe the changes in three different networks. The change in these networks is well defined by the event stream and a set of change rules. As the state of these networks is known for each point in time, this approach processes a lot more information than, for example, models using aggregated panel data.

To analyze the private message dynamics in the database, four different event types were identified and transformed into an event stream with more than 5 million entries. Each of these entries (a row in the resulting database table) describes one event and consists of a time-stamp, a sender, a receiver and an event type. An exemplary snapshot of the event stream is given in table 5.1.

In this dataset, senders are always actors, while receivers may either be actors or questions. The first event type is *question_opened* which indicates that an actor poses a question (which is identified by a unique number). Event type *answer* shows that an actor responds to a question, while *question_closed* indicates that a question is closed either by the question opener, by an administrator, or because the maximum question life time of seven days has been reached. Though the different event streams are not independent, this paper focuses on the dynamics of the fourth event type *message*, which shows that one actor writes a private message to another actor.

5.3 Modeling the Case Study

The decision making of actors regarding private message sending can be modeled as a Markov process with three different levels of decisions. First, this process has to decide, which actor writes a private message. Second, the conceptual decision about the receiver is split into two sub-decision: It is decided whether the possible receiver should be *active* or *non-active*. This "decision" is included for computational reasons since it considerably decreases the size of the multinomial choice model and reflects the fact that many actors leave the community after a short while. If the receiver is of non-active type the chosen

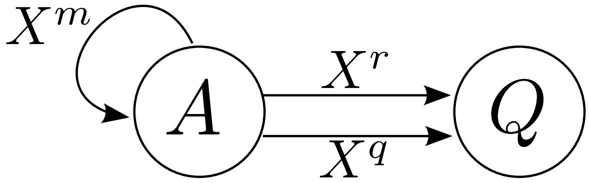


Figure 5.2: Three different graphs are defined on the two node sets A (actors) and Q (questions). The weighted one-mode graph X^m represents recent private communication. X^q and X^r are binary two-mode networks and show which actors have asked questions (X^q) and which actors have responded to them (X^r) on the platform.

sender picks the sender with equal probability. If the receiver is of active type, then the chosen sender decides whom of all *active* actors he or she sends the private message. This decision depends on the network structures that surround sender and receiver. Whether certain structures are relevant for actors' decisions can be tested with a multinomial logit model based on the observed behavior in the dataset. The network structures are a result of all events having been observed in the past and a set of change rules. These rules are briefly explained in section 5.3.1. Section 5.3.2 shows how the regression statistics look like that are tested for influence on the decision process of actors. Section 5.3.3 introduces a heuristic that distinguishes active and non-active actors. In section 5.3.4 the global Markov process is modeled. It consists of many individual decision processes that are explained in section 5.3.5. The econometric evaluation of the model is based on certain assumptions that are listed in section 5.3.6.

5.3.1 Transforming Events into Graphs

A state of the whole process is named x. x is a realization of a random variable X. x is defined by the state of three graphs at a certain point in time. These graphs are defined on two sets of nodes (two modes), which are the set of actors A (with elements $a_1, a_2, \dots$) and the set of questions Q (with elements $q_1, q_2, \dots$). The three graphs describing private messages, question asking and responses to questions are named X^m, X^q and X^r. A realization of x equals (x^m, x^q, x^r). X^m reflects the recent intensity of message writing between actors and has directed, weighted ties in $\mathbb{R}_0^+$. X^q shows which actors have posed a question that has not been closed, yet. X^r is a similar graph that connects actors with active questions they have responded to. The last two of these graphs have directed, binary ties and are bipartite two-mode graphs (affiliation networks). Figure 5.2 shows how the node sets are connected by the three different graphs.

The event stream is an ordered sequence Ω with elements $\omega_1, \omega_2, \dots, \omega_\nu, \dots, \omega_{|\Omega|}$. If the position within the sequence is not of interest, the elements will just be named ω. The variables ω.*time*, ω.*sender*, ω.*receiver* and ω.*type* indicate the four attributes of events as introduced in section 5.2.2 and shown in table 5.1. Depending on these characteristics, events change certain ties of the graphs that define the Markov state X.

		event types			
		question_open	*question_close*	*answer*	*message*
graphs	X^m	–	–	–	tie value +1
	X^q	add tie	remove tie	–	–
	X^r	–	remove ties	add tie	–

Table 5.2: Overview of probabilistic (event triggered) graph change rules

An overview of the used probabilistic (event triggered) change rules is given in table 5.2. If an event of type *question_open* is observed, a new tie from an actor node to a question node is added to the two-mode network X^q. If this question is responded to (event type *answer*), a binary tie is added between the answering actor and the question node in graph X^r. If the question is closed (event type *question_close*), all attached ties are removed from graph X^q and X^r. X^m is a weighted graph (although the later used statistics dichotomize the observed ties). So, if an event of type *message* is observed in the data stream, the corresponding directed network tie from message sender to recipient is increased by 1 ($x_{ij}^{m\prime} = x_{ij}^m + 1$).

But even if no event takes place, the values of ties change due to deterministic, time dependent processes. In this case only one deterministic change rule is applied. The tie values of the private message graph X^m decrease over time with an exponential decay function (see Greiner et al. (1993)). Introducing such a natural decay function seems reasonable, as otherwise the communication intensity between actors could only increase. Even, if there was no communication between actors for very long periods, this value would remain stable and we would still consider the private communication level as being very high.

In an exponential decay function only a parameter *half-life* $t_{1/2}$ needs to be specified. It gives the time after which each tie value decreases by 50%. Ties that have a value $\leq \varepsilon$ are reset to zero. ε is a small value, in this paper it was set to 0.01. The half-life was defined as one week. This is done for computational reasons to reduce the set of *active* actors that are considered to be potential receivers of messages (see section 5.3.3). Note, that a decay plus a threshold value is similar to dichotomizing networks using a threshold in longitudinal models.

Figure 5.3 shows how a directed communication tie (representing the recent private message writing intensity) between two actors changes over time driven by events of type *message* with a_i being the sender and a_j being the receiver. Whenever such an event takes place (at times $t_1, t_2, \ldots, t_7$) the tie value is increased by one. Between the events, the tie value decreases due to the exponential decay.

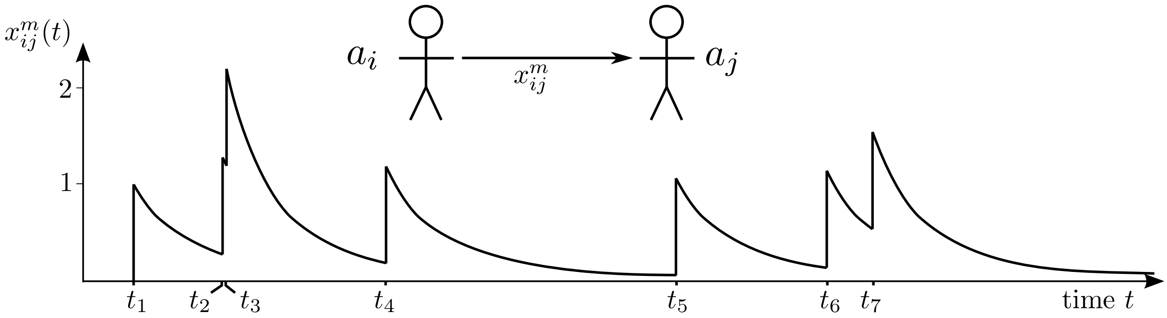


Figure 5.3: Change of a directed private message tie from actor a_i to a_j. At each point in time $t_1,\ldots,t_7$ there is an event ω with $\omega.sender = a_i$ and $\omega.receiver = a_j$. Each event increase the tie value by 1 (probabilistic rule). Between events there is an exponential decay of ties with a fixed half-life (deterministic rule).

5.3.2 Decision Statistics

Ties in social networks do not emerge randomly. Existing network structures are often a good predictor for how people connect with others. These structures (see figures 5.4 and 5.5) can be measured and the resulting decision statistics be understood as independent variables of the receiver choice process modeled by a multinomial logit model (see section 5.3.5). Parameter estimates describing the relevance of these structures can be interpreted similarly to estimates in exponential random graph models (ERGM, see Robins et al. (2007a)) and SIENA models (Snijders, 2005).

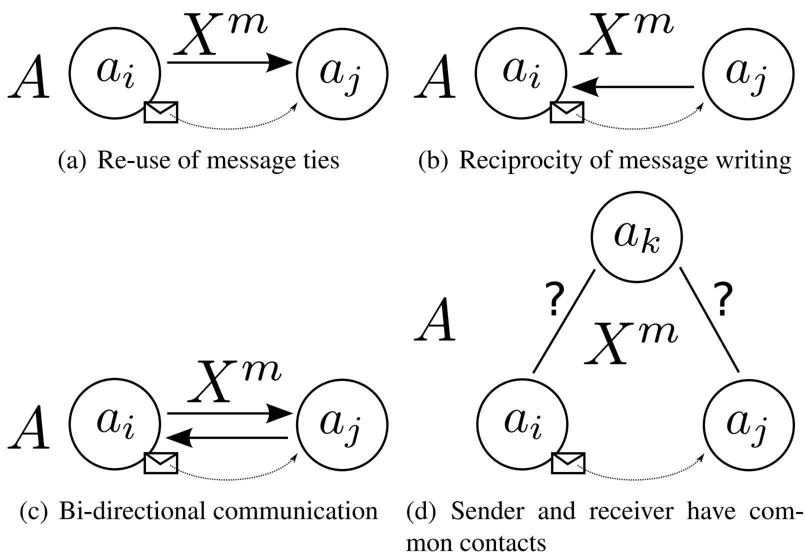
For each decision, this model evaluates the structures in the local environment of the senders and receivers of private messages. Decision statistics are functions $s_d(x,i,j)$ which map the state of the Markov process $x = (x^m, x^q, x^r)$ and the indices i and j of the sender a_i and the receiver a_j into $\mathbb{R}$. Structures can be measured within the communication network itself (endogenous structures) or in other one- or two-mode networks. Structures can in general incorporate actor attributes, multi-network structures or any combination of these (Stadtfeld, 2010).

In this paper, we are interested in whether endogenous (private communication is driven by previous private communication) structures in the message graph and two-mode structures measuring question affiliation influence the choice of event receivers. As mentioned before, only structures in the local environment of sender a_i and receiver a_j in A are evaluated.

Endogenous One-mode Statistics

Figure 5.4 shows four endogenous one-mode structures on the private message graph X^m.

The statistic of the structure in sub-figure 5.4(a) measures whether there is a tendency to re-use message ties, thus to repeatedly communicate with the same actors. Creating new ties is costly and most people have a smaller set of receivers they regularly communicate with, so we assume that this structure will have a positive influence on the probability of choosing the corresponding actor. This effect is only measured binary – the actual weight



(a) Re-use of message ties

(b) Reciprocity of message writing

(c) Bi-directional communication

(d) Sender and receiver have common contacts

Figure 5.4: Four endogenous one-mode network structures that might influence private message receiver decisions. Actor a_i is the sender, a_j the receiver.

of the tie has no effect. Formally, it is defined with the function

$$s_1(x,i,j) = \begin{cases} 1 & \text{, if } x_{ij}^m > 0 \\ 0 & \text{, else} \end{cases}. \tag{5.1}$$

The statistic of the structure in sub-figure 5.4(b) measures whether actors tend to reciprocate previous, incoming private communication. A positive estimate for this statistic indicates that people prefer communication partners that have written a private message themselves before. It is measured by

$$s_2(x,i,j) = \begin{cases} 1 & \text{, if } x_{ji}^m > 0 \\ 0 & \text{, else} \end{cases}. \tag{5.2}$$

The third of the endogenous structures is given in figure 5.4(c). It is a combination of the first two statistics and measures whether actors communicate repeatedly. It should, therefore, not be interpreted without regarding the first two statistics. This structure covers all the reciprocated messages from figure 5.4(b) except the very first response to a private message (or similar structures due to a decay of ties). It also covers all re-uses of a tie from figure 5.4(a) except re-uses of ties that are not reciprocated. It is only measured if the first two statistics are measured with a statistic of 1. This structure is, therefore, an interaction of the first two structures. It indicates whether actors tend to communicate bi-directionally with messages going back and forth for a long time, or whether private conversations are rather short. This could be expected in a rather topic oriented online community like a Q&A platform. The statistic is defined as

$$s_3(x,i,j) = \begin{cases} 1 & \text{, if } x_{ij}^m > 0 \wedge x_{ji}^m > 0 \\ 0 & \text{, else} \end{cases}. \tag{5.3}$$

The fourth structure in figure 5.4(d) may reveal whether sender and receiver of the private message are embedded in community-like structures. The statistic counts the number of actors that sender and recipient are both connected to by previous private messages. The private communication with the counted third actors does not have to be bi-directional. So, it includes all types of transitive triangles, circles and their combinations. It is not taken into account whether a_i and a_j are connected on the message graph x^m. For each third actor, this structure is measured as the binary function f_4. The sum of these measurements (in $\mathbb{N}$) is the statistic of this structure.

$$s_4(x,i,j) = \sum_{a_k \in A \setminus \{a_i,a_j\}} f_4(x,i,j,k) \tag{5.4}$$
$$f_4(x,i,j,k) = \begin{cases} 1 & \text{, if } \left(x^m_{ik} > 0 \vee x^m_{ki} > 0\right) \wedge \left(x^m_{jk} > 0 \vee x^m_{kj} > 0\right) \\ 0 & \text{, else} \end{cases}$$

Two-mode Statistics Measuring Question Affiliation

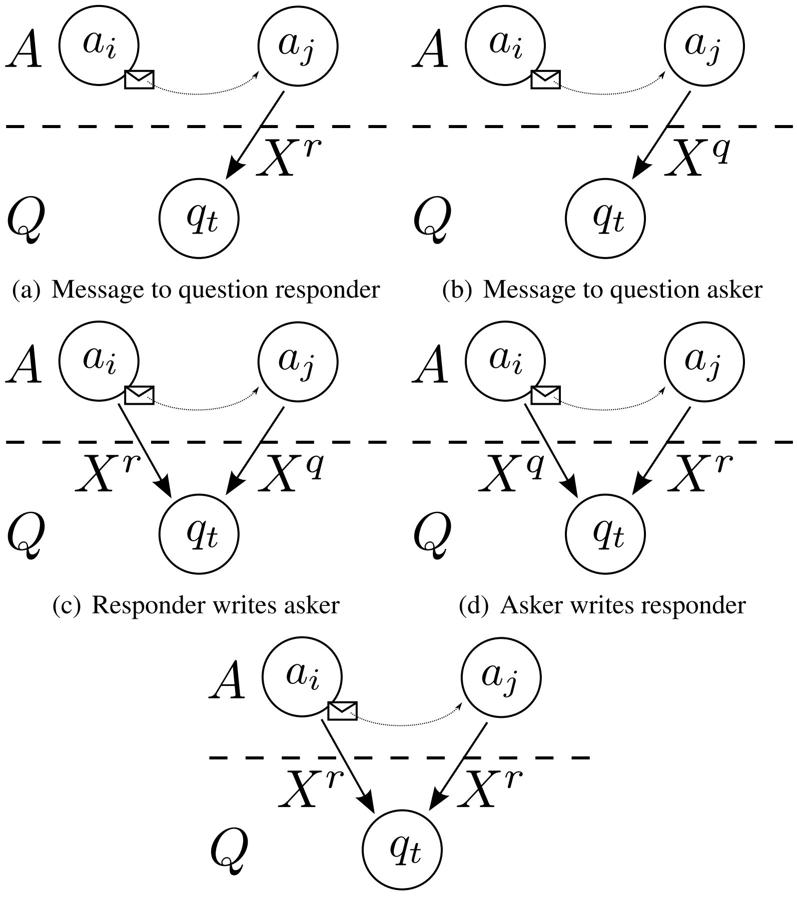
Figure 5.5 shows five structures that are measured to test, whether question affiliation has an effect on private communication. All five structures are two-mode structures with sender a_i and receiver a_j in A and questions from the set of questions Q. Structures in the asker graph X^q and in the responder graph X^r are evaluated.

All five structures measure binary ties. The first two structures (figures 5.5(a) and 5.5(b)) evaluate whether the receiver is connected to questions. Statistic $s_5(x,i,j)$ measures the tendency to write private messages to question askers, statistic $s_6(x,i,j)$ the tendency to write private messages to question responders:

$$s_5(x,i,j) = \sum_{q_t \in Q} f_5(x,j,t) \tag{5.5}$$
$$f_5(x,j,t) = \begin{cases} 1 & \text{, if } x^r_{jt} = 1 \\ 0 & \text{, else} \end{cases}$$

$$s_6(x,i,j) = \sum_{q_t \in Q} f_6(x,j,t) \tag{5.6}$$
$$f_6(x,j,t) = \begin{cases} 1 & \text{, if } x^q_{jt} = 1 \\ 0 & \text{, else} \end{cases} .$$

The structures in figures 5.5(c), 5.5(d) and 5.5(e) evaluate whether actors tend to write private messages to receivers who are connected to the same questions as the sender. It is differentiated between private messages from responder to asker (statistic $s_7(x,i,j)$), asker to responder ($s_8(x,i,j)$) and between responders of the same questions ($s_9(x,i,j)$). The



(a) Message to question responder

(b) Message to question asker

(c) Responder writes asker

(d) Asker writes responder

(e) Responder writes responder

Figure 5.5: Two-mode structures measuring the affiliation of actors to questions. Actor a_i is the sender, a_j the receiver.

statistics are defined as follows:

$$s_7(x,i,j) = \sum_{q_t \in Q} f_7(x,i,j,t) \tag{5.7}$$
$$f_7(x,i,j,t) = \begin{cases} 1 & \text{, if } \left(x^r_{it} = 1\right) \wedge \left(x^q_{jt} = 1\right) \\ 0 & \text{, else} \end{cases}$$

$$s_8(x,i,j) = \sum_{q_t \in Q} f_8(x,i,j,t) \tag{5.8}$$
$$f_8(x,i,j,t) = \begin{cases} 1 & \text{, if } \left(x^q_{it} = 1\right) \wedge \left(x^r_{jt} = 1\right) \\ 0 & \text{, else} \end{cases}$$

$$s_9(x,i,j) = \sum_{q_t \in Q} f_9(x,i,j,t) \tag{5.9}$$
$$f_9(x,i,j,t) = \begin{cases} 1 & \text{, if } \left(x^r_{it} = 1\right) \wedge \left(x^r_{jt} = 1\right) \\ 0 & \text{, else} \end{cases} .$$

Note, that more complex structures could be tested as well. Also, it is possible to combine binary structures with weighted structures. More information is given in Stadtfeld (2010).

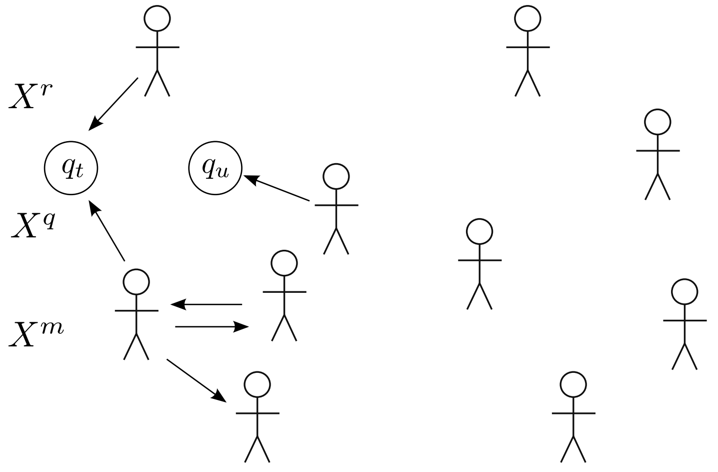
5.3.3 Active Actors

A lot of accounts on the Q&A platform are only used for short time spans, e.g. just to pose one question. Therefore, active actors and non-active actors are distinguished. The set of actors A is split into the subsets A^+, the set of active, and A^-, the set of non-active actors with $A^+ \cup A^- = A$. These sets vary over time. We use a simple heuristic to define the set of active actors based on the three graphs X^m, X^q and X^r. Active actors are those that are connected to a non-closed question (as asker or responder) or have at least one in- or outgoing message tie with a value > 0. Formally, for all actors $a_i \in A^+$ the following condition holds for each point in time:

$$\begin{aligned} a_i \in A^+ \Leftrightarrow & (\exists q_k \in Q : x^q_{ik} = 1 \vee x^r_{ik} = 1) \\ & \vee (\exists a_j \in A : x^m_{ij} > 0 \vee x^m_{ji} > 0) \end{aligned} \tag{5.10}$$

A graphical representation of the idea is shown in figure 5.6.

This heuristic reduces the computational complexity of the estimation significantly as A^+ is much smaller than A and in the model only actors in A^+ are considered as potential receivers of a private message (with high probability). The development of the size of set A^+ over time and an evaluation of the precision of the heuristic are given in section 5.4.2.



(a) Active actors (A^+) (b) Inactive actors (A^-)

Figure 5.6: It is distinguished between active actors in A^+ and non-active actors in A^-.

5.3.4 Markov Process

The proposed Markov process models the occurrence of private message events in the event stream (see table 5.1) from a global perspective. It is a composition of three individual choice models that are explained in section 5.3.5.

Let $\{X(t)|t \geq 0\}$ with state space $\mathscr{X}$ be a Markov process (a continuous-time Markov chain) with right-continuous realizations. The state space $\mathscr{X}$ is defined by all combinations of the possible states of the three graphs X^m (private message graph), X^q (posed questions) and X^r (responses to questions) – see figure 5.2. The Markov process describes updates of ties in X^m due to the occurrence of private message events in the event stream (see table 5.1) as state changes. Formally, it holds that the state space is

$$\mathscr{X} = \left\{(x^m, x^q, x^r) : x^m \in (\mathbb{R}_0^+)^{n \times n}; x^q, x^r \in \{0,1\}^{n \times m}\right\} \tag{5.11}$$

where n is the size of the set of actors A, m the number of questions in set Q and the small x denote concrete realizations of random variables X. Recall that X^m describes a weighted graph on the set of actors, while X^q and X^r are binary and bipartite graphs with ties connecting actors and questions as explained in section 5.3.1.

A Markov process (or continuous time Markov chain) is a process ”without memory” which means that all relevant information for the next process change is represented by the current state. Therefore, the emergence of new ties in the event stream is assumed to depend only on current network structures in the three graphs and a set of constant parameters. In this case, it does not matter by which sequence of events the graphs actually evolved. However, the state space can be extended to model this fact.

For two subsequent private message events $\omega_\nu = (i_\nu, j_\nu, t_\nu)$, $\omega_{\nu+1} = (i_{\nu+1}, j_{\nu+1}, t_{\nu+1})$

the Markov property holds:

$$\begin{aligned}&P\big(X(t_{\nu+1}) = x_{\nu+1}|X(t_\nu) = x_\nu, X(t_{\nu-1}) = x_{\nu-1}, \ldots, X(t_0) = x_0\big)\\ =&P\big(X(t_{\nu+1}) = x_{\nu+1}|X(t_\nu) = x_\nu\big)\end{aligned} \tag{5.12}$$

For each possible message event from a sender $a_i \in A$ to a receiver $a_j \in A$ a “tendency“ for its occurrence is defined as a Poisson process with a rate λ_{ij} (explained in equation 5.13). This is similar to the statement that the time between two consecutive messages from a_i to a_j is λ_{ij}-exponentially distributed.

These rates vary with the sending and receiving actors. λ_{ij} can be understood as the propensity of actor a_i to write a private message to a_j. It depends, first, on the general activity of a_i and, second, on the network structures that a_i, a_j and all other potential event receivers are embedded in. Based on these structures, a_i is assumed to make a choice about the receiver. A rather artificial third additional decision on the type of actor is included: It is distinguished between the two cases that the event receiver may be active or non-active at the time of the event. We understand all three decision levels of the Markov process transition rates as driven by individual choices of the sending actor.

5.3.5 The Individual Choice Model

The transition rates of the Markov process are based on individual choices. They are described by a Poisson parameter $\lambda_{ij}(x;\rho_i,\beta,p^+)$ which models the decision of actor a_i to write a message and to choose actor a_j as the receiver, given the process state x and a set of stable parameters (ρ_i,β,p^+). The process state is only stable for short time intervals, as several not explicitly modeled processes change it: the exponential decay, new questions and answers in the community, and the closing of questions. Therefore, the transition rate is defined as an approximation:

$$\lambda_{ij}(x;\rho_i,\beta,p^+) \approx \begin{cases} \rho_i p^?_{ij}(x;\beta)p^+ & ,\text{if } a_j \in A^+ \textbf{ (i)} \\ \rho_i \frac{1}{|A^-|}(1-p^+) & ,\text{if } a_j \in A^- \textbf{ (ii)} \end{cases} \tag{5.13}$$

with ρ_i the parameter which describes the general activity level of actor a_i, p^+ denoting the probability $P(\omega.receiver \in A^+)$, and $p^?_{ij}$ a multinomial logit model describing the choice of receivers and explained in equation 5.14. It depends on x and a weight vector β. The set A^+ changes with the state of the Markov process and can directly be derived from x as explained in equation 5.10. The rationale for the parameters is explained below.

ρ_i is a parameter of a Poisson process. It describes the general activity of an actor a_i regarding the sending of private messages. The parameters ρ_i of this process can be interpreted as the expected number of messages sent by actor a_i in a defined time span. In equation 5.13, the Poisson rate ρ_i is split into sub-Poisson rates of independent sub-processes in two different ways:

(i) For receivers in the set of active actors A^+, ρ_i is split by multiplying with the probability $p^?_{ij}$ from a multinomial logit model which describes a_i’s choice of a receiver from this set. This case is weighted with p^+.

(ii) For receivers in the set of non-active actors A^-, ρ_i is split – for reasons of simplicity – into equal sub-rates by multiplying with $\frac{1}{|A^-|}$. This case is weighted with $(1-p^+)$.

p^+: A case distinction is made depending on whether the receiver of the private message is active ($a_j \in A^+$) or not ($a_j \in A^-, A^- = A \backslash A^+$). Probability p^+ is equal to $P(\omega.receiver \in A^+)$. p^+ is assumed to be considerably higher than $1-p^+$. So, most receivers of private messages are actually active. For all other receivers $\in A^-$ the probability of a selection is just equally distributed. The probability p^+ is a Bernoulli probability and is assumed to be independent from the size of the set of active actors A^+ that considerably changes over time (see results in figure 5.8). For any period in the dataset it is computed by the fraction of the number of messages sent to active actors to the number of all messages.

$p^?_{ij}$: In case of an active receiver, the probability for choosing a specific receiver $a_j \in A^+$ depends on the network structures sender a_i and receiver a_j are embedded in. $p^?_{ij}(x;\beta)$ is a multinomial logit model (see McFadden (1974), Cramer (2003, p. 107–108), Hosmer and Lemeshow (2000, p. 260–263)) on the set of all active receivers in A^+.

$$p^?_{ij}(x;\beta) = \frac{1}{c^+} \exp\left(\beta^T s(x,i,j)\right) \tag{5.14}$$

with

$$c^+ = \sum_{a_k \in A^+} \exp\left(\beta^T s(x,i,k)\right) \tag{5.15}$$

$s(x,i,j)$ is a vector of network statistics that includes statistic functions like those in section 5.3.2 (equation 5.1 – 5.9):

$$s(x,i,j) = \begin{pmatrix} s_1(x,i,j) \\ s_2(x,i,j) \\ \vdots \end{pmatrix} \tag{5.16}$$

Each statistic $s_d(x,i,j)$ in vector $s(x,i,j)$ is weighted with a corresponding parameter $\beta_d \in \mathbb{R}$. The vector β is unknown but can be calculated using a maximum likelihood estimation. β and $s(x,i,j)$ have the same dimension. The linear function $\beta^T s(x,i,k)$ describing a possible decision is transformed with an exponential function. The resulting value, giving a "weight" for the structures surrounding the sender and the observed receiver is normalized with the characteristics of all those weights that might have occurred, given that a_i decided to write the message to any active actor $a_k \in A^+$. This assures that $p^?_{ij}(x;\beta)$ is a proper probability distribution. The denominator c^+ ("+" indicates that only *active* actors are evaluated) is given in equation 5.15. A^+ directly follows from the process state x and the heuristic in equation 5.10.

5.3.6 Assumptions for Estimating the Individual Choice Models

There are three important assumptions connected to the formulation of this stochastic process as a Markov process:

1) *No phase transitions in analyzed windows*: From figure 5.1 it follows that there are (at least four) different phases with specific characteristics. The characteristics of a phase should have some influence on the emergence of private message ties and should, therefore, be part of the Markov process state. However, we assume that within a shorter analyzed time window certain parameters are constant and therefore do not need to be encoded as part of the process states. This holds for the individual activity of actors, the probability to write messages to active actors and also for structural effects determining the choice of event receivers. We assume that there is no phase transition within one of the analyzed periods of the Markov process.

2) *Local homogeneity*: The Markov process is assumed to be homogeneous (to be independent from the concrete point in time t) within an analyzed period as long as it is small enough. This is reasonable at least within phases II to IV (see figure 5.1) as the distribution of possible states would only marginally depend on the initial state with three empty graphs. Growth processes within a phase are not further considered. We assume different behavior patterns between the different phases II to IV and, in addition, homogeneous behavior in each "small" period analyzed.

3) *Stability in short time spans*: A concrete process state $x = (x^m, x^q, x^r)$ is influenced by a decay of message ties of graph x^m and also by events of other type that change the graphs x^q and x^r. We define the Markov process transition rates only for very short time spans so that we can assume an only marginal relevance of those factors. As the general activity of private message writing is high we consider this a reasonable assumption.

5.4 Estimation and Results

The proposed event model is an actor-driven three-level decision process (see equation 5.13). First, the general actor activity is modeled with the Poisson parameter ρ_i. Second, the probability for choosing an active actor instead of an unconnected actor is given by the probability p^+. Third, if an active actor is chosen, the multinomial choice is given by the probability $p^?_{ij}(x;\beta)$. Results of the parameter estimates are given for each level separately.

As the four phases of the Q&A community seem to have different characteristics (see figure 5.1), a subsequence of each phase II to IV was evaluated separately. The whole event stream has more than five million events including 120,000 private message events. Therefore, it is already sufficient to analyze smaller samples of the stream to get statistically significant results. Also, the process assumes a homogeneity in behavior within shorter time spans as mentioned in section 5.3. This makes it reasonable to look at smaller, stable subsets of the stream within three of the four phases. Subsequences of two days to two weeks were chosen for estimation that included enough message events to get stable results. Analyzing bigger sub-streams is possible though, as memory complexity and computational complexity only increase linearly with the number of estimated events for each simulation iteration step (due to a preprocessing of network statistics). Some characteristics of the three windows chosen are provided in table 5.3.

	phase II	**phase III**	**phase IV**
windows starts	August 1, 2007	December 3, 2007	March 10, 2008
window ends	August 14, 2007	December 9, 2007	March 11, 2008
length	14 days	7 days	2 days
messages	1,465	1,323	1,227
sending actors	288	297	217
avg. size of A^+	8,710.96	6,725.97	10,102.90

Table 5.3: Three sub-streams of phases II to IV were selected.

5.4.1 Estimated Actor Activities ρ_i

The Poisson rates ρ_i are given separately for each of the defined phases II to IV. Only those actors that wrote a message within such a time span are considered. The estimates were calculated with a maximum likelihood estimation. The time unit is days. The average individual Poisson rates $\overline{\rho_i}$ are presented in table 5.4. On average, the actors that send messages in phase II at all wrote 0.372 messages per day. In phase III this rate almost doubled to 0.657 messages per day. In phase IV actors wrote (on average) 2.903 messages per day if they wrote private messages at all. In this phase, we observed a big growth of private messages in the community. Partly, this is reflected by a higher average activity of actors.

phase	$\overline{\rho_i}$	(s.e.)
II	0.372	(0.036)
III	0.657	(0.047)
IV	2.903	(0.116)

Table 5.4: Average Poisson rates of all actors that wrote at least one message in one of the phases II to IV (see figure 5.1) with standard errors of estimates.

In section 5.3 the Poisson rates ρ_i were defined as individual parameters of actors a_i. Plots of the individual actor activities in the selected windows are shown in figure 5.7. The x-axis shows individual actor rates, the y-axis gives the fraction of actors. Both axes are logarithmic. The different phases are indicated by three different symbols. It can be seen that the frequency distribution of parameters seems similar (many actors have a low rate and only few actors have very high rates, the logarithm of the curve is almost linear), but the absolute values are significantly higher in the later phases. In phase IV almost 7% of all

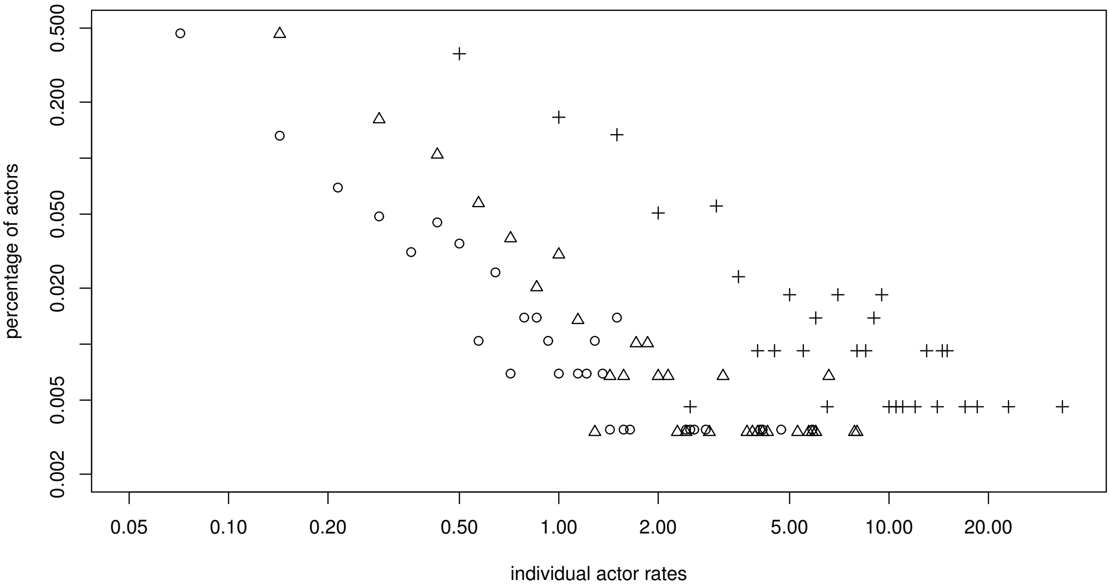


Figure 5.7: Relative number of individual Poisson rates in the sub-windows of phase II (○), phase III (△), and phase IV (+). Both scales are logarithmic.

actors had an activity rate in the range from 10 to 30 expected messages per day. In phase II the highest observed rate was 5.86 only. In phase III the maximum was 7.86 only.

5.4.2 Probability for Choosing Active Actors p^+

The average probability p^+ of choosing an active actor in A^+ over a non-active actor in A^- was 96.27% in the observed event stream. Of 112,811 messages in the completely logged months from September 2006 to May 2008 only 4,208 were observed to be sent to inactive actors, which shows that our heuristic (see equation 5.10) worked quite well.

In phases II, III and IV the observed probability p^+ had values between 96.92% and 97.67%. The estimates of this Bernoulli probability are shown in table 5.5.

phase	all messages	messages to any $a_j \in A^+$	p^+
II	1,500	1,465	97.67%
III	1,365	1,323	96.92%
IV	1,260	1,227	97.38%

Table 5.5: Observed probabilities of p^+ in the sub-streams of phases II to IV.

The average number of active actors $|A^+|$ per month is shown in figure 5.8. The x-axis indicates the beginning of the months. As in figure 5.1, the different phases are highlighted with different shades of gray. In phase I less than 2,000 active actors were in the network.

The number increased to an average of almost $10,000$ active actors at the end of phase II. We showed in figure 5.1 that also the number of questions and answers increased significantly in this time span. In phase III, the number of active actors varied between $6,000$ and $10,000$. In December we observed a much lower activity on the platform, probably due to Christmas break. In phase IV, the number of active actors slightly increased with the increasing number of private messages and goes up to about $12,000$ in the last two months. The set of actors A is constant and has a size of 87,824.

Figure 5.9 shows how the parameter p^+ changes over time. As in figure 5.8 the x-axis indicates the months between September 2006 to June 2008. The y-axis represents the percentage of messages sent to active actors (the monthly average of p^+). Except from low values at the end of phase I and the beginning of phase II the value is rather stable and always higher than 92.5%. In phase IV which includes most of the messages, the percentage of messages written to active actors is at a even higher level of about 97.5%.

5.4.3 Choice of Event Receivers $p^?$

The best fitting parameters $\hat{\beta}$ of probability $p^?_{ij}(x;\beta)$ (see equation 5.14) are calculated by applying a maximum likelihood (ML) estimation.

$$\max_{\beta} \log L = \sum_{v=1}^{|\Omega|} \log p^?_{i_v j_v}(x_v;\beta) \tag{5.17}$$

Ω is the event stream with ordered events $\omega_1,\ldots,\omega_v,\ldots\omega_{|\Omega|}$. i_v and j_v are the indices of actors $a_{i_v} = \omega_v.sender$ and $a_{j_v} = \omega_v.receiver$. The event triggered changes have not yet been applied on x_v.

For each event $\omega_v \in \Omega$ the decision probabilities $p^?_{i_v j_v}(x_v;\beta)$ were assumed to depend only on the network structures at that time (being conditionally independent given x_v). We assumed that those structures have a stable stationary distribution at least for shorter time windows within the event stream and are not significantly influenced by previous events taking place in other environments.

Standard errors of parameter estimates were estimated using a bootstrapping approach with a sample size of 50 (Efron and Tibshirani, 1986).

The log likelihood function in equation 5.17 is concave and can therefore be estimated using a Newton-Raphson algorithm (see Deuflhard (2004)). The software for network statistics preprocessing and estimation was developed in Java, partly using software packages from Apache Commons [1].

The estimation results for the sub-windows in phases II to IV are shown in tables 5.6, 5.7 and 5.8. The figure references and the names of the statistics are given in the first column. Nine different models were tested. The first model only includes the one parameter that improved the log likelihood most compared to a random decision model. This base line model

[1] http://commons.apache.org/

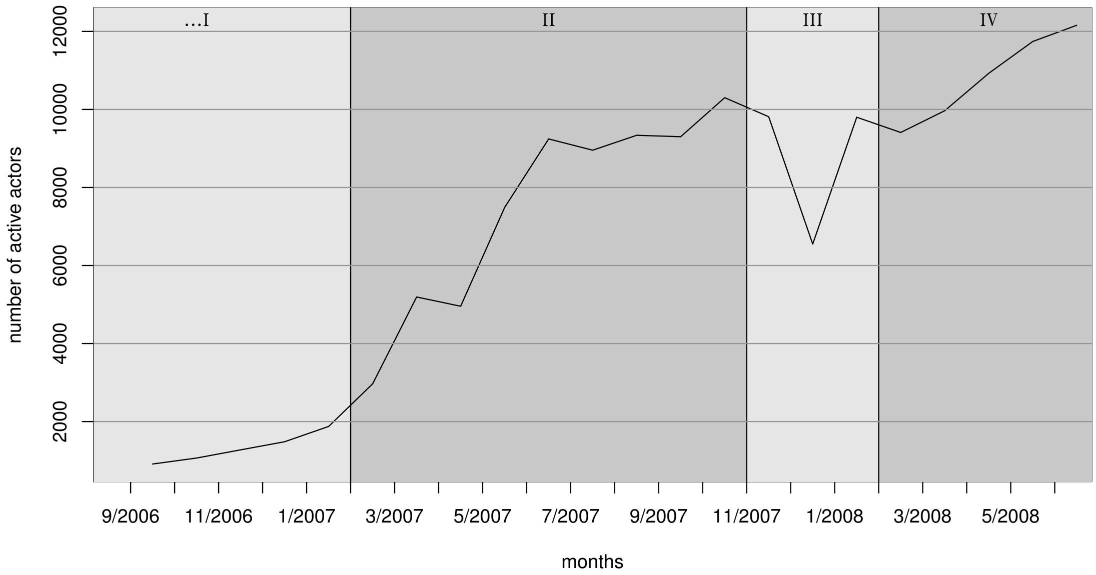


Figure 5.8: Number of active actors in the analysis over time. The plotted values are average values of a month. Phases I to IV are indicated by the background color. A tick on the x-axis indicates the first day of a month. The total number of actors is $|A| = 87,824$.

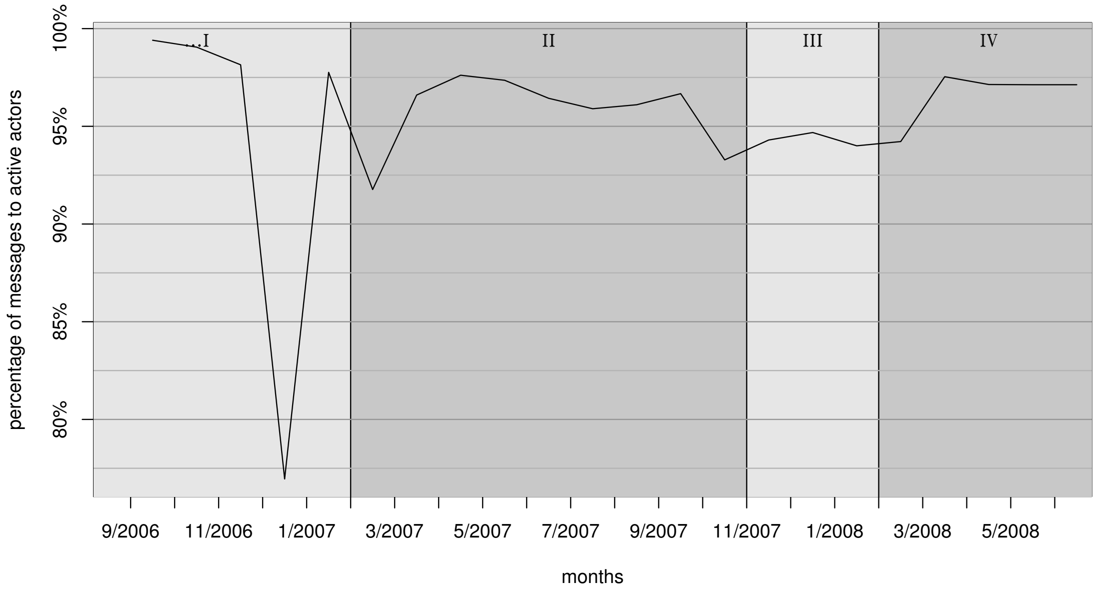


Figure 5.9: Percentage of messages sent to active actors (p^+) per month. Phases I to IV are indicated by the background color. A tick on the x-axis indicates the first day of a month.

includes no parameters and returns a probability of $\frac{1}{A_v^+}$ (uniform probability distribution over all potential receivers) for each event ω_v in the window. The log likelihood of these random decision models is given in the captions of the tables. The additional eight parameters were included stepwise by the additional improvement of the log likelihood (forward selection, see Miller (2002)). The log likelihood of a model is shown in the first row under the model together with Akaike's information criterion (AIC, see Akaike (1974)). In the second row, for each model the deltas of the log likelihood ($\Delta \log L$) and the AIC (Δ AIC) are given. The values are compared to the model with the highest log likelihood and the model with the lowest AIC. In all three tables most parameters are significant with a level of $p < 0.001$. Less significant parameters are italicized, non-significant parameters are marked with an "x". More details are given in the captions of the three tables 5.6, 5.7 and 5.8. The best model regarding the AIC is highlighted by a gray background. This means that all subsequent models with more parameters (and a higher log likelihood) did not further reduce the AIC. The best model has the minimum AIC which is defined by $\frac{-2\log L}{n} + \frac{2k}{n}$ with k is the number of parameters and n the number of private messages in the observed window (see table 5.3). The AIC values are given in the same row as the model log likelihood.

For all models the fit seems to be quite good. Compared with the log likelihood of the random decision reference models (II: -13,290.974; III: -11,660.567; IV: -11,313.649) all models in tables 5.6, 5.7 and 5.8 have a considerably higher log likelihood (in the range from -3,993.563 to -6,111.000). This indicates that the models made a contribution towards explaining the receiver choices of the private communication behavior. The estimates are discussed in section 5.5.

5.5 Discussion of Receiver Choice Parameters

In all sub-windows of phases II to IV we discovered both significant endogenous effects and significant question affiliation effects. Most statistics turned out to be significant on a very high level. In total, only three statistics were not included in a best fitting model. The most important effect in all phases turned out to be the tendency of actors to re-use ties. Reciprocity and bi-directional communication were always relevant but seemed to be more common in the last analyzed phase. Common contacts also increased the probability for communication. The two-mode structures helped a lot to explain the model variance as well. In the first two analyzed phases, question affiliation structures were even the second and third most important independent variables contributing to model fit.

In the following, we discuss the findings for endogenous one-mode statistics and two-mode statistics measuring question affiliation separately.

5.5.1 Endogenous One-mode Statistics

In all three windows we observed highly significant estimates of the three *dyadic*, endogenous one-mode statistics: *Re-use of ties* and *Reciprocity* were always positive, while *Bi-directional communication* was negative. It is interesting to see that *Re-use of ties* always

figure:	**name**	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
		Model II-1		**Model II-2**		**Model II-3**	
5.4(a):	Re-use of ties	9.071	0.113	8.711	0.123	8.598	0.142
5.5(a):	Message to responder			0.008	7.E-4	*0.008*	*6.E-4*
5.5(d):	Asker writes responder					0.275	0.080
$\log L$ / AIC		-6,111.000/8.344		-5,862.794/8.007		-5,822.067/7.952	
$\Delta \log L$ / Δ AIC		-430.774/+0.579		-182.568/+0.242		-141.834/+0.187	
		Model II-4		**Model II-5**		**Model II-6**	
5.4(a):	Re-use of ties	8.508	0.112	8.425	0.159	8.605	0.141
5.5(a):	Message to responder	0.008	6.E-4	0.008	6.E-4	0.007	9.E-4
5.5(d):	Asker writes responder	0.253	0.07	*0.249*	*0.094*	*0.232*	*0.072*
5.4(d):	Common contacts	0.114	0.020	0.102	0.020	0.113	0.020
5.4(b):	Reciprocity			0.663	0.190	4.837	0.429
5.4(c):	Bi-directional comm.					-4.423	0.482
$\log L$ / AIC		-5,787.983/7.907		-5,773.338/7.834		-5,687.623/7.773	
$\Delta \log L$ / Δ AIC		-107.754/+0.142		-93.112/+0.069		-7.397/+0.008	
		Model II-7		**Model II-8**		**Model II-9**	
5.4(a):	Re-use of ties	8.574	0.151	8.567	0.135	8.565	0.154
5.5(a):	Message to responder	0.007	6.E-4	0.007	7.E-4	0.007	7.E-4
5.5(d):	Asker writes responder	*0.224*	*0.077*	0.232	0.070	0.230	0.069
5.4(d):	Common contacts	0.108	0.02	0.108	0.025	0.108	0.024
5.4(b):	Reciprocity	4.860	0.471	4.863	0.521	4.861	0.321
5.4(c):	Bi-directional comm.	-4.448	0.564	-4.435	0.554	-4.433	0.405
5.5(b):	Message to asker	*0.011*	*0.004*	*0.014*	*0.005*	*0.014*	*0.005*
5.5(c):	Responder writes asker			-0.025	0.022x	-0.026	0.028x
5.5(e):	Responder writes responder					0.004	0.012x
$\log L$ / AIC		-5,681.133/7.765		-5,680.348/7.766		-5,680.226/7.767	
$\Delta \log L$ / Δ AIC		-0.907/0.000		-0.122/+0.001		0.000/+0.002	

Table 5.6: Nine models with detailed parameters for a sub-window of Phase II. The last two models were excluded as the additional parameters are insignificant and the additional log likelihood improvement does not reduce the AIC. The random decision log likelihood is -13,290.974. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.01$ only are *italicized.* Parameters with a lower significance level are indicated by a “x”.

figure: name		$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
		Model III-1		**Model III-2**		**Model III-3**	
5.4(a):	Re-use of ties	8.648	0.170	8.203	0.149	8.176	0.156
5.5(e):	Responder writes responder			0.360	0.036	0.281	0.036
5.5(d):	Asker writes responder					0.340	0.049
$\log L$ / AIC		-5,697.398/8.614		-5,308.846/8.028		-5,220.517/7.896	
$\Delta \log L$ / Δ AIC		-721.462/+1.079		-332.910/+0.493		244.581/+0.361	
		Model III-4		**Model III-5**		**Model III-6**	
5.4(a):	Re-use of ties	8.067	0.172	8.220	0.190	8.246	0.201
5.5(e):	Responder writes responder	0.26	0.030	0.256	0.036	0.189	0.033
5.5(d):	Asker writes responder	0.330	0.051	0.340	0.050	0.294	0.039
5.4(b):	Reciprocity	1.477	0.236	5.944	0.414	6.447	0.460
5.4(c):	Bi-directional comm.			-4.729	0.410	-5.464	0.463
5.4(d):	Common contacts					0.065	0.017
$\log L$ / AIC		-5,166.512/7.816		-5,076.189/7.681		-5,018.122/7.595	
$\Delta \log L$ / Δ AIC		-190.576/+0.281		-100.253/+0.146		-42.186/+0.060	
		Model III-7		**Model III-8**		**Model III-9**	
5.4(a):	Re-use of ties	8.221	0.145	8.081	0.165	8.080	0.186
5.5(e):	Responder writes responder	0.166	0.036	0.112	0.038	*0.111*	*0.041*
5.5(d):	Asker writes responder	0.280	0.040	0.293	0.043	0.294	0.042
5.4(b):	Reciprocity	6.411	0.335	6.234	0.335	6.227	0.411
5.4(c):	Bi-directional comm.	-5.348	0.432	-5.174	0.415	-5.156	0.447
5.4(d):	Common contacts	*0.058*	*0.020*	*0.060*	*0.026*	*0.059*	*0.024*
5.5(c):	Responder writes asker	*0.234*	*0.074*	*0.242*	*0.069*	*0.226*	*0.086*
5.5(a):	Message to responder			0.004	0.001	0.004	7.E-4
5.5(b):	Message to asker					0.005	0.007x
$\log L$ / AIC		-4,996.916/7.564		-4,976.391/7.535		-4,975.936/7.536	
$\Delta \log L$ / Δ AIC		-20.980/+0.029		-0.455/0.000		0.000/+0.001	

Table 5.7: Nine models with detailed parameters for a sub-window of Phase III. The last model was excluded as the additional parameter is insignificant and the additional log likelihood improvement does not reduce the AIC. The random decision log likelihood is -11,660.567. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.05$ only are *italicized.* Parameters with a lower significance level are indicated by a "x".

figure: name		$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
		Model IV-1		**Model IV-2**		**Model IV-3**	
5.4(a):	Re-use of ties	8.820	0.112	6.306	0.522	7.967	0.204
5.4(b):	Reciprocity			3.692	0.529	8.380	0.224
5.4(c):	Bi-directional comm.					-5.984	0.276
$\log L$ / AIC		-5,242.867/8.547		-4,592.861/7.490		-4,186.123/6.828	
$\Delta\log L$ / Δ AIC		-1,249.304/+2.032		-599.298/+0.966		-192.560/+0.304	
		Model IV-4		**Model IV-5**		**Model IV-6**	
5.4(a):	Re-use of ties	7.875	0.223	7.665	0.218	7.645	0.181
5.4(b):	Reciprocity	8.213	0.275	8.051	0.218	7.948	0.257
5.4(c):	Bi-directional comm.	-5.877	0.303	-5.72	0.264	-5.672	0.265
5.5(c):	Responder writes asker	0.473	0.059	0.474	0.060	0.332	0.068
5.5(a):	Message to responder			0.004	6.E-4	0.004	7.E-4
5.5(b):	Message to asker					0.033	0.007
$\log L$ / AIC		-4,105.352/6.698		-4,067.848/6.639		-4,046.842/6.606	
$\Delta\log L$ / Δ AIC		-111.789/+0.174		-74.285/+0.115		-53.279/+0.082	
		Model IV-7		**Model IV-8**		**Model IV-9**	
5.4(a):	Re-use of ties	7.580	0.234	7.525	0.197	7.516	0.243
5.4(b):	Reciprocity	7.997	0.249	7.967	0.259	8.015	0.248
5.4(c):	Bi-directional comm.	-5.744	0.292	-5.727	0.289	-5.763	0.293
5.5(c):	Responder writes asker	0.319	0.078	0.300	0.082	0.332	0.084
5.5(a):	Message to responder	0.004	8.E-5	0.003	9.E-4	0.005	9.E-4
5.5(b):	Message to asker	0.034	0.008	0.033	0.005	0.029	0.006
5.4(d):	Common contacts	0.082	0.018	0.082	0.016	0.085	0.019
5.5(d):	Asker writes responder			*0.224*	*0.071*	0.258	0.060
5.5(e):	Responder writes responder					-0.114	0.026
$\log L$ / AIC		-4,028.618/6.578		-4,012.096/6.553		-3,993.563/6.524	
$\Delta\log L$ / Δ AIC		-35.055/+0.054		-18.533/+0.029		0.000/0.000	

Table 5.8: Nine models with detailed parameters for a sub-window of Phase IV. The random decision log likelihood is -11,313.649. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.01$ only are *italicized*.

explained most of all parameters and as soon as *Reciprocity* was included in a model, the statistic *Bi-directional communication* increased the log likelihood significantly as the next included parameter (and increased the significance of the first two). As mentioned in section 5.3.2, the third structure is interpreted as an interaction effect.

The statistics *Re-use of ties* and *Reciprocity* are the two main effects and *Bi-directional communication* models the interaction between the two main effects. Therefore, the interpretation of the third effect has to take the estimates of the first two effects into account. This can best be understood by defining an "equivalent" model with two new statistics replacing *Re-use of ties* and *Reciprocity*:

In this equivalent model, *Re-use of ties* ($s_1(x,i,j)$) can be substituted by an effect that measures the re-use of a tie only if there is no in-coming tie from the sender. This effect is named s'_1. The standard *Reciprocity* effect ($s_2(x,i,j)$) can be replaced with an effect that measures reciprocity only if there is no re-usable communication tie from sender to receiver. This effect is named s'_2. Equations 5.18 to 5.23 show how these effects are formally defined.

$$s_1 := s_1(x,i,j) \tag{5.18}$$
$$s_2 := s_2(x,i,j) \tag{5.19}$$
$$s_3 := s_3(x,i,j) \tag{5.20}$$
$$s'_1 := s_1 - s_3 \tag{5.21}$$
$$s'_2 := s_2 - s_3 \tag{5.22}$$
$$s'_3 := s_3 \tag{5.23}$$

Table 5.9 shows how the different dyadic statistics measure the four possible states of the communication dyad between sender a_i and receiver a_j in the private message communication network x^m.

The four rows show first a complete dyad with ties in both directions, then, second and third, two dyads with just one directed tie either from sender to receiver or from receiver to sender, and, fourth, an empty dyad with no positive communication tie. It can be seen that in a model with the three statistics s'_1, s'_2, s'_3 (the last three columns of table 5.9) the three non-empty structures in the first three rows are measured disjointly. This has an effect on the interpretation of parameter estimates as we will demonstrate in the following.

In table 5.10 we compare the estimates (without s.e.) of model IV-3 from table 5.8 (with statistics s_1, s_2 and s_3) with an equivalent model IV-3' (with statistics s'_1, s'_2 and s'_3) which is based on the "equivalent" definition given above. The results are shown in two sub tables.

In the first variant the interaction effect acts as a correction of the two main effects. In the second variant the three statistics measure disjoint structures (see table 5.9). Therefore, the estimates measure the influence of each dyadic structure on receiver choices separately. The log likelihood of both models is the same.

What we learn from this comparison is that in a model with structures s_1, s_2 and s_3 included (model IV-3, for example) the first two structures can – as in model IV-3' – be interpreted as the influence of non-bi-directional structures on the choice of receivers. This

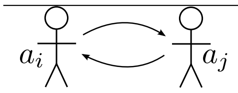
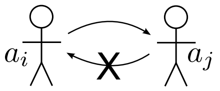
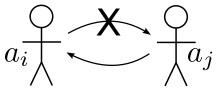
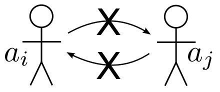
dyad state	s_1	s_2	s_3	s'_1	s'_2	s'_3
	1	1	1	0	0	1
	1	0	0	1	0	0
	0	1	0	0	1	0
	0	0	0	0	0	0

Table 5.9: Overview of the values for the dyadic statistics including the three alternative statistics s'_1, s'_2 and s'_2 ($= s_3$) for each possible state of the dyad between event sender a_i and receiver a_j in the private message communication network x^m. In the figures on the left a crossed arc indicates that this tie is explicitly missing in the measured dyad. All other directed ties have a value > 0.

model **IV-3**	
$\hat{\beta}_1$	7.967
$\hat{\beta}_2$	8.380
$\hat{\beta}_3$	-5.984

model **IV-3'**	
$\hat{\beta}'_1$	7.967
$\hat{\beta}'_2$	8.380
$\hat{\beta}'_3$	10.363

Table 5.10: The two sub tables show estimates of model IV-3 and an equivalent model IV-3' in which the statistics s_1, s_2 and s_3 were replace with statistis s'_1, s'_2 and s'_3 (equals s_3) as introduced in equations 5.18 to 5.23. The sum over all parameters from model IV-3 is 10.363 which equals the parameter $\hat{\beta}'_3$ of statistic s'_3 in model IV-3'.

follows from the fact that the values of the first two parameters in model IV-3 and IV-3' are the same (see table 5.10). The third parameter of model IV-3 can best be understood by interpreting it as a correction of the sum of the first two parameters which is $7.967 + 8.380 - 5.984 = 10.363$. This is exactly the estimate of the complete sender-receiver-dyad on the choice of event receivers in the disjoint alternative model. Therefore, bi-directional pure communication is actually best be understood as enforcing communication choices although the correcting parameter is negative.

Although a model with the equivalent statistics s'_1, s'_2 and s'_3 would have been more straightforward to interpret we chose to use the non-disjoint statistics. The reason is that with no bi-directional communication statistic s_3 or s'_3 included in a model statistics s_1 and s_2 explain more of the overall model (they generate a higher log likelihood). In the following, the interpretation of the absolute dyadic parameter estimates will be discussed.

The absolute values of the three dyadic parameters can be interpreted by comparing the probability for the choice of receivers *with* a certain structure to the choice of receivers without any dyadic structure (the fourth row in table 5.9). The probability of sender a_i for choosing a certain receiver a_j over any other receiver was defined in equation 5.14 as $p^?_{ij}(x;\beta)$. Compared to the base line decision with no dyadic structures it is θ times higher assuming that all other structures influencing the decision are equivalent:

$$\begin{aligned} p^?_{ij}(x;\beta) &= e^{\beta_1 s_1+\beta_2 s_2+\beta_3 s_3} \times \frac{e^{\beta_4 s_4+\ldots}}{c^+} \\ &= \theta\, \overbrace{e^{\beta_1\times 0+\beta_2\times 0+\beta_3\times 0}}^{=1} \times \frac{e^{\beta_4 s_4+\ldots}}{c^+} \\ \Leftrightarrow \theta &= e^{\beta_1 s_1+\beta_2 s_2+\beta_3 s_3} \end{aligned} \tag{5.24}$$

As stated before, people in the dataset tended to re-use ties, they tended to reciprocate and they even more tended to communicate within stable bi-directional communication patterns. In model IV-3, for example, the existence of a re-usable tie that was not part of a bi-directional communication structure (row two in table 5.9) increased the probability times $\theta = e^{7.697} = 2,201.73$ compared to a receiver without any dyadic structure.

If there was an incoming message tie from an actor the sender had no outgoing tie to (as in row three of table 5.9), the probability for choosing this actor was $\theta = e^{8.380} = 4,359.01$ times higher than in the base line model.

If there were both an incoming and an outgoing tie (row one in table 5.9), the probability for choosing such a receiver was $\theta = e^{7.697+8.380-5.984} = e^{10.363} = 21,099.40$ times higher which is more than in one of the dyads with just one private message tie.

For the used dyadic statistics holds: Only if the negative value of parameter *Bi-directional communication* would have had a higher absolute value than one of the two other dyadic structures, we could have inferred that certain "unclosed" dyads were preferred over complete communication dyads. This was not the case in any of the models. This implies that bi-directional communication with repeated message writing in both directions was preferred to short conversations with just one private message written in each direction. Still, repeated message writing to the same receivers without reciprocation and reciprocating incoming message events *once* were very important predictors of receiver choice behavior. We argue that all these observations indicate that private communication in the dataset was not only functional, like to give additional information about questions or to say "thank you" if a question was answered but has a (dyadic) social component.

The absolute values of the dyadic parameters differ between the window. While *Re-use of ties* is similar in the best models of the three phases but decreases slightly (from 8.574 (0.141) to 7.516 (0.243)), *Reciprocity* significantly increases over time. The estimate is 4.860 (0.471) in phase II, grows to 6.234 (0.335) in phase III, and hits 8.015 (0.248) in phase IV. Together with the only slightly decreasing statistic *Bi-directional communication* (from -4.448 (0.564) in phase II to -5.763 (0.293) in phase IV) we conclude that actors tend to communicate bi-directionally more and more: Compared to choosing a potential receiver that is not connected to the sender in the message graph, the probability for choosing a receiver connected bi-directionally was 7,990.43 times higher in phase II, 9,330.09 times higher in phase III and 17,465.80 times higher in phase IV. We found those values by adding all three dyadic endogenous statistics in the best fitting models of each phase.

The observed increase of bi-directional communication is underlined by the fact that the relative importance of these parameters in the models increased. *Reciprocity* and *Bi-directional communication* were only the fifth and sixth statistic included in phase II. In phase III they were included as fourth and fifth statistic. In phase IV, finally, the three dyadic statistics were the three most important variables in the model. This means that dyadic structures were observed (and could be well explained) in an increasing number of cases. However, this change of relative importance could also be related to an interaction effect that we did not measure.

This supports the hypothesis that actors increasingly write messages within closed dyads. Partly, this effect can be explained with the higher rate of written messages per person that prevents a decay of ties. However, this effect alone is probably less significant, as the minimum time before a tie is removed from the dataset is more than six weeks and in most cases there is an earlier tie update if actors communicate bi-directionally. The absolute value of ties was not considered as long as the tie was not removed from graph x^m because it had decayed under a certain threshold. We argue that these observations indicate an "emergence" of social (in contrast to functional) behavior over the life-time of the Q&A community.

The three dyadic structures are the only statistics in this analysis that are rather independent from the state of the process. They are not influenced by varying factors like network density or number of active actors. They are also not strongly affected by the general activity in the dataset. So, we can always compare these parameters to a decision without the corresponding structure and thereby give an interpretation of the absolute values and its changes over time as long as possible interactions with other effects are kept in mind.

This is different with the endogenous statistic *Common contacts*. It was significantly positive (with varying significance levels) in all models. This means that actors tended to communicate with others who they had (many) common contacts with. The absolute parameter is harder to interpret. The probability for choosing a receiver increases in model II-7, for example, with each additional common contact by 11.4% ($e^{0.1805} = 1.114$).

However, in this model we implicitly assume that changes of the probability depend linearly on the number of common contacts. This is probably not the case. So, if we observe different numbers of common contacts in the local environments in different windows of the event stream the absolute value of the estimate is influenced by that fact and is therefore less straightforward to interpret.

The rank of the *Common contacts* statistic decreased from rank 4 in phase II to rank 6 in phase III to rank 7 in phase IV. The reason could be an interaction with the dyadic statistics with increasing rank. The inclusion of this parameter, however, never had a huge effect on the estimates of the dyadic statistics (or any other statistic), so we assume that the interaction is not too strong. In general, however, the existence of common contacts was an important predictor for communication choices. We argue that this indicates a social component in private communication behavior on the platform.

5.5.2 Two-mode Statistics Measuring Question Affiliation

Whenever one of the five two-mode statistics was significant at all it had a positive weight in almost all models. Only three times (in models II-7, II-8 and III-9) structures were insignificant. We observed a tendency for communication with askers or question responders and also a high tendency for communication between actors connected to the same questions in any way. This supports the hypothesis that private communication on the platform was also driven by question affiliation – affiliated receivers were preferred over others, especially if the sender was connected to the same question. The more question affiliations were counted, the higher was the probability for choosing the corresponding receiver.

The only exception is the statistic *Responder writes responder* in the last model IV-9. Here, a negative effect was observed. Once again, the reason could be an interaction with other effects. We observed that in model IV-9 the significance of the estimates of *Message to responder* and *Asker writes responder* increased (compared to model IV-8) when the negative effect *Responder writes responder* was included. This is similar to what happened in case of dyadic, endogenous statistics.

The rank of the two-mode statistics changed a lot between the three analyzed windows. This is not surprising, as these structures are not independent. We cannot say whether the general tendency for certain types of question related private messages increased or decreased over time. All we learned was that in general the endogenous structures explained more in the last phase compared to the included two-mode structures. Still, beside *social* aspects in private communication choices question affiliation as a rather *functional* parameter was important for explaining private communication choices in the analyzed web community.

5.6 Conclusions and Further Research

In this paper the structural dynamics of private message communication in a German speaking Q&A community were analyzed. We introduced the dataset and the event stream and defined four different phases in the community development. A generic actor oriented Markov process model was introduced that can be applied to describe event formation in social environments. To demonstrate the application, we analyzed the sending of private messages within the Q&A community. The model was constructed as a three-level decision process. First, actors were assumed to decide about the time of private message sending based on individual Poisson rates. Second, in case of sending a private message they were assumed to choose whether to send an event to a currently active actors. This decision level was included to take into account that only a smaller subset of all actors was active in the community at the same time and would therefore be considered as a communication partner. The third decision is about the choice of private message receivers. This last decision was modeled as a multinomial logit model. Different endogenous communication structures and two-mode question affiliation structures were included as independent variables. We estimated different models (using a newly developed software package) to learn how these structures influence the decision about receivers of private messages. We also tried to find out whether we could identify differences in the different phases of the community.

It turned out that private communication dynamics in the analyzed community depended on dyadic and triadic endogenous structures in the private message graph, but also on two-mode question affiliation structures of senders and receivers. We found, for example, a high tendency for repeated private communication with the same actors, for bi-directional private communication, for triadic private communication structures and for choosing receivers that are answering or asking questions.

It could be shown that the estimates are slightly different in the different phases of the community and explain a different amount of the overall variance. Dyadic, endogenous effects seemed to get more relevant in the later phases. We learned that private communication in the analyzed community was driven both by social structures and functional aspects.

We could show how the proposed model framework can be applied on big event streams with different types of events and modes. Possible extensions of this framework were mentioned. Network structures, for example, can also incorporate the values of ties, actor attributes or multi-network structures. The multi-level decision process can be extended by decisions about different event types or event intensities. The simple actor activity rate used in this paper can also be parameterized to model, for example, the influence of structures, attributes or time on actor activity. Due to the richness of event stream data, the multinomial decisions could be estimated on an individual level if the research question was targeting individual behavior patterns instead of general group behavior. The current model only describes a small part of the overall dynamics in the Q&A community. It could, for example, be extended to measure co-evolutionary dynamics of private messages, questions and answers.

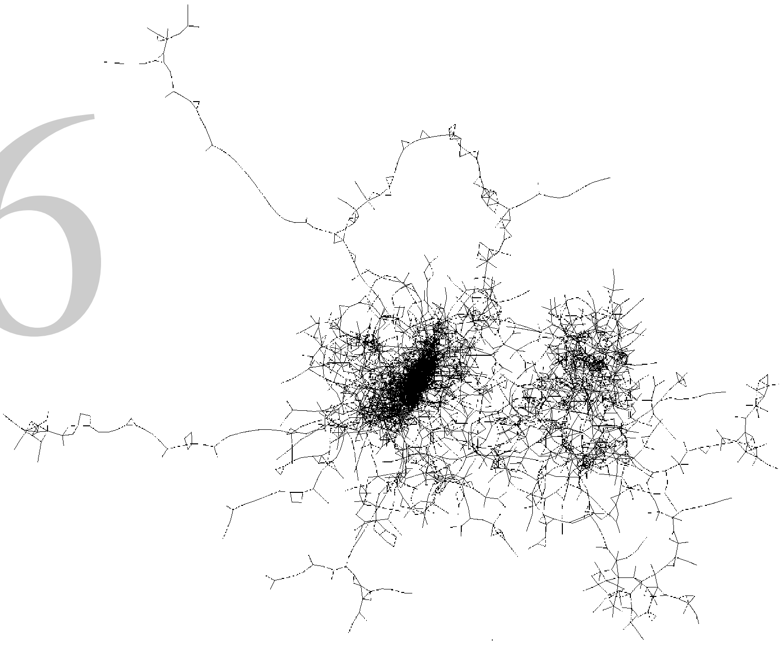
In future work we plan to apply a more structured model fitting algorithm. There is a huge number of possible independent variables in structural network models with many possible interactions. It would be interesting to find an algorithm that uses structural dependencies in the graphs to explore the space of possible network structures in a more sophisticated way.

Furthermore, we want to test the methodology on more, interesting datasets to learn more about robustness, interpretation of results and good model fitting strategies. Some of the mentioned extensions of the model framework may make sense when modeling different event stream datasets.

One advantage of event stream analyses is that the analyzed periods do not have to be predefined by an experimental setting but can be chosen ex post. If it was possible to define standardized structures that do not strongly depend on how the networks look like at a certain point in time, sliding window analyses could be applied to reveal the periods where structural breaks or slow changes in the underlying structural dynamics occur[2].

[2] Some helpful remarks by two anonymous reviewers contributed to this idea.

6 Application II: The Influence of Distance, Time, and Communication Network Structures on the Choice of Communication Partners



This chapter is published in the proceedings of the *Third IEEE Conference on Social Computing 2011* (Stadtfeld et al., 2011).

6.1 Introduction

Phone calls are costly. Individuals must invest money, time, and social effort when communicating with others. Consequently, the individual decisions on when and with whom to communicate are not random. Instead, the communication dynamics observed in a community reflects underlying individual decision processes driven by rational choice. These individual choices can never be understood in their full complexity. Still, researchers may be interested in finding models that explain communication behavior in an observed community as well as possible. An important goal is to identify the main drivers of communication choices.

Recently, actor-oriented decision models for social networks were of increasing interest. New model frameworks for social network data in the form of panel data were developed (Snijders (2005)). An increasing availability of network data and advances in estimation algorithms led to many publications using SAOMs. In earlier publications (Stadtfeld and Geyer-Schulz (2010); Stadtfeld (2010)) it has been demonstrated that these models can be adapted to event data streams (like e-mail or phone logs) and that the underlying models can be estimated efficiently even for very long data streams and big networks. *ESNA* (Stadtfeld, 2011b) is a new java software tool developed for the estimation of multinomial logit models in network structures.

The aim of this paper is to apply this new methodology to the phone call data stream of the MIT Reality Mining project Eagle et al. (2009). It is tested whether certain network structures are important drivers of communication choices. These structures are measured in three graphs describing the recent communication patterns (who communicated with whom in the past?), the time since recent interactions (when did the last interaction take place?), and a heuristic defining the distance between communication partners (in cell tower hops). The major contributions of this paper are to show that structures in all three graphs significantly influence communication choices and that the model selection process reveals the relative importance of these structures. Beside these substantial contributions this paper will extend the previously introduced model framework by a parameterization of the individual activity rates and by investigating multiple network structures as independent variables of the receiver choices.

In section 6.2 we introduce the MIT Reality Mining project, inform about the available data and the graphs of specific interest for actor-oriented models. In section 6.3 we explain the two-level decision model. It is defined as a Markov process that consists of actor Poisson processes determining the general propensity of actors to make phone calls and a multinomial choice model describing the choice of call receivers based on network structures. In section 6.4 we present results for different models. We present the results for both decision levels separately. The estimated parameters of the multinomial choice model are discussed in section 6.5. Section 6.6 summarizes our results and gives an outlook on future research questions.

6.2 Case Study: MIT Reality Mining Data

The MIT Reality Mining project was conducted in the academic year from 2004 to 2005 at the MIT Media Lab Eagle et al. (2009). 94 students and staff members were equipped with cell phones. All communication events using these cell phones (calls and text messages/SMS) were logged. In addition, the data set includes information about bluetooth connections, cell towers that cell phones are connected to and survey data incorporating information about friendship relations.

In this paper, the publicly available data set (`http://reality.media.mit.edu/download.php`, SQL data base from 2009) is used to show how communication choices in big event streams can be explained by network structures. Instead of using all information available in the data set, use is restricted to the communication network, information about the time since the last interaction on a dyadic level, and a heuristic of spatial distances between actors as independent variables to estimate a model describing communication choices.

The database contains 599,097 phone calls. Each entry consists of several fields, including a *timestamp*, a *sender id* (the person who started the call) and a *receiver id* (the person who received the call). Hereinafter, this triplet (timestamp, sender, receiver) will be referred to as an *event*. After filtering out duplicates the table is reduced to 83,747 entries. Then, all phone calls directed to cell phones outside of the set of experiment participants are removed (35,026 events remaining). From these we pick all events for which the cell towers were known, to which sender and receiver were connected at the time of the call (10,689 events remaining). To avoid a bias in our estimates only those participants were included in the data set as a *potential receiver* of a phone call for whom the current cell tower was known as well. By applying this selection it is implicitly assumed that the missing data about cell tower connections are totally random as long as they are not caused by an intentional absence (e.g. cell phone shutdown). Especially at the end of the experiment, the data of many participants are not logged any more which leads to a decreasing number of potential receivers. Events that incorporated at least 20 potential receivers are evaluated only (final size: 10,027 events). The cell tower data (32,656 entries in the data set) are preprocessed as well. After filtering out cell towers with obviously faulty names (like “ERROR“), cell tower entries representing the same cell tower with the same id, but different spelling of the provider names (e.g. “T-Mobile”, T - Mobile“, “TMO”) are identified. This reduced the data set to 20,549 distinct cell towers.

6.2.1 The Three Observed Networks

Choices of communication partners are explained by structures from three different networks that are represented by the graphs X^c, X^t and X^d shown in figure 6.1. They connect actors $a_1, a_2, \ldots$ from the set of actors A and change over time. Together they define one part of the process state as $X = (X^c, X^t, X^d)$. A concrete realization of X at time t is denoted by $x(t)$, a vector of three graphs at time t. Possible statistics of the structures in figure 6.1 are given in equation 6.11.

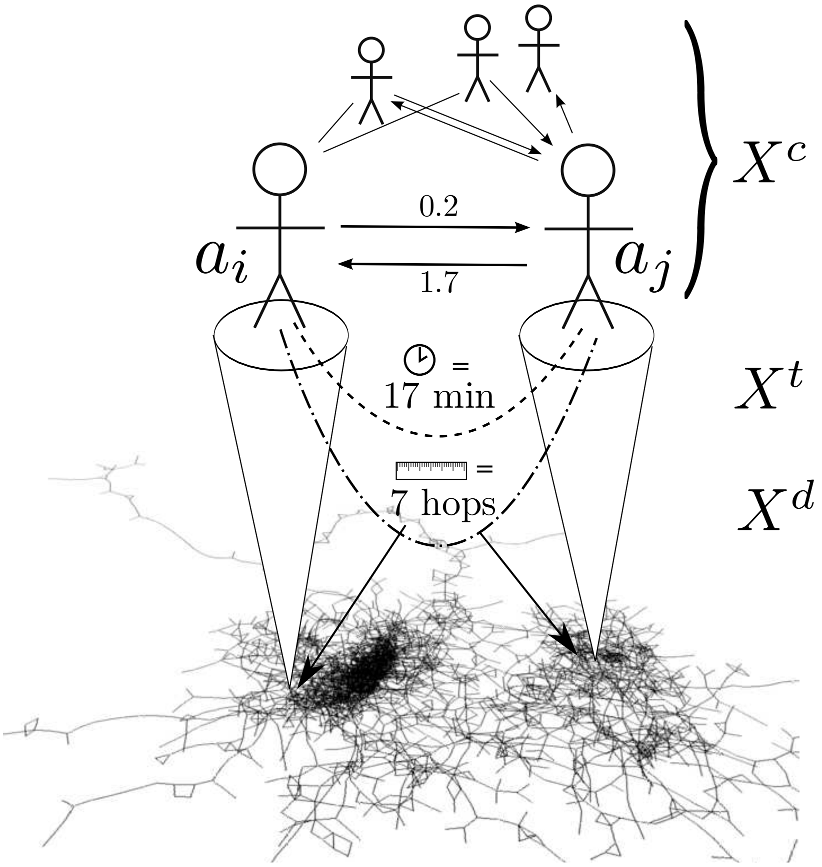


Figure 6.1: Actor a_i's choice of call receiver a_j.

X^c: The *communication graph* X^c represents recent cell phone communication between actors. The graph is directed and weighted. Whenever actor a_i calls actor a_j the communication tie x^c_{ij} is increased by 1. During the time between two phone calls tie values decay exponentially with a half-life of one week. Very weak ties with a value $< \frac{1}{32}$ are removed from the graph (after 5 weeks, if there is no further update). In previous tests on simulated data streams, these model parameters turned out to be robust on changes of these specifications. To define the state of graph X^c at a certain point in time, we applied the changes of all 35,026 phone calls observed among the 94 experiment participants. However, our multinomial receiver choice submodel only evaluates those events for which the cell tower connections of sender and receiver are known.

X^t: The *last communication time graph* X^t measures the time since the last phone call for each pair of actors a_i, a_j. It is undirected ($x^t_{ij} = x^t_{ji}$) and ties are reset to zero whenever a phone call between the corresponding actors takes place.

X^d: The *distance graph* X^d represents an approximate distance between two actors measured in “cell tower hops“. As no precise location information is available in the data set (but only information about cell tower connection of actors) *cell tower neighborship* is defined as an indicator of the immediate distance of two cell towers. Cell towers are neighbors if at least one actor was connected to both cell towers consecutively without any time in between. A partial validation of this structure was possible by comparing user-assigned tags with observed clusters. A spring embedder visualization of the main component of the cell tower neighborship graph is shown in the lower part of figure 6.1. The cluster on the left represents Boston, the one on the right New York.

The distance between two actors is defined by the shortest path between their current cell towers in the cell tower neighborship graph. Hereinafter, this distance this distance is referred to as the number of hops. The actual spatial distance may vary depending on whether the cell towers are in rural or urban areas, for example.

6.3 Model

The model used in this paper is a SAOM for dyadic event data streams as introduced by Stadtfeld and Geyer-Schulz (2010); Stadtfeld (2010). This class of models is an extension and adaptation of the class of SAOMs for panel data introduced by Snijders (2005). The main extension is its applicability to constantly measured event networks in which an update of ties is observed rather than an explicit creation and dissolving of ties. The dissolving process of ties is modeled by an external exponential decay function. Moreover, the evaluation of network structures does not take place on a global level but within the ego-network of the actor who starts a dyadic event (e.g. a phone call). This class of models is also related to exponential random graph models and to other dynamic event models as discussed by Stadtfeld and Geyer-Schulz (2010).

6.3.1 Two Individual Decision Levels

The occurrence of new phone calls in the data set is modeled as a two-level actor-oriented decision process as proposed by Snijders (2005). Both decision levels are estimated separately and are assumed to be independent:

First decision: Actors in the data set are assumed to have an individual *activity rate* determining the points in time at which to start a phone call. For each actor, the activity rate has been estimated separately for each hour h (24 activity rates/actor). This decision is modeled as a Poisson process which means that the time spans between two consecutive events of the same actor a_i are exponentially distributed with an individual parameter $\rho_i(h)$.

Second decision: Whenever an actor starts a phone call he/she evaluates the set of potential receivers and picks the receiver based on a multinomial logit choice model. The choice of receivers is influenced by a set of independent variables – in this paper the structures measured in the graphs of section 6.2.1 are used as predictors. They are explained in detail in section 6.3.2.

6.3.2 Multinomial Choice Statistics

The choice of event receivers is modeled as a multinomial actor choice logit model as introduced by McFadden McFadden (1974) and explained in detail in Hosmer and Lemeshow (2000); Cramer (2003). This class of models allows to test whether a discrete choice (in

Table 6.1: The multinomial choice statistics

ID	Name	Fig.	Equ.	Sec.
0	Repeated phone calls	6.2(a)	(6.1)	6.3.2
1	Reciprocity	6.2(b)	(6.2)	6.3.2
2	Receiver Indegree	6.2(c)	(6.3)	6.3.2
3	Receiver Outdegree	6.2(d)	(6.4)	6.3.2
4	Shared neighbors	6.2(e)	(6.5)	6.3.2
5	Distance in hops $\in [0,2]$	6.3(a)	(6.6)	6.3.2
6	Last call $\in [0\ min, 1\ min]$	6.3(b)	(6.7)	6.3.2
7	Last call $\in [1\ min, 30\ min]$	6.3(b)	(6.8)	6.3.2
8	Last call $\in [30\ min, 12\ hrs]$	6.3(b)	(6.9)	6.3.2

this case: The choice of a communication partner from among all potential communication partners in the data set) is influenced by independent variables measured in the sample. The independent variables of the choice model in this paper are of three different kinds: Endogenous communication structures, time network structures, and distances between actors in the cell tower neighborship network. Apart from to these classes of independent variables, any interaction between the effects is allowed to be part of the models. Table 6.1 gives an overview of the basic independent variables discussed below.

Endogenous Communication Structures

Five structures are measured as independent, endogenous variables. *Repeated phone calls* and *Reciprocity* test whether people tend to communicate with those they have an outgoing communication tie to or a incoming communication tie from. *Receiver indegree* and *Receiver outdegree* measure the tendency of choosing actors with many in- or outgoing communication ties. The last endogenous independent variable *Shared neighbors* measures the number of contacts between both the sender and receiver of the event. The direction of this connection is not considered. The endogenous statistics are shown in figures 6.2(a) to 6.2(e). The first two statistics

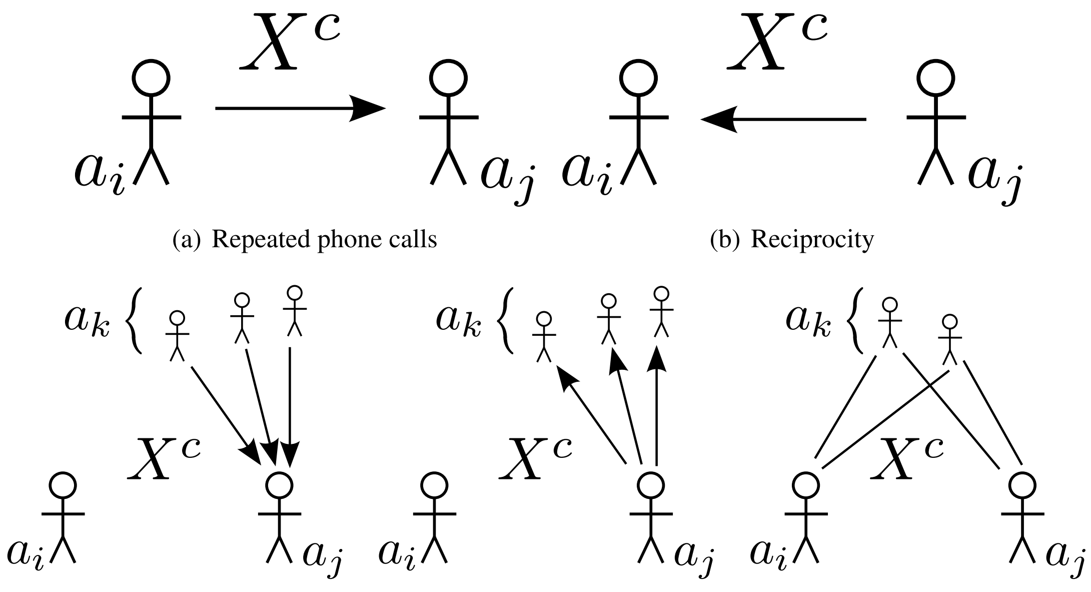
$$s_0(x,i,j) = \mathbf{I}_{\mathbb{R}^+}(x^c_{ij}) \tag{6.1}$$

$$s_1(x,i,j) = \mathbf{I}_{\mathbb{R}^+}(x^c_{ji}) \tag{6.2}$$

measure the existing of a tie in the sender-receiver dyad and correspond to figures 6.2(a) and 6.2(b). The two degree statistics (in-degree and out-degree of the event receiver) from figures 6.2(c) and 6.2(d) are defined by statistics s_2 and s_3.

$$s_2(x,i,j) = \sum_{a_k \in A} \mathbf{I}_{\mathbb{R}^+}(x^c_{kj}) \tag{6.3}$$

$$s_3(x,i,j) = \sum_{a_k \in A} \mathbf{I}_{\mathbb{R}^+}(x^c_{jk}) \tag{6.4}$$



(a) Repeated phone calls

(b) Reciprocity

(c) Indegree of receiver

(d) Outdegree of receiver

(e) Shared neighbors

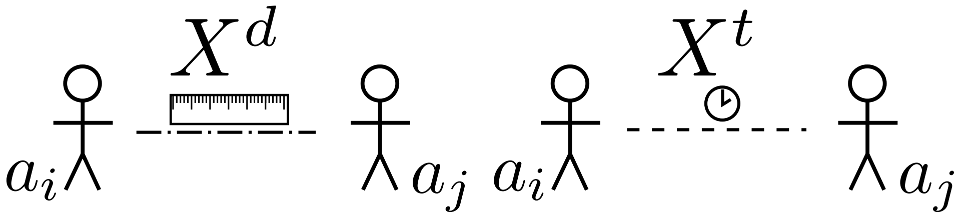
Figure 6.2: Endogenous communication potentially influencing the choices of call receivers. Actor a_i calls, a_j receives.

The last endogenous statistic that is shown in figure 6.2(e) counts the number of actors that both event sender and event receiver are connected to in the communication graph (shared neighbors):

$$s_4(x,i,j) = \sum_{a_k \in A} \left(\mathbf{I}_{\mathbb{R}^+}\left(\max(x^c_{ik}, x^c_{ki})\right) \cdot \mathbf{I}_{\mathbb{R}^+}\left(\max(x^c_{jk}, x^c_{kj})\right) \right) \tag{6.5}$$

with the indicator function $\mathbf{I}_C(b) = 1$ only if $b \in C$. For example, $\mathbf{I}_{\mathbb{R}^+}(b) = 1$ if $b > 0$, otherwise it is 0.

Distance Between Actors



(a) Distance in cell tower hops

(b) Time since last call

Figure 6.3: Variables measured on graphs X^t, X^d

It was decided to measure the distance as a dichotomized variable instead of using the number of hops as a metric variable. The hop distance only is a rough approximation of the real spatial distance of two actors. Therefore, it is only distinguished between ”rather close“ and ”rather far away“. There are many separate components in the cell tower neighborship network – an unknown distance is expressed by ”rather far away“. Figure 6.3(a) shows the idea of the statistic. Formally, it is defined by

$$s_5(x,i,j) = \mathbf{I}_{[0,2]}(x^d_{ij}). \tag{6.6}$$

where a distance of ≤ 2 hops is measured as 1. It indicates whether actors tend to communicate in ”rather close“ distance.

Time Since the Last Event

Assuming that the time intervals between two calls of one person are roughly exponentially distributed, it is tested whether the choice of receivers depends on the time since the last interaction (see figure 6.3(b)):

$$s_6(x,i,j) = \mathbf{I}_{[0,1min]}(x^t_{ij}) \tag{6.7}$$

$$s_7(x,i,j) = \mathbf{I}_{(1min,30min]}(x^t_{ij}) \tag{6.8}$$

$$s_8(x,i,j) = \mathbf{I}_{(30min,12hrs]}(x^t_{ij}) \tag{6.9}$$

The dummy variable for time intervals of more than 12 hours is suppressed because of multicollinearity.

Interaction Effects

Besides the nine basic independent variables proposed, we also test for interaction effects. An independent variable for the interaction effect of two variables can be defined as follows

$$s_m(x,i,j) = s_k(x,i,j) \times s_l(x,i,j) \tag{6.10}$$

with k,l,m being indexes. Multiple interactions are possible. If an interaction has a significant effect in a model it can be inferred that the two underlying statistics are not linear independent but that the co-existence has an additional positive or negative effect on communication choices.

A vector of statistics $s(x,i,j)$ defines a concrete model of the multinomial second decision. A model with all basic variables without interactions would yield the following values for the choice of a_j by actor a_i in figure 6.1:

$$s(x,i,j) = \begin{pmatrix} s_0(x,i,j) \\ \vdots \\ s_8(x,i,j) \end{pmatrix} = (1,1,2,2,2,0,0,1,0)^T \tag{6.11}$$

6.3.3 Markov Process

The state of the Markov process is defined by the state of the three graphs $x = (x^c, x^t, x^d)$, the subset of actors who are connected to a cell tower ($A^+ \subseteq A$), and the time of the day h that is measured in 24 discrete steps ($h \in \{1, \ldots, 24\}$). Changes of the process state caused by the decay of ties in x^c, by the increase of ties in x^t, by changes of cell tower connections or position, or by the change of h are not described explicitly in this Markov process. Moreover, the process state is assumed to be stable for short time spans. This is reasonable as the mentioned external processes are rather slow compared to the rate of phone calls in the data set (especially, as most independent variables were measured on a binary scale).

The occurrence of events is modeled as a Poisson process. The time span between two consecutive events therefore is exponentially distributed. This holds both for the time between two calls of the same actor a_i without regarding who is called ($a_i \rightarrow a_?$) and for the timespan until a_i makes a phone call to the specific receiver a_j ($a_i \rightarrow a_j$). The probabilities of these process changes from an actor's perspective are

$$P(a_i \rightarrow a_? \text{ in } [t, t+\varepsilon]) = \rho_i e^{-\rho_i \varepsilon} \tag{6.12}$$

and

$$P(a_i \rightarrow a_j \text{ in } [t, t+\varepsilon]) = \lambda_{ij} e^{-\lambda_{ij} \varepsilon}. \tag{6.13}$$

Equation 6.12 is related to the first actor decision from section 6.3.1 (*when* to start a phone call). Equation 6.13 is related to a combination of the first and the second actor decision (*who* is called). The combination of all individual rates for all potential call receivers determines the transitions of the Markov process:

$$\lambda_{ij}(x, \beta, h, A^+) \approx \rho_i(h) p_{ij}(x, \beta, A^+) \tag{6.14}$$

λ_{ij} is the Poisson rate (the propensity) of any actor a_i to call actor a_j and, thus, to change the process state.
$\rho_i(h)$ is the general activity of actor a_i to start phone calls. It depends on h, the time of the day.
$p_{ij}(x, \beta, A^+)$ is the probability of actor a_i choosing a_j as a receiver over any other actor in the set of currently active actors in $A^+ \subseteq A$. It is a multinomial logit choice model as introduced by McFadden (1974). The choice of the actual event receiver depends on the independent variables in $s(x, i, j)$ from section 6.3.2, which are measured on the current networks in x. The statistics are weighted with the parameter vector β:

$$p_{ij}(x, \beta, A^+) = \frac{\exp\left(\beta^T s(x, i, j)\right)}{\sum_{a_k \in A^+ \setminus \{a_i\}} \exp\left(\beta^T s(x, i, k)\right)}. \tag{6.15}$$

6.4 Results

The results for the two decision levels from section 6.3.1 (actor activity and choice of receivers) are presented separately.

6.4.1 Actor Activity Rates

The individual Poisson rates $\rho_i(h)$ of all actors $a_i \in A$ that determine the general activity of actors (the first decision in the two-level decision processes) were calculated as maximum likelihood estimates. The distributions of individual rates for each hour of the day h are presented as box plots for each hour of the day in figure 6.4. The horizontal lines show the minimum and maximum values without extreme outliers and the 25%-, 50%- and 75%-quantiles.

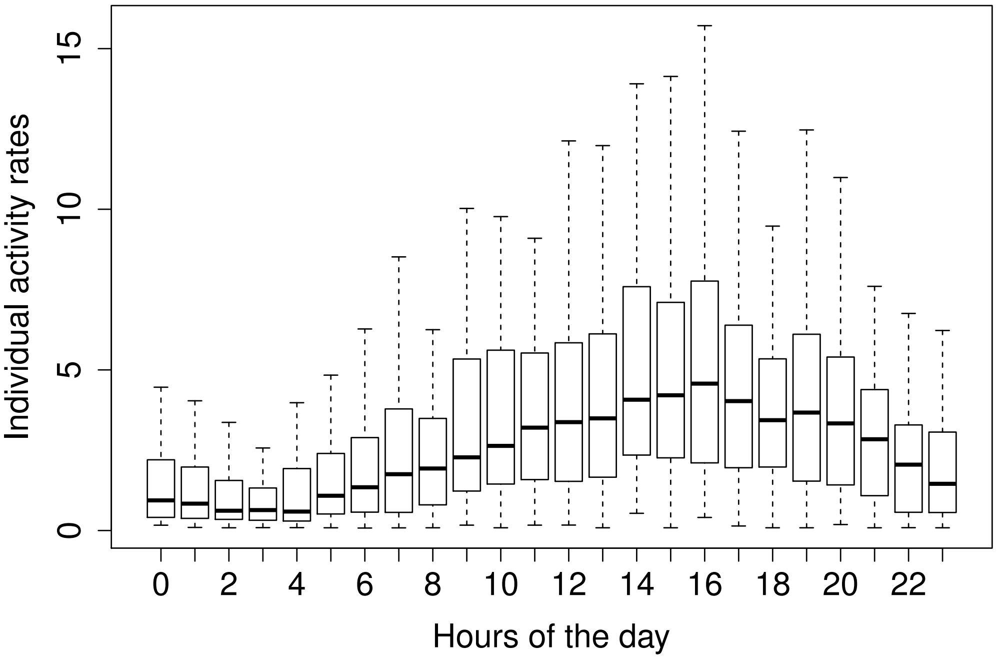


Figure 6.4: Distribution of individual activity Poisson rates at different day times.

The mean value of the Poisson rates varies between $\overline{\rho}(h=2) = 0.62$ for the time between 2 and 3am and $\overline{\rho}(h=16) = 4.57$ for the time between 4 and 5pm. The values are interpreted as follows: The average actor tends to start 0.62 calls between 2 and 3 am within 24 days (as the time slot is only one of 24 hours of a day), while the average actor is expected to start 4.57 phone calls between 4 and 5pm within 24 days. The variance of the individual rates observed is higher at day times.

6.4.2 Multinomial Receiver Choices and Model Selection

The parameters of the multinomial actor choices were calculated as maximum likelihood estimates using a Newton Raphson algorithm (see Young and Smith (2005)). They are given in table 6.2 using the IDs of the basic statistics in table 6.1. Interaction effects between two variables a and b are indicated by axb. Multiple interactions are possible. $\log L$ indicates the log likelihood of a model. A concrete model M is defined by selecting variables from the basic decision statistics and their interaction effects ($M = \{s_l, s_m, \dots\}$, see section 6.3.2).

The model selection process is defined by selecting the variable with the largest reduction in Akaike's Information Criterion (AIC; Hurvich and Tsai (1989)). For performance reasons, interaction variables were only considered eligible for selection if their parts were already included in the model. The model was fitted by an iterative forward inclusion algorithm with M_k being the model of iteration step k after k inclusions of parameters. The model fitting process started with the empty model $M_0 = \{\}$ as base line which represents the likelihood of a random choice between potential receivers. Table 6.2 shows the results of the first 14 iteration steps of the model fitting algorithm.

It turned out that of the nine basic parameters in table 6.1 eight were part of the best 11-parameter model M_{11} plus three interaction effects. Only the number of shared neighbors did not have an effect on communication choices at this level. The significance level of almost all parameters has a value of $p < 0.001$. Insignificant parameters or parameters with a lower level of significance are italicized. The parameter estimates are discussed in more detail in section 6.5.

It is well known that the AIC criterion tends to overspecify the model selected. Hurvich and Tsai (Hurvich and Tsai, 1989, p.300) suggested the corrected AIC criterion (AIC$_C$) with the additional penalty term $+\frac{2k(k+1)}{n-k-1}$ as a termination criterion for the model selection process. Table 6.2 shows one possibly overspecified models (M_{11}) with M_{10} having the best fit regarding the AIC$_C$. Note that additional parameters included (e.g. the *Distance in hops* (5) in M_{11} and further parameters in the subsequent models that are not shown) remain highly significant.

6.5 Discussion of Receiver Choice Parameters

The parameter explaining most of each model was *Repeated phone calls* (0), meaning that people in the data set tended to call those they had called before. In model M_1 the probability of choosing a sender the receiver was connected with a directed communication tie to is $e^{3.517} = 33.68$ times higher than without this structure. The next three parameters are related to the time since the last communication (no matter who of the communication partners started the last call). These effects are measured on a categorical scale. The effect of communication more than 12 hours ago was left out. It can be seen that all three effects are significantly positive. The effects of the very short time span (6) and the short time span (7) have very high absolute values. Compared to a decision in favor of calling a receiver the sender has communicated with by phone more than 12 hours ago, the probability of calling an actor with a very short or a short time span is $e^{4.285} = 72.602$ and $e^{3.416} = 30.447$ times higher. It is concluded that actors in the data set tend to communicate with the same communication partners several times within shorter time spans. The very short time span effect (6) might often be driven by technical problems, while the short time span effect (7) indicates regular repeated communication. The time effects are closely related to *Repeated phone calls* (0), as in most cases of repeated interactions most phone calls take place on

Table 6.2: Results of the model fitting of receiver choice parameters

IDs from table 6.1	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
	Model M_0		**Model** M_1		**Model** M_2		**Model** M_3	
0 (Repeated calls)			3.517	0.035	3.316	0.0351	2.922	0.042
7 (Last call $\in [1\ min, 30\ min]$)					2.904	0.040	3.337	0.0362
8 (Last call $\in [30\ min, 12\ hrs]$)							1.883	0.0278
$\log L$	-37,655.329		-27,872.821		-25,391.170		-23,326.535	
AIC / AIC$_C$	7.510 / 7.510		5.559 / 5.560		5.064 / 5.066		4.653 / 4.655	
	Model M_4		**Model** M_5		**Model** M_6		**Model** M_7	
0 (Repeated calls)	2.841	0.037	2.563	0.040	3.030	0.050	2.944	0.0506
7 (Last call $\in [1\ min, 30\ min]$)	3.416	0.042	3.352	0.042	3.322	0.042	3.321	0.042
8 (Last call $\in [30\ min, 12\ hrs]$)	1.968	0.028	1.892	0.028	1.874	0.028	1.867	0.028
6 (Last call $\in [0\ min, 1\ min]$)	4.285	0.116	4.303	0.115	4.242	0.114	4.257	0.114
1 (Reciprocity)			0.511	0.029	1.624	0.067	1.573	0.067
0x1					-1.300	0.073	-1.261	0.073
2 (Receiver Indegree)							0.015	0.001
$\log L$	-22,540.723		-22,383.586		-22,238.576		-22,177.638	
AIC / AIC$_C$	4.496 / 4.500		4.465 / 4.471		4.436 / 4.445		4.425 / 4.436	
	Model M_8		**Model** M_9		**Model** M_{10}		**Model** M_{11}	
0 (Repeated calls)	2.897	0.0508	2.903	0.051	2.830	0.052	2.830	0.052
7 (Last call $\in [1\ min, 30\ min]$)	3.309	0.042	4.432	0.168	4.516	0.168	4.516	0.169
8 (Last call $\in [30\ min, 12\ hrs]$)	1.847	0.028	1.843	0.028	2.191	0.057	2.185	0.057
6 (Last call $\in [0\ min, 1\ min]$)	4.239	0.115	4.240	0.115	4.242	0.115	4.242	0.115
1 (Reciprocity)	1.655	0.068	1.548	0.071	1.462	0.073	1.461	0.073
0x1	-1.261	0.073	-1.152	0.077	-0.963	0.082	-0.963	0.082
2 (Receiver Indegree)	0.036	0.002	0.036	0.002	0.036	0.002	0.034	0.002
3 (Receiver Outdegree)	-0.026	0.002	-0.026	0.002	-0.025	0.002	-0.024	0.002
0x7			-1.184	0.172	-1.279	0.172	-1.283	0.173
0x1x8					-0.445	0.063	-0.448	0.063
5 (Distance in hops $\in [0, 2]$)							-0.201	0.033
$\log L$	-22,096.724		-22,074.686		-22,050.560		-22,032.674	
AIC / AIC$_C$	4.409 / 4.423		4.404 / 4.422		4.400 / 4.422*		4.396 / 4.423	

previously used communication paths. This is represented by the importance of the three interaction effects 0x7, 0x1x8 and 0x1x6 which are the 9th, 10th and 12th parameter included during the model fitting.

Reciprocity (1) turned out to be the fourth important parameter, the interaction with *Repeated phone calls* was the fifth parameter. It is significantly positive ("actors tend to reciprocate communication"). The interaction effect 0x1 is negative but as the sum of all three parameters in model M_6, for example, is $3.030+1.624-1.300 = 3.354$ which is more than one of the basic parameters, it may be stated that actors even more tend to communicate in bi-directional structures. Similar results were found by Stadtfeld and Geyer-Schulz (2010).

The next (also significant) effects included in models M_7 and M_8 measure the tendency of actors in the data set to communicate with others who have a high in- or outdegree. Compared to a random choice people tended to communicate with others who have many incoming communication ties. The probability of choosing a certain communication partner in model M_6 is increased by 1.51% ($e^{0.015} = 1.0151$) with each additional incoming tie. With each additional outgoing communication tie, the probability of choosing an actor decreased by 2.63% ($e^{0.026} = 1.0263$). These effects can be understood as follows: Some people tend to call many different people regularly – if a choice of communication partners is made, the necessity to call these people back is smaller, as there is a high chance of one of the many different previous calls being made with the current sender. Therefore, the outdegree effect (3) is negative. The tendency of choosing communication partners with many incoming ties may be related to the "popularity" of actors. However, we do not know enough about the data to make any further statements: "Popularity" may be related to social status, to general willingness to communicate with cell phones or to status within the university community.

The eleventh effect included is the binary measured distance (5). It is included in model M_{11} and has – as long it is not altered by further interaction effects – a significantly negative weight. Given the choice between two structurally equivalent potential event receivers, the probability of a choice of an actor in the range of 0 to 2 cell tower hops (this would, for example, cover all positions on the MIT campus) is 22.26 times smaller compared to a receiver further away ($\exp^{0.201} = 22.26$). This effect may be interpreted by parts of the close distance cell phone communication being substituted by other means of communication, e.g. face-to-face communication or local telephone systems in university buildings or dormitories. With the inclusion of two interaction effects with the in- and outdegree of the receiver the pure distance effects becomes insignificant as it seems to be explained well by the two interactions.

The *Shared neighbors* statistic (4) was not part of any of the first 14 models. Previous research of a closed online community (see Stadtfeld and Geyer-Schulz (2010)), however, revealed that this structure may have an important effect. In this context it must be noted that the participants do not form any closed social communities. 58.18% of the communication events were directed to non-participants of the experiments. This implies that important actors of the social groups of the participants are missing in the data set. This fact is less relevant to dyadic structures and degree structures. The *Shared neighbors* statistic(4) was only included in model M_{25} (not shown in table 6.2). However, it had a significantly positive effect on the choice of communication partners.

6.6 Conclusions and Further Research

This article presented a new model framework that can be used to obtain a deeper understanding of dynamic communication patterns in communities that are driven by individual choice. The methodology was applied to the cell phone communication data stream of the MIT Reality Mining data set. The choices of individuals were modeled as a two-level decision process: The general activity of actors was estimated separately from the multinomial choices determining communication partners. The multinomial logit model describing the second decision was fitted using local network structures in three different graphs as independent variables. It was found that dyadic structures in the communication network and time effects are important predictors for communication choices. But also the degree distribution of the call receiver and the distance between actors play a role in this decision process. Overall, all models improved the log likelihood of the random choice base line model M_0 significantly.

The description of the stochastic process as a Markov process comes with certain assumptions that are a simplification of human behavior. It is implicitly assumed that all actors in the observed community roughly follow homogeneous communication patterns. The MIT data set was collected in a university environment with students, researchers and staff members. People do have roles within such a community. These roles probably lead to specific communication patterns. As these roles were latent, roles were not incorporated as actor attributes.

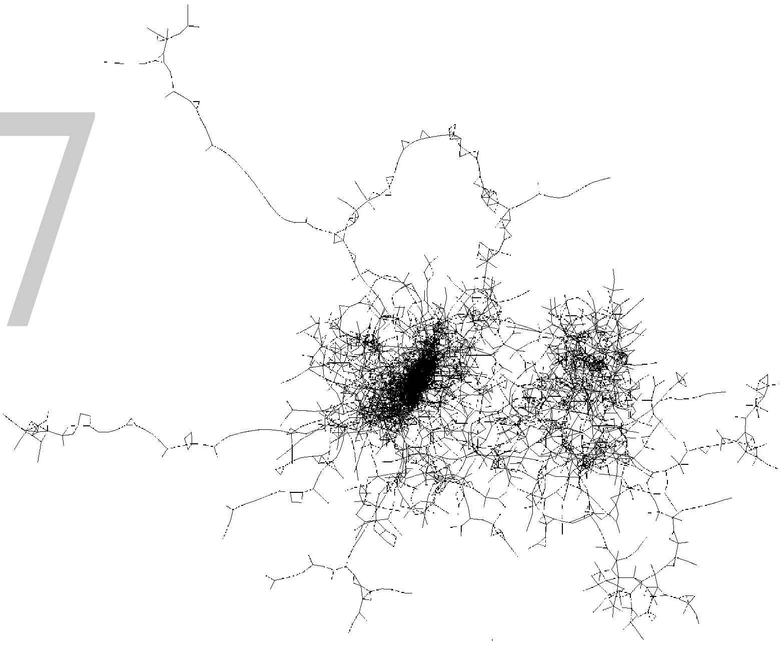
The independence assumption between the two levels of the Markov process is critical. This holds even more when a time related parameterization of the Poisson rates and of the multinomial decisions is applied.

As mentioned in section 6.5, the observed data set does not represent a closed community but only a sample of a much bigger communication network. Probably, many unobserved, but important communication patterns exist. Structures measuring the effect of third actors that both sender and receiver are connected to are blurred due to this restriction of the data set.

Despite these shortcomings, we show that the proposed methodology helps to gain interesting insights in dynamic communication behavior of communities.

In the future research the methodology is planned to be applied to other communication data sets in order to learn more about actor roles and network structures. Furthermore, it is planned to investigate model selection strategies for structural network models.

7 Conclusions and Outlook



The increasing availability of event data (e.g., collected by computer mediated communication and social network sites) during the last decade has given rise to new research questions from various disciplines. This development has been accompanied by a call for appropriate statistical tools that enable modeling such data. In this thesis, we presented a new stochastic actor-oriented framework for the analysis of dynamic event streams in social networks. This book aimed at answering research questions of three different types:

1. *Methodical*: How can the occurrence of events be described by an actor-oriented model? Can we express different levels of decisions that lead to an event? How can the model be extended and specified? How can the model be fitted?

2. *Computational*: How can model parameters be estimated efficiently? Which optimization algorithms can be used? Are there heuristics and means of preprocessing that help to reduce computational complexity?

3. *Substantive*: Given a real data set, can we learn more about individual behavior with the new methodology? Do actors tend to communicate within dense clusters? Is there an effect of entity affiliation on the choice of event receivers? How do time intervals and spatial distance influence communication behavior? Do communication patterns in a community change over time?

We showed that event stream dynamics can be modeled in a similar way as the stochastic actor-oriented model for network panel data. Different individual decisions are embedded in a Markov process and estimated separately. The newly developed framework is highly flexible and may be specified in multiple ways. We applied a new software tool on two different data sets to demonstrate the estimation process of the framework. Thereby, we showed that the framework is a useful tool for gaining deeper insights into the dynamic decisions in social networks.

7.1 Summary

In chapter 2 we discussed previous work related to our framework. After introducing multinomial logit models, three structural network models were discussed: Exponential random graph models, stochastic actor-oriented models and event models based on event history modeling can all be applied to understand structural effects in social networks. However, none of these models was found to be both actor-oriented and applicable on event data sets.

Therefore, in chapter 3 we introduced a new event framework that can be seen as an extension and adaption of stochastic actor-oriented models for panel data. Events in an event stream are assumed to generate structures that predict future events. The basic event framework was introduced as a Markov process with two individual decision levels, which were modeled separately: The activity of actors and the choice of event receivers were defined as two independent sub-processes. The receiver choice sub-process was investigated in more detail: It was shown that the model can be estimated efficiently and that parameter estimates can be interpreted in a meaningful way. We explained every step of the basic event framework using a short exemplary event stream. An exemplary model was specified with basic, dyadic parameters.

In chapter 4 the basic framework was extended by a variety of possible extensions and new specifications. We discussed that event streams may sometimes incorporate valuable information that exceeds the basic form presented in chapter 3. This information could be exploited to answer more elaborate research questions. We proposed to define an extended process state that includes actor covariates (i.e., attributes) and dyadic covariates (i.e., states of other graphs). This state can be changed by introducing additional decision levels and additional transition rates. The wealth of information in event streams allows to estimate parameterized activity rates and individualized choice models. We showed that the choice of event receivers can be specified by a multitude of variables that are measured in the extended process state. We also presented some basic ideas for systematic model fitting. Some of the extensions proposed in this chapter were applied in chapters 5 and 6.

In chapter 5 the new event framework was applied to an event stream, namely the sending of private messages in a question and answer web community. The influence of different types of structural variables on the choice of event receivers was tested: It was revealed that both endogenous structures in the communication graph and exogenous structures in a question affiliation two-mode graph were important when choosing receivers. The relative importance of endogenous effects increased over time. In this application, we introduced an additional decision level to distinguish between active and inactive user accounts, which reduced the computational complexity substantially.

In chapter 6 cell phone communications within a group of students was analyzed. In this application, we additionally parameterized the actor activity rates that determine timestamps of events. We found a variety of significant parameters in the receiver choice submodel. In addition to significant endogenous effects, time and spatial distances had a positive influence on the choices of event receivers.

7.2 Future research

The last two application chapters shed light on possible applications of the new event framework. An important issue in future research will be applying the new event framework to further data sets in order to explore the following topics:

1. Model extensions and new specifications
2. Robustness of the estimation process and comparability of estimates
3. Model fitting strategies
4. Dynamics of attribute changes
5. Substantive research on individual choice behavior

Each of these five topics is discussed in more detail in the following.

First, although the two applications demonstrated a range of possible extensions and new specifications, some of the proposed extensions of chapter 4 remained untested. For instance, we did not yet analyze a data set with different types of events that are defined by competing transition rates. This was proposed in section 4.3.3. In communication data sets this may be reasonable, as some means of communication substitute others. For example, text messages and phone calls both indicate dyadic information exchange and could be modeled together in the same communication graph. Furthermore, the parameterization of actor activity rates was kept rather simple and did not evaluate structural network effects or attributes. We proposed this extension in section 4.3.1. It could be tested whether actor attributes like gender or formal role predict the event activity of actors. In section 4.3.4, we proposed to estimate *individual* choice models to investigate individual differences in choice behavior. In marketing-related data sets this approach could be interesting: One could, for example, analyze which actors are most likely to spread product information to other actors within their local social environment. Also our proposed weighted statistics in section 4.4.5 may be applied in future case studies. The influence of edge weights on the choice of event receivers can generate interesting insights. For example, the *Repeated communication* effect can probably be improved significantly once it takes the *weight* of "re-usable" ties into account.

Second, further elaboration regarding the robustness of the estimation process and the comparability of the estimates is needed. Event stream data has the major advantage that the analyzed periods do not have to be defined in advance, as it is the case with panel data analyses. In an event data set, arbitrary time spans can be defined ex-post. This allows a comparison of individual decision models between different sub-phases of the same data set. However, when comparing estimates of different event streams we have to deal with the difficulty that the estimated values are influenced by varying factors, like network density, number of observations, or number of potential receivers per event. To make estimates comparable, standardization is necessary. Standardized estimates can be analyzed with a sliding window approach. This enables the visualization of behavioral patterns over time. This may be especially interesting in phases where changes from outside the network influence the observed set of actors. If the data set describes an organization, it could be tested whether changes in the formal organizational structure influence informal communication patterns over time.

Third, the fitting of structural network models is a widely researched topic. In section 4.5 we proposed some preliminary ideas on fitting models. These ideas should be extended in further research. A model fitting algorithm can exploit the fact that structural covariates are not independent of each other but nested in hierarchical structures, e.g. dyadic ties are embedded in triangular structures.

Fourth, research on disentangling selection and influence effects has become a popular topic. In social networks, a high similarity is often observed among connected actors. Selection on the one hand is understood as creating ties to receivers who are similar to the actor. Influence on the other hand is understood as an attribute change: Actors who are connected become similar in their attributes over time. Combined selection and influence models could be defined for dynamic event stream data sets as well. One possible application scenario is similarity in opinions in web log networks. Do web logs form link-clusters with authors who have similar political opinions? Or do connected web logs adopt opinions of influential others? Both linking events and attribute changes can be studied over time.

Finally, we are confident that applying the presented methodology to novel research questions and interesting data sets will generate new insights into the dynamics of human behavior in social networks. The aim of this book was to provide the appropriate statistical tools to researchers from different disciplines studying events in social networks. With these tools researchers may target many more research questions of substantive nature, most importantly in the fields of social sciences, political sciences, marketing research and economics.

Bibliography

Agneessens, F., Wittek, R., 2008. Social capital and employee well-being: Disentangling intrapersonal and interpersonal selection and influence mechanisms. Revue française de sociologie 3 (1), 613–637.

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. John Wiley, Hoboken, New Jersey.

Agresti, A., 2007. An Introduction to Categorical Data Analysis, 2nd Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716 – 723.

Anderson, C. J., Wasserman, S., Crouch, B., 1999. A p* primer: Logit models for social networks. Social Networks 21 (1), 37 – 66.

Anton, H., 1994. Lineare Algebra. Spektrum Akademischer Verlag, Heidelberg, Berlin.

Arrow, K. J., 1951. Social choice and individual values. Monographs / Cowles Commission for Research in Economics ; 12. John Wiley & Sons, Inc., New York; Chapman & Hall, London.

Ben-Akiva, M., McFadden, D., Gärling, T., Gopinath, D., Walker, J., Bolduc, D., Börsch-Supan, A., Delquié, P., Larichev, O., Morikawa, T., Polydoropoulou, A., Rao, V., 1999. Extended framework for modeling choice behavior. Marketing Letters 10, 187–203.

Berman, A., Plemmons, R. J., 1979. Nonnegative matrices in the mathematical sciences. Academic Press Inc., New York.

Besag, J., 1975. Statistical analysis of non-lattice data. Journal of the Royal Statistical Society. Series D (The Statistician) 24 (3), 179–195.

Besag, J. E., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) 36 (2), 192–236.

Blossfeld, H.-P., Rohwer, G., 1995. Techniques of event history modeling : New approaches to causal analysis. Erlbaum, Mahwah, NJ.

Box-Steffensmeier, J. M., Jones, B. S., 1997. Time is of the essence: Event history models in political science. American Journal of Political Science 41 (4), 1414–1461.

Brandes, U., Erlebach, T., 2005. Network Analysis: Methodological Foundations. Springer.

Brandes, U., Lerner, J., Snijders, T. A. B., 2009. Networks evolving step by step: Statistical analysis of dyadic event data. In: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009). IEEE Computer Society, pp. 200–205.

Bronstein, I., Semendjajew, K., Musiol, G., Mühlig, H., 2001. Taschenbuch der Mathematik, 5th Edition. Verlag Harri Deutsch.

Burk, W. J., Steglich, C. E. G., Snijders, T. A. B., 2007. Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. International Journal of Behavioral Development 31 (1), 397–404.

Burnham, K. P., Anderson, D. R., 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33 (2), 261–304.

Butts, C. T., 2008. A relational event framework for social action. Sociological Methodology 38 (1), 155–200.

Carrington, P., Scott, J., Wasserman, S., 2005. Models and Methods in Social Network Analysis. Vol. 27 of Structural Analysis in the Social Sciences. Cambridge University Press, New York.

Casella, G., Berger, R. L., 2002. Statistical Inference, 2nd Edition. Duxbury Advanced Series. Duxbury, Pacific Grove.

Centola, D., Macy, M., 2007. Complex contagions and the weakness of long ties. American Journal of Sociology 113 (3), 702–734.

Christakis, N. A., Fowler, J. H., 2007. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine 357 (1), 370–379.

Christakis, N. A., Fowler, J. H., 2008. The collective dynamics of smoking in a large social network. The New England Journal of Medicine 358 (1), 2249–2258.

Coleman, J. S., 1964. Introduction to mathematical sociology. Free Press of Glencoe, New York.

Coleman, J. S., 1981. Longitudinal Data Analysis. Basic Books, Inc.

Coleman, J. S., 1990. Foundations of social theory. Belknap Press of Harvard University Press, Cambridge, Mass.

Contractor, N. S., Wasserman, S., Faust, K., 2006. Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. Academy of Management Review 31 (3), 681 – 703.

Corander, J., Dahmström, K., Dahmström, P., 1998. Research report 1998:8. Maximum likelihood estimation for markov graphs. Tech. rep., Department of Statistics, University of Stockholm.

Cosslett, S. R., 1981. Efficient estimation of discrete-choice models. In: Manski and McFadden (1981b), Ch. 2, pp. 51–111.

Cox, D. R., Hinkley, D. V., 1974. Theoretical statistics. Chapman & Hall, London.

Cramer, J., 2003. Logit Models from Econometrics and other fields. Cambridge University Press, Cambridge.

Davison, A. C., 2003. Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

de Nooy, W., 2008. Signs over time: Statistical and visual analysis of a longitudinal signed network. Journal of Social Structure 9 (1), 1.

de Nooy, W., 2011. Networks of action and events over time. A multilevel discrete-time event history model for longitudinal network data. Social Networks 33 (1), 31–40.

Deuflhard, P., 2004. Newton Methods for Nonlinear Problems. No. 35 in Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg.

Duller, C., 2008. Einführung in die nichtparametrische Statistik mit SAS und R: Ein anwendungsorientiertes Lehr- und Arbeitsbuch. Physica-Verlag (Springer), Heidelberg.

Dunbar, R. I. M., 1992. Neocortex size as a constraint on group size in primates. Journal of Human Evolution 22 (6), 469–493.

Durrett, R., 2007. Random Graph Dynamics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Eagle, N., Pentland, A. S., Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 106 (36), 15093–15514.

Efron, B., Hinkley, D. V., 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika 65 (3), 457–482.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1 (1), 54–77.

Erdős, P., Rényi, A., 1959. On random graphs I. Publicationes Mathematicae 6 (1), 290–297.

Erdős, P., Rényi, A., 1960. On the evolution of random graphs. Publications of the Hungarian Academy of Sciences 5 (1), 17 – 61.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, Berlin.

Fowler, M., 2002. Refactoring: Improving the design of existing code, 10th Edition. The Addison-Wesley Object Technology Series. Addison-Wesley, Boston.

Fowler, M., 2003. Analysis Patterns: Reusable Object Models, 13th Edition. The Addison-Wesley Object Technology Series. Addison-Wesley, Boston.

Franceschetti, M., Meester, R., 2007. Random Networks for Communication: From Statistical Physics to Information Systems. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Frank, O., 1991. Statistical analysis of change in networks. Statistica Neerlandica 45, 283–293.

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

Freedman, D. A., 2009. Statistical Models : Theory and Practice, revised edition Edition. Cambridge University Press, New York.

Freeman, L. C., 2004. The Development of Social Network Analysis : A Study in the Sociology of Science. Empirical Press, Vancouver.

Gamma, E., Helm, R., Johnson, R., Vlissides, J., 1995. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional Computing Series. Addison-Wesley Publishing Company, Reading.

Geyer, C. J., Thompson, E. A., 1992. Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society, Series B 54 (3), 657–699.

Goodreau, S. M., 2007. Advances in exponential random graph (p*) models applied to a large social network. Social Networks 29 (2), 231 – 248, special Section: Advances in Exponential Random Graph (p*) Models.

Goos, G., 2000. Vorlesungen über Informatik. Band 1: Grundlagen und funktionales Programmieren (Springer-Lehrbuch), 3rd Edition. Vol. 1 of Vorlesungen über Informatik. Springer-Verlag, Berlin.

Greene, W. H., 2008. Econometric Analysis, 6th Edition. Pearson Prentice Hall, Upper Saddle River.

Greiner, W., Neise, L., Stöcker, H., 1993. Theoretische Physik Bd. 9, Thermodynamik und Statistische Mechanik, 2nd Edition. Vol. 9. Verlag Harri Deutsch, Frankfurt am Main.

Holland, P. W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.

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

Holland, P. W., Leinhardt, S., 1977b. Social structure as a network process. Zeitschrift für Soziologie 6 (4), 386–402.

Holland, P. W., Leinhardt, S., 3 1981. Exponential family of probability distributions for directed graphs. Journal of the American Statistical Association 76 (373), 33–50.

Hosmer, David, W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Hunter, D. R., 2007. Curved exponential family models for social networks. Social Networks 29 (2), 216–230, special Section: Advances in Exponential Random Graph (p*) Models.

Hurvich, C. M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), pp. 297–307.

Johnson, J. C., Luczkovich, J. J., Borgatti, S. P., Snijders, T. A., 2009. Using social network analysis tools in ecology: Markov process transition models applied to the seasonal trophic network dynamics of the chesapeake bay. Ecological Modelling 220 (22), 3133–3140, special Issue on Cross-Disciplinary Informed Ecological Network Theory - Selected Papers from the Third Workshop on Ecological Network Analysis, University of Georgia, Athens, GA, USA, April 2008.

Knuth, D. E., 2000. The Art of Computer Programming: Fundamental Algorithms, 3rd Edition. Vol. 1. Addison-Wesley, Reading.

Koopman, B. O., 1936. On distributions admitting a sufficient statistic. Transactions of the American Mathematical Society 39 (3), 399–409.

Koskinen, J., Edling, C., 2010. Modelling the evolution of a bipartite network – peer referral in interlocking directorates. Social NetworksIn Press.

Koskinen, J. H., Snijders, T. A. B., 2007. Bayesian inference for dynamic social network data. Journal of Statistical Planning and Inference 137 (12), 3930–3938, 5th St. Petersburg Workshop on Simulation, Part II.

Liben-Nowell, D., Kleinberg, J., 2003. The link prediction problem for social networks. In: Proceedings of the 12th International Conference on Information andKnowledge Management (CIKM). ACM, New York.

Lindsey, J., 2004. Statistical Analysis of Stochastic Processes in Time. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Lubbers, M. J., Molina, J. L., Lerner, J., Brandes, U., Àvila, J., McCarty, C., 2010. Longitudinal analysis of personal networks. The case of Argentinean migrants in spain. Social Networks 32 (1), 91–104, special Issue on Dynamics of Social Networks.

Lubbers, M. J., Snijders, T. A., 2007. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. Social Networks 29, 489–507.

Luce, D., 1977. The choice axiom after twenty years. Journal of Mathematical Psychology 15 (1), 215–233.

Luce, R. D., 1959. Individual choice behavior. Wiley, Chapman & Hall.

Macy, M. W., Willer, R., 08 2002. From factors to actors: Computational sociology and agent-based modeling. Annual Review of Sociology 28 (1), 143–166.

Maddala, G. S., 1983. Limited-dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Maddala, G. S., 2001. Introduction to Econometrics, 3rd Edition. John Wiley, Chichester.

Maddala, G. S., Kim, I.-M., 1998. Unit Roots, Cointegration, and Structural Change. Cambridge University Press, Cambridge, UK.

Manski, C. F., 1977. The structure of random utility models. Theory and Decision 8 (3), 229–254.

Manski, C. F., McFadden, D., 1981a. Alternative estimators and sample designs for discrete choice analysis. In: Manski and McFadden (1981b), Ch. 1, pp. 2–50.

Manski, C. F., McFadden, D. (Eds.), 1981b. Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, Massachusetts.

Marschak, J., 1960. Binary-choice constraints and random utility indicators. In: Arrow, K. J., Karlin, S., Suppes, P. (Eds.), Proceedings of the First Stanford Symposium on Mathematical Methods in the Social Sciences, 1959. Vol. 4 of Stanford Mathematical Studies. Stanford University Press, Stanford, Ch. 21, pp. 312– 329.

McCullagh, P., Nelder, A., 1992. Generalized Linear Models, 3rd Edition. Vol. 37 of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press Inc., New York, Ch. 4, pp. 105–142.

McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski and McFadden (1981b), Ch. 5, pp. 198–272.

McFadden, D., 2001. Economic choices. American Economic Review 91 (3), 351 – 378, nobel prize acceptance speech.

Miller, A., 2002. Subset Selection in Regression, 2nd Edition. Vol. 95 of Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton.

Monge, P. R., Contractor, N. S., 2003. Theories of Communication Networks. Oxford University Press, New York, USA.

Moreno, J. L., 1954. Die Grundlagen der Soziometrie (Original Title: Who shall survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama). Westdeutscher Verlag, Köln und Opladen, German translation by Grete Leutz.

Moreno, J. L., Jennings, H. H., 1938. Statistics of social configurations. Sociometry 1 (3/4), 342–374.

Myers, R. H., Montgomery, D. C., Vining, G. G., 2002. Generalized Linear Models. Wiley Series In Probability And Statistics. John Wiley & Sons, Inc., New York.

Newman, M., 2010. Networks: An Introduction. Oxford University Press, Oxford.

Norris, J. R., 1997. Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

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

Powell, W. W., White, D. R., Koput, K. W., Owen-Smith, J., 1 2005. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. The American Journal of Sociology 110 (4), 1132–1205.

Ripley, R. M., Snijders, T. A. B., Preciado Lopez, P., 2011. Manual for SIENA 4.0. Nuffield College and Department of Statistics, University of Oxford.

Robins, G., 2009. Exponential random graph (p*) models for social networks. In: Myers, R. (Ed.), Encyclopaedia of Complexity and System Science. Springer, New York, pp. 8319–8333.

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

Robins, G., Pattison, P., Wang, P., 2009. Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. Social Networks 31 (2), 105 – 117.

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

Robins, G., Snijders, T. A. B., Wang, P., Handcock, M., Pattison, P., 2007b. Recent developments in exponential random graph (p*) models for social networks. Social Networks 29 (2), 192–215.

Schweinberger, M., Snijders, T. A. B., 2007. Markov models for digraph panel data: Monte Carlo-based derivative estimation. Computational Statistics & Data Analysis 51 (1), 4465–4483.

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

Snijders, T. A. B., 2001. The statistical evaluation of social network dynamics. In: Sobel, M., M.P., B. (Eds.), Sociological Methodology. Vol. 31. Boston and London: Basil Blackwell, pp. 361–395.

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

Snijders, T. A. B., 2005. Models for longitudinal network data. In: P., C., Scott, J., Wasserman, S. (Eds.), Models and methods in social network analysis. Vol. 27 of Structural Analysis in the Social Sciences. Cambridge University Press, New York, Ch. 11, pp. 215–247.

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

Snijders, T. A. B., Bosker, J. R., 1999. Multilevel Analysis: An introduction to basic and advanced multilevel modelling. SAGE, New Delhi.

Snijders, T. A. B., Koskinen, J., Schweinberger, M., 2010a. Maximum likelihood estimation for social network dynamics. The Annals of Applied Statistics 4 (2), 567–588.

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

Snijders, T. A. B., Steglich, C. E. G., Schweinberger, M., Huisman, M., 2008. Manual for SIENA version 3.2. ICS, University of Groningen; Department of Statistics, University of Oxford.

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

Stadtfeld, C., 2010. Who communicates with whom? Measuring communication choices on social media sites. In: Proceedings of the 2010 IEEE International Conference on Social Computing (SocialCom). Minneapolis, USA, pp. 564 –569.

Stadtfeld, C., 2011a. Measuring the influence of network structures on social interaction over time. In: Gaul, W., Geyer-Schulz, A., Schmidt-Thieme, L., Kunze, J. (Eds.), Challenges at the Interface of Data Analysis, Computer Science, and Optimization: Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V., Karlsruhe. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, in press.

Stadtfeld, C., 2011b. Website ESNA. `http://www.em.uni-karlsruhe.de/ref/esna`, visited: August 5, 2011.

Stadtfeld, C., Geyer-Schulz, A., 2010. Analysing event stream dynamics in two mode networks. `http://www.insna.org/awards/student.html`; visited August 5, 2011, INSNA Best Student Paper 2010.

Stadtfeld, C., Geyer-Schulz, A., 2011. Analyzing event stream dynamics in two mode networks: An exploratory analysis of private communication in a question and answer community. Social Networks 33 (4), 258–272.

Stadtfeld, C., Geyer-Schulz, A., Allmendinger, O., 2011. The influence of distance, time, and communication network structures on the choice of communication partners. In: Proceedings of the 2011 IEEE International Conference on Social Computing (SocialCom). Boston, USA, pp. 402–409.

Stadtfeld, C., Geyer-Schulz, A., Waldmann, K.-H., 2010. Estimating event-based exponential random graph models. In: Dreier, T., Krämer, J., Studer, R., Weinhardt, C. (Eds.), Information Management and Market Engineering. Vol. 2 of Studies on eOrganisation and Market Engineering. KIT Scientific Publishing, pp. 79–94.

Steglich, C., Snijders, T. A. B., Pearson, M., 2010. Dynamic networks and behavior: Separating selection from influence. Sociological Methodology 40 (1), 329–393.

Strauss, D., 11 1992. The many faces of logistic regression. The American Statistican 46 (4), 321–327.

Strauss, D., Ikeda, M., 3 1990. Pseudolikelihood estimation for social networks. Journal of the American Statistical Association 85 (409), 204–212.

van de Bunt, G. G., Van Duijn, M. A., Snijders, T. A., 1999. Friendship networks through time: An actor-oriented dynamic statistical network model. Computational & Mathematical Organization Theory 5, 167–192.

van Duijn, M. A. J., Gile, K. J., Handcock, M. S., 2009. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. Social Networks 31 (1), 52–62.

van Duijn, M. A. J., Zeggelink, E. P. H., Huisman, M., Stokman, F. N., Wasseur, F. W., 2003. Evolution of sociology freshmen into a friendship network. The Journal of Mathematical Sociology 27 (2-3), 153–191.

Waldmann, K.-H., Stocker, U., 2004. Stochastische Modelle: Eine anwendungsorientierte Einführung. Springer, Berlin.

Wang, P., Sharpe, K., Robins, G. L., Pattison, P. E., 2009. Exponential random graph (p*) models for affiliation networks. Social Networks 31 (1), 12–25.

Wasserman, S., 1980. Analyzing social networks as stochastic processes. Journal of the American Statistical Association 75 (370), pp. 280–294.

Wasserman, S., Faust, K., 1994. Social Network Analysis, 1st Edition. Vol. 8 of Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge.

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

Wasserman, S., Robins, G., 2005. An introduction to random graphs, dependence graphs, and p*. In: Carrington, P. J., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Vol. 27 of Structural Analysis in the Social Sciences. Cambridge University Press, New York, Ch. 8, pp. 148–161.

Whittle, P., 2007. Networks : Optimisation and Evolution. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Wilf, H. S., 1994. generatingfunctionology, 2nd Edition. Academic Press Inc., Boston.

Young, G. A., Smith, R. L., 2005. Essentials of Statistical Inference. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Zenk, L., Stadtfeld, C., 2010. Dynamic organizations. How to measure evolution and change in organizations by analyzing email communication networks. Procedia – Social and Behavioral Sciences 4, 14–25.

Zenk, L., Stadtfeld, C., Windhager, F., 2010. How to analyze dynamic network patterns of high performing teams. Procedia - Social and Behavioral Sciences 2 (4), 6418–6422.

Interactions between people are ubiquitous. When people make phone calls, transfer money, connect on social network sites, or visit each other, these actions can be collected as dyadic, directed, relational events.

Each of those events can be understood as driven by multiple individual decisions that at least partially involve rational considerations. As a whole, the many individually driven event decisions form social networks. In turn, these networks influence future event decisions. This book aims at developing models that allow to understand individual event decisions in the context of large social networks. In two extensive case studies, a new class of models is applied to empirical data.

Christoph Stadtfeld studied Information Engineering and Management in Karlsruhe, Strasbourg and Oxford. He received his doctoral degree in Economics from Karlsruhe Institute of Technology (KIT) and is now working as a postdoctoral researcher at the University of Groningen / ICS.

ISBN 978-3-86644-792-9

9 783866 447929 >